Fuzzy Natural Language Similarity Measures Through Computing With Words

N ADEL

PhD 2022

Fuzzy Natural Language Similarity Measures Through Computing With Words

NAEEMEH ADEL

A thesis submitted in partial fulfilment of
the requirements of
Manchester Metropolitan University
for the degree of Doctor of Philosophy

Department of Computing and Maths
Faculty of Science and Engineering
Manchester Metropolitan University

2022

Declaration of Authorship

I, Naeemeh Adel, declare that this thesis titled “Fuzzy Natural Language Similarity Measures Through Computing With Words” and the work presented within are my own and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification at Manchester Metropolitan University or any other university or institution of learning.

I have maintained professional integrity during all aspects of my research degree and, I have complied with the Institutional Code of Practice and the Regulations for Postgraduate Research Degrees.

Where I have consulted the published work of others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. This research received no external funding, other than the PGTA funding by Manchester Metropolitan University and I declare no conflict of interest.

Abstract

A vibrant area of research is the understanding of human language by machines to engage in conversation with humans to achieve set goals. Human language is naturally fuzzy by nature, with words meaning different things to different people, depending on the context. Fuzzy words are words with a subjective meaning, typically used in everyday human natural language dialogue and often ambiguous and vague in meaning and dependent on an individual's perception. Fuzzy Sentence Similarity Measures (FSSM) are algorithms that can compare two or more short texts which contain fuzzy words and return a numeric measure of similarity of meaning between them.

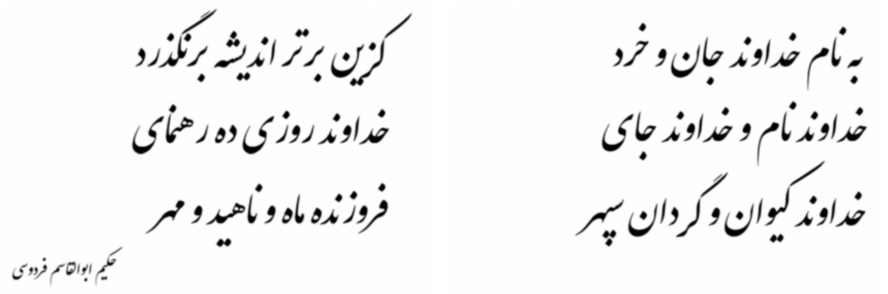
The motivation for this research is to create a new FSSM called FUSE (FUzzy Similarity mEasure). FUSE is an ontology-based similarity measure that uses Interval Type-2 Fuzzy Sets to model relationships between categories of human perception-based words. Four versions of FUSE (FUSE_1.0 – FUSE_4.0) have been developed, investigating the presence of linguistic hedges, the expansion of fuzzy categories and their use in natural language, incorporating logical operators such as 'not' and the introduction of the fuzzy influence factor.

FUSE has been compared to several state-of-the-art, traditional semantic similarity measures (SSM's) which do not consider the presence of fuzzy words. FUSE has also been compared to the only published FSSM, FAST (Fuzzy Algorithm for Similarity Testing), which has a limited dictionary of fuzzy words and uses Type-1 Fuzzy Sets to model relationships between categories of human perception-based words. Results have shown FUSE is able to improve on the limitations of traditional SSM's and the FAST algorithm by achieving a higher correlation with the average human rating (AHR) compared to traditional SSM's and FAST using several published and gold-standard datasets.

To validate FUSE, in the context of a real-world application, versions of the algorithm were incorporated into a simple Question & Answer (Q&A) dialogue system (DS), referred to as FUSION, to evaluate the improvement of natural language understanding. FUSION was tested on two different scenarios using human participants and results compared to a traditional SSM known as STASIS. Results of the DS experiments showed a True rating of 88.65% compared to STASIS with an average True rating of 61.36%. Results showed that the FUSE algorithm can be used within real world applications and evaluation of the DS showed an improvement of natural language understanding, allowing semantic similarity to be calculated more accurately from natural user responses.

The key contributions of this work can be summarised as follows: The development of a new methodology to model fuzzy words using Interval Type-2 fuzzy sets; leading to the creation of a fuzzy dictionary for nine fuzzy categories, a useful resource which can be used by other researchers in the field of natural language processing and Computing with Words with other fuzzy applications such as semantic clustering. The development of a FSSM known as FUSE, which was expanded over four versions, investigating the incorporation of linguistic hedges, the expansion of fuzzy categories and their use in natural language, inclusion of logical operators such as 'not' and the introduction of the fuzzy influence factor. Integration of the FUSE algorithm into a simple Q&A DS referred to as FUSION demonstrated that FSSM can be used in a real-world practical implementation, therefore making FUSE and its fuzzy dictionary generalisable to other applications.

Acknowledgments



به نام خداوند جان و خرد
خداوند نام و خداوند جای
خداوند کیوان و گردان سپهر

کزین برتر اندیشه برنگذرد
خداوند روزی ده رهنمای
فروزنده ماه و ناهید و مهر

حکیم ابوالقاسم فردوسی

In the name of God of life and wisdom
The God of fame in whom powers reside
Creator of the world & the orderly universal run

A worthier notion shall not arise
Provider, sustainer, the ultimate guide
The light giver to the Moon, Venus & the Sun

Shahnameh, an epic Persian poem
written by Ferdousi between 977 and 1010 C.E.

I must firstly acknowledge **Allah** for he is the one who has given me the strength during all my challenging moments in life, especially in the completion of this thesis.

I would like to dedicate this thesis to my beloved parents, without whom I would not be the person I am. My beautiful mother, **Narges Moussavi**, for her endless devotion, support, prayers, and encouragement. My caring father, **Dr. Morteza Adel**, for always inspiring me and offering me advice in times of need. My dear brother, **Vahid Adel** for his support. I am very fortunate to be your daughter, and you, my dear parents, have both been behind me through all my life's choices and with your love, pushed me to succeed in everything I put my heart into.

My sincere gratitude is expressed to my respected and inspiring supervisor **Professor Keeley Crockett**. She has given me her time so generously and expertly guided me through my PhD journey. Her masterful knowledge, wisdom, patience, inspiring support, belief in me and insightful feedback have steered me during all the stages of my PhD journey. I am so grateful and honoured to have been supervised by her, a true role model and inspiration.

Special acknowledgment and thanks go to Manchester Metropolitan University, Department of Computing and Maths, for offering me the Post Graduate Teaching Assistant (PGTA) position and subsequently the financial support it provided me.

Professor Darren Dancey, Head of Department, and **Professor Rob Aspin**, Deputy Head of Department for providing me with such rewarding teaching opportunities and the many prospects it presented. It has made me a more confident teacher and further fuelled my passion for academia. My countless students, who have shaped my journey as an academic, it has been so fruitful and rewarding teaching you all.

I would like to thank my extended supervisory team, **Professor Joao Paulo Carvalho;** early support from **Dr Alan Crispin** (now retired), and **Dr Annabel Latham,** who joined the supervisory team at a later stage. Sincere appreciation is extended to **Dr Daria Livesey,** who's invaluable guidance and advice have been of immense benefit to this research.

My dear friend, colleague and fellow PGTA **Lewis Evans,** whom I have shared many happy and fond memories with throughout our PhD/PGTA journeys.

Finally, all my incredible and loving friends and family, who with their love, patience, words of encouragement and support enabled me to get to the end, even when I thought I never would!

Table of Contents

List of Algorithms

List of Equations

List of Figures

List of Tables

List of Abbreviations

AHR = Average Human Rating

AI = Artificial Intelligence

CA = Conversational Agent

CASE = Computer Aided Software Engineering

COG = Centre of Gravity

CWW = Computing With Words

DS = Dialogue Systems

EIA = Enhanced Interval Approach

FAST = Fuzzy Algorithm for Similarity Testing

FI = Fuzzy Influence

FLS = Fuzzy Logic System

FOU = Footprint of Uncertainty

FS = Fuzzy Sets

FSSM = Fuzzy Semantic Similarity Measure

FUSE = FUzzy Similarity mEasure

HMA = Hao–Mendel Approach

HSP = Hedge Sentence Pair

IA = Interval Approach

IC = Information Content

ICC = Intra-Class Correlation Coefficient

LCS = Lowest Common Subsumer

LSA = Latent Semantic Analysis

MF = Membership Function

MFWD = Multi Fuzzy Word Dataset

NLP = Natural Language Processing

Per-C = Perceptual Computing

Q&A = Question and Answering

RQ = Research Question

SEMILAR = the SEMantic simILARity toolkit

Sentence Pair = SP

SFWD = Single Fuzzy Word Dataset

SSM = Sentence Similarity Measure

STSS = Short Text Semantic Similarity

SVD = Singular Value Decomposition

WEM = Word Embedding Models

WFH = Working From Home

WSD = Word Sense Disambiguation

List of Author Publications

The following publications have been generated as a result of the work in this thesis:

List of Conference Publications:

1. **Fuzzy Influence in Fuzzy Semantic Similarity Measures.** N Adel, K Crockett, JP Carvalho, V Cross. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, 2021. DOI: 10.1109/FUZZ45933.2021.9494535
2. **Using Fuzzy Set Similarity in Sentence Similarity Measures**. V Cross, V Mokrenko, K Crockett, N Adel. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), UK (virtual congress), 2020. DOI: 10.1109/FUZZ48607.2020.9177836
3. **Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures**. N Adel, K Crockett, D Chandran, JP Carvalho. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), UK (virtual congress), 2020. DOI: 10.1109/FUZZ48607.2020.9177605
4. **Ontological and fuzzy set similarity between perception-based words**. V. Cross, V. Mokrenko, K. Crockett, N. Adel. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). New Orleans, USA, 2019. DOI:978-1-5386-1728-1/19/
5. **Human Hedge Perception – and its Application in Fuzzy Semantic Similarity Measures**. N. Adel, K. Crockett, A. Crispin, JP. Carvalho, D. Chandran 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). New Orleans, USA, 2019. DOI: 978-1-5386-1728-1/19/
6. **FUSE (Fuzzy Similarity Measure)-A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets.** N. Adel, K. Crockett, A. Crispin, D. Chandran, JP. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Brazil, 2018. DOI: 10.1109/FUZZ-IEEE.2018.8491641
7. **Application of fuzzy semantic similarity measures to event detection within tweets.** K. Crockett, N. Adel, J. O'Shea, A. Crispin, D. Chandran, et al. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Italy, 2017. DOI: 10.1109/FUZZ-IEEE.2017.8015488, 2017

List of Submitted Journal Publications

The following journal paper has been published in IEEE Access on 28th July 2022:

1. **An Interval Type-2 Fuzzy Ontological Similarity Measure.** N. Adel, K. Crockett, D. Livesey and J.P. Carvalho

CHAPTER 1

CHAPTER 1: INTRODUCTION

1.1 - Background

Natural language processing (NLP) is a well-known sub-field of artificial intelligence and linguistics (Nadkarni et al., 2011). The ultimate aim of NLP is to process meaning from chunks of naturally occurring text (Liddy, 2001). To achieve this, NLP must distinguish between nouns, verbs, and adjectives etc, in a given text. NLP techniques, effectively led to the development of systems that could communicate with words and exchange dialogue with humans. Some examples of NLP use include Weizenbaum's ELIZA (O'Dell and Dickson, 1984) built to replicate the conversation between a psychologist and a patient, simply by permuting or echoing the user input. Winograd's SHRDLU (Winograd, 1973) simulated a robot that manipulated blocks on a tabletop. LUNAR was developed by Woods (Woods, 1973) as an interface system that gave information about lunar rock samples. PARRY (Colby, 1981) attempted to symbolise a theory of paranoia in a system. Instead of single keywords, it used groups of keywords, and used synonyms if keywords were not found.

According to Feldman (Feldman, 1999) there are seven levels of natural language processing, which when combined, allow the extraction of meaning from text or spoken language, thus giving a more capable NLP system (Liddy, 2001):

1. **Phonetic** - which deals with pronunciation.
2. **Morphological** - which deals with the smallest parts of words that carry meaning, such as suffixes and prefixes.
3. **Lexical** - which deals with the meaning of words specially if words have more than one meaning.
4. **Syntactic** - which deals with the grammatical structure of the sentence.
5. **Semantic** - which deals with the meaning of words and sentences.
6. **Discourse** - which deals with the structure of different kinds of text using document structures.
7. **Pragmatic** - which deals with the knowledge that comes from the outside world, i.e., from outside the content of the document.

There are many applications of NLP, but ones particularly relevant to this work are (Liddy, 2001):

- **Question & Answering (Q&A) Systems** - which provides the user with either just the text of the answer itself or answer-providing passages. A Q&A system typically asks the user a question in natural language and returns the right answer to this question as opposed to returning a set of documents/links/images etc. which are deemed relevant to the search query (Deriu et al., 2021).
- **Dialogue Systems (DS)** - which focus on a narrowly defined application. Dialogue systems usually take turns with the user and based on the user's response or utterance; the next dialogue response is activated. The conversation can vary in method such as text based, image based or voice based (Car et al., 2020).

The term semantic similarity refers to the similarity between two objects (Crockett et al., 2017). Semantic similarity is, therefore, a complex concept with a long history in cognitive psychology and linguistics (Rubenstein and Goodenough, 1965), which can analyse the deep semantic structure of a short text to convey meaning. Semantic similarity methods usually give a ranking or percentage of similarity between texts as opposed to a binary decision (Chandrasekaran and Mago, 2021). The task of assessing the semantic similarity between short texts has been a central problem in NLP, due to its importance in a variety of applications (Alnajran, 2019). Some early text similarity applications used for text classification include information retrieval (Rocchio, 1971), automatic word sense disambiguation (Lesk, 1986), and extractive text summarization (Salton and Buckley, 1988). Semantic similarity is an important and fundamental concept in Artificial Intelligence (AI) and has applications in word-sense disambiguation, image retrieval and conversational agents (D. Lin, 1998). Measuring semantic similarity can be performed at various levels, ranging from words, phrases and sentences to paragraphs and documents (Alnajran, 2019). Each of these categories employ different methods and techniques to gauge the underlying meaning at that particular level (Alnajran, 2019). A problem in the field of semantic sentence similarity is the inability of existing measures to capture the meaning of fuzzy words. **A fuzzy word can be defined as a natural language word with subjective meanings, e.g., *huge*, *small*, *hot* and *cold*, that is characteristically used in every day human natural language dialogue.** Fuzzy words are often ambiguous in meaning, since they are based on an individual's perception (Chandran et al., 2013). This can best be explained using Figure 1 and Figure 2 as an example.

Figure 1 - Mount Everest

Figure 2 - Malham Cove

Figure 1[1] is an image of Mount Everest, which is the highest mountain on Earth above sea level located in the Mahalangur Himal sub-range of the Himalayas, with a height of 29,031.7 ft. Figure 2[2] is an image of Malham Cove, which is a large, curved limestone formation north of the village of Malham, North Yorkshire, England, with a height of 260 ft.

If the reader was asked, "*How would you describe the mountain?*" in each of these two figures in terms of height, there will be different answers given. One might say Figure 1 is *Huge*, *Tall*, *Big*. One might say Figure 2 is *Huge*, *Tall*, *Big*. Now if the same question was asked of a 6-year-old who was standing at the bottom of Figure 2 looking up, "*How would you describe this mountain?*" They might reply its *Massive*, *Enormous*, *Gigantic*. If asked to compare the two figures together, the reader might say Figure 1 is *Gigantic*, but Figure 2 is *Big*. A 6-year-old might say Figure 1 is *Enormous* and Figure 2 is *Massive*. This simple example alone can explain how fuzzy words differ based on an individual's perception.

Fuzzy set theory has been used in the field of Computing with Words (CWW) to represent people's perceptions of fuzzy words. "*CWW is a methodology in which the objects of computation are words and propositions drawn from natural language*" (Zadeh, 1996). Full details about CWW will be provided in Section 3.2.

Original work was limited to Type-1 Fuzzy Sets (FS), which caused limitations for CWW due to the linguistic uncertainty of Type-1 FS. Since real world applications are often faced with

[1] - Source: https://www.independent.co.uk/climate-change/news/glacier-ice-melt-mount-everest-b2011157.html

[2] - Source: https://www.bbc.co.uk/news/uk-england-york-north-yorkshire-35026529

multiple sources of uncertainty (J. M. Mendel, 2007), Type-2 FS were introduced. A normal Type-2 FS is three dimensional, where the third dimension is the value of the membership function, known as the Footprint of Uncertainty (FOU). However, for Interval Type-2 FS, this third dimension has a consistent value, meaning no new information is stored in this third dimension (J. M. Mendel, 2007). The benefit of Type-2 FS is its three-dimensional nature, which provides additional degrees of freedom, that makes it possible to directly model uncertainties. However, this also makes them difficult to understand and use because:

(i) There is no simple collection of well-defined mathematical terms that allow effective communication and representation of words.

(ii) Derivations of the formulas is a difficult concept to understand.

(iii) Type-2 FS are computationally more complicated than Type-1 FS (J. M. Mendel and John, 2002).

Models of fuzzy words produced using Type-2 FS have not been used within semantic similarity measures (SSM) (to the best of the authors knowledge whilst undertaking this research) and this research provides an opportunity to exploit their ability to represent uncertainty in the modelling of fuzzy words within short texts. This is further explained in Chapter 3.

Recent works in the field of fuzzy natural language processing include (Yang, 2021) who has attempted to translate Chinese literature to English while maintaining its fuzziness. (B. Wang et al., 2020) proposed a fuzzy computing model to improve the performance of sentiment classification in different online reviews. (Phan et al., 2020) created a fuzzy model to improve the performance of Tweet Sentiment Analysis (TSA) in order to have a better understanding of a user's emotions when they upload a tweet.

Integration of fuzzy words in semantic similarity algorithms can therefore allow a way to capture and measure human similarity in a given context.

A dialogue system, sometimes referred to as a conversational agent (CA) is a computer program which interacts with a user through natural language dialogue and provides some form of service (J. O'Shea et al., 2013). However, they typically suffer from high maintenance in updating dialogue patterns for new scenarios due to the huge number of language patterns within the scripts. The key impact of the work presented in this thesis will lie in its ability as a machine, to semantically comprehend and thus, communicate effectively with a human in a specific domain using natural language. Incorporation of the semantic meaning of fuzzy words within the dialogue system will improve the quality of human-machine interaction. Dialogue systems match human utterances to machine utterances to engage the user in some form of conversation. Traditionally, this was done using pattern matching algorithms (J. O'Shea et al., 2013) which were cumbersome, thus more recently semantic similarity measures have been used (K. O'Shea, 2012) which has led to a reduction of patterns and therefore less maintenance costs.

1.2 - Motivation and Problem Statement

To date, to the best of the authors knowledge whilst conducting this research, there is only one fuzzy semantic similarity algorithm that has been developed, named Fuzzy Algorithm for Similarity Testing (FAST) (Chandran, 2013). FAST is an ontology-based similarity measure that uses concepts of fuzzy and CWW to allow for the accurate representation of fuzzy based words. Through human experimentation, fuzzy sets were created for six categories of words using Type-1 FS (*Size & Distance*, *Age*, *Goodness*, *Frequency*, *Temperature* and *Level of Membership*). The use of Type-1 FS causes a weakness for FAST, since these words are not a true representation of each category, because the rating of the words is still the subjective opinion of those individuals (J. M. Mendel and John, 2002). This adversely affects the accuracy of the categories by the potential bias of the individual's views that are used to quantify fuzzy words, which is further discussed in Chapter 3.

The motivation behind this research is the development of a new Type-2 Fuzzy Semantic Similarity Measure (FSSM). This new FSSM will include a wider coverage of fuzzy words and the inclusion of linguistic hedge measurements and negation words such as '*not*'. The new FSSM will first be compared against other known similarity measures over a number of published and gold standard datasets. The FSSM will then be evaluated in a Question and Answering type dialogue system to investigate whether the ability to model and interpret fuzzy words can enable an improved machine dialogue response to a human utterance.

Linguistic hedges map a fuzzy set to another and modifies the shape of fuzzy sets (H. Wang et al., 2018). Typically, hedges can be classified into two categories, which are intensified hedges (such as '*very*') and weakened hedges (such as '*more*') (H. Wang et al., 2018). Hedges include adverbs such as *very*, *somewhat*, *quite*, *more* and *slightly* (Negnevitsky, 2005). Linguistic hedges can help further improve the precision of sentence similarity. This is achieved by obtaining a higher correlation of similarity with human ratings, since hedges can help reflect human thinking. Thus, making them an important part of measuring human perceptions of the similarity of short texts (Adel et al., 2019). This is further explained in Chapter 6.

1.3 - Aims and Objectives

The aim of this research is to develop a new natural language Fuzzy Semantic Similarity Measure (FSSM) based on Type-2 Fuzzy Sets (FS) for integration into Dialogue Systems (DS), to provide an improved language understanding and learning ability.

The research, which is to create a new FSSM and evaluate through integration in a Q&A dialogue system will address the following primary research questions:

RQ1. Investigate the feasibility of utilising Type-2 Fuzzy Sets and their representation of an individual's perception of fuzzy words and evaluate the suitability of the resulting fuzzy word models for incorporation into a Fuzzy Semantic Similarity Measure (FSSM).

RQ2. Can a Type-2 FSSM be embedded into a Q&A dialogue system with an improved success rate of utterance - response matches compared to traditional Semantic Similarity Measures (SSM)?

To be able to answer these research question's, the following project objectives have been set:

1. Conduct research into Type-2 FS representations in the context of natural language interpretation and modelling, and review existing semantic similarity measures, dialogue systems (engines, scripting languages and applications) and FSSM benchmark evaluation datasets (Chapter 2 & 3).
2. Investigate, develop and implement a new Fuzzy Semantic Similarity Measures (FSSM) using Type-2 FS for measuring similarity between natural language user utterances, by employing techniques from CWW (Chapter 4).

3. Evaluate the new FSSM using Type-2 FS with benchmark FSSM datasets to produce a comparative study (Chapter 5).
4. Investigate, develop and evaluate the representation of linguistic hedges and negation values in FSSM using correlations with human ratings (Chapter 6).
5. Develop a methodology for the scripting of a semantic Q&A dialogue system using a short text semantic similarity approach, using the new FSSM (Chapter 7).
6. Implement a prototype dialogue system, for a specific application domain, using the methodology developed in (5), and evaluate the new FSSM within the prototype dialogue system using prototypical scenarios and human participants (Chapter 7).

1.4 - Research Methodology

This research was granted ethical approval by Manchester Metropolitan Universities Science and Engineering Research Ethics and Governance Committee (EthOS Reference Number: 11759).

To achieve this research project, a number of iterations is needed to develop the new proposed FSSM algorithm. The name of the core FSSM algorithm is FUSE (FUzzy Similarity mEasure). Following a thorough evaluation of the literature and state of the art on SSM, and current work in the field of CWW, the framework for the FUSE algorithm was established. The FUSE algorithm was developed in three core phases and then incorporated into a dialogue system in the final phase.

Phase 1: The aim of Phase 1 was to investigate and develop a method of modelling human fuzzy words using Interval Type-2 Fuzzy sets and use them to build ontologies. The methods used to achieve this are defined as:

1. Conducting a systematic literature review of modelling words using Type-2 and Interval Type-2 fuzzy sets and state-of-the-art methods for capturing and analysing human ratings.
2. Developing a method to create six initial fuzzy categories of words using existing categories of words developed in FAST (Chandran, 2013) and expanding them to improve natural language coverage. The method included the creation of a fuzzy dictionary (Section 4.3).

3. Designing and conducting a series of experiments (adopting a methodology designed by O’Shea) (J. O’Shea et al., 2013) to obtain scales for each fuzzy word from human participants.
4. Modelling the fuzzy words in each category using Interval Type-2 Fuzzy Sets using techniques developed by Hao-Mendel known as the HMA approach (Hao and Mendel, 2015).
5. Designing and developing a series of six fuzzy category ontologies based on the synthesis of ideas from existing semantic similarity measures such as STASIS (Li et al., 2003), WordNet (Miller, 1995) and FAST (Chandran, 2013).

Phase 2 - The aim of Phase 2 is to create the first version of the FUSE algorithm, referred to as FUSE_1.0. The methods used to achieve this are defined as:

1. Design and development of a fuzzy semantic similarity measure for FUSE_1.0 by:
    a. Short text similarity determined by word similarity, path depth referred as the Lowest Common Subsumer and path length in the ontology
    b. Fuzzy word similarity was determined using fuzzy ontologies
    c. Non fuzzy words were determined using the WordNet ontology
    d. FUSE_1.0 computes overall similarity through a combination of syntactic and semantic weighting
2. Providing an illustrative example of FUSE_1.0 with sample sentence pairs.
3. FUSE_1.0 evaluation on three published datasets and result correlation with human ratings was compared with two other measures, STASIS which is a traditional SSM and FAST which is the only other available FSSM; fully described in Chapter 5.

Phase 3 - The aim of this phase was to improve the performance of FUSE_1.0 following results of empirical experiments; therefore, a number of different versions were created each addressing an issue:

1. **FUSE_2.0**, the introduction of linguistic hedges to the FUSE algorithm and expansion from six categories to nine (Section 6.2 and 6.3).

2. **FUSE_3.0**, the introduction of negation operators to the FUSE algorithm (Section 6.4).
3. **FUSE_4.0**, the introduction of a Fuzzy Influence factor to the FUSE algorithm, which caters for fuzzy words not in the same category (Section 6.5).
4. Evaluation of the performance of each version of FUSE, through a series of experiments. FUSE_2.0 was evaluated on five datasets and result correlation with human ratings compared with four other SSM. This evaluation was run at each stage to test the versions and improvements compared to human ratings (Section 6.3.3). FUSE_3.0 was fully evaluated when effects of negation on natural language utterances were explored via incorporation in the dialogue system (Section 7.4). FUSE_4.0 was evaluated on three datasets and result correlation with human ratings compared to four other SSM's as well as previous versions of FUSE (Section 6.5).

Phase 4 - The aim of this phase was to incorporate FUSE_2.0 and FUSE_4.0 into a dialogue system known as FUSION and test its correlation with human ratings using two separate scenarios (Chapter 7):

1. **FUSION_V1** - Design of the first version of the Dialogue System, referred to as FUSION_V1 through the development of a Q&A scenario using FUSE_2.0; The scenario selected was based on a set of questions used within the context of rating a participants experience of visiting a local café.
2. **Evaluation of FUSION_V1** - A dataset of participant results and a set of prototypical answers for each question was created. Participants were recruited and asked to visit a local café, purchase a drink of their choice, and observe their surroundings. They were then asked to evaluate their experience by answering a set of questions asked by FUSION_V1 Dialogue System. The dataset was used to evaluate FUSE_2.0 in the context of FUSION_V1 and results were compared with those of a traditional SSM, STASIS.
3. **FUSION_V2** - Design of the second version of the Dialogue System, referred to as FUSION_V2 by incorporating it to a Q&A scenario using FUSE_4.0 which utilised the negation operator and the fuzzy influence factor (two novel contributions within the FUSE algorithm). The scenario selected was based on two sets of questions used within

the context of rating your experience surrounding working from home (WFH) conditions. Participants accessed and evaluated FUSION_V2 online due to Covid-19 and social distancing restrictions.

4. **Evaluation of FUSION_V2** - Two datasets of participant results and two sets of prototypical answers reflecting each set of questions was created. Participants were asked to evaluate their experience of working from home by answering two sets of questions asked by FUSION_V2 Dialogue System. The two datasets were used to evaluate FUSE_4.0 (which also incorporated the negation operator and the fuzzy influence factor) in the context of FUSION_V2 and results were compared with those of a traditional SSM, STASIS.

1.5 - List of Contributions

The contributions made from this research are listed below, a full breakdown can be seen in (Section 1.8 and Figure 3):

- A new methodology for modelling fuzzy words was created which utilised Interval Type-2 fuzzy sets to represent human perception-based words. This work led to the creation of a fuzzy dictionary for six fuzzy categories which contained defuzzified numerical measures derived from average human ratings obtained using Interval Type-2 fuzzy set approach (Chapter 4). The fuzzy dictionary is a useful resource which can be used by other researchers in the field of NLP.
- Development of a fuzzy semantic similarity measure known as FUSE (FUzzy Similarity mEasure), with its first version (FUSE_1.0) using Interval Type-2 fuzzy sets and the inclusion of the newly developed fuzzy dictionary for six fuzzy categories using Interval Type-2 fuzzy sets (Chapter 4).
- Development of four versions of the FUSE algorithm which includes the incorporation of linguistic hedges and category expansion to nine fuzzy categories (FUSE_2.0). The inclusion of negation operators (FUSE_3.0) which permits a novel ability to apply fuzzy complement operators to fuzzy words modelled by Interval Type-2 Fuzzy Sets. Up to this point, fuzzy word similarity was only computed using the fuzzy category

ontologies if fuzzy words belonged to the same fuzzy category. The introduction of a fuzzy influence factor (FUSE_4.0) allowed the fuzzy measure of a word to contribute to the overall similarity measure regardless of the fuzzy words in a pair of sentences belonging to the same fuzzy category or not (Chapter 6).

- The development of three new fuzzy categories resulting in an expansion of the fuzzy dictionary for nine fuzzy categories used for the FUSE_4.0 algorithm. This presents fuzzy words and their defuzzified numerical measure derived from average human ratings obtained using Interval Type-2 fuzzy set approach. The fuzzy dictionary of FUSE_4.0 can be used by other researchers in the field of NLP with other fuzzy applications such as semantic clustering (Appendix C).
- Comparison of the different versions of the FUSE algorithm with other state of the art Semantic Similarity Measures (SSM), across several published and newly created datasets (Chapter 5 and 6).
- Integration of FUSE_2.0 and FUSE_4.0 into two versions of a simple Q&A Dialogue System referred to as FUSION_V1 and FUSION_V2 respectively. Textual human responses were captured using two different scenarios (visit to a local café for FUSION_V1 and working from home for FUSION_V2). The integration of the FUSE algorithm into the FUSION dialogue system demonstrated that FSSM can be used in a real-world practical implementation, by incorporation into two different scenarios of a Q&A Dialogue System. Evaluation of the FUSION Dialogue Systems was achieved through comparison with traditional semantic similarity measures, and results showed that a FSSM incorporated into a dialogue system is able to improve language understanding (Chapter 7).

1.6 - Thesis Overview

Figure 3 shows the research overview with the objectives of the research and the chapters they are linked to along with the associated resulting publications.

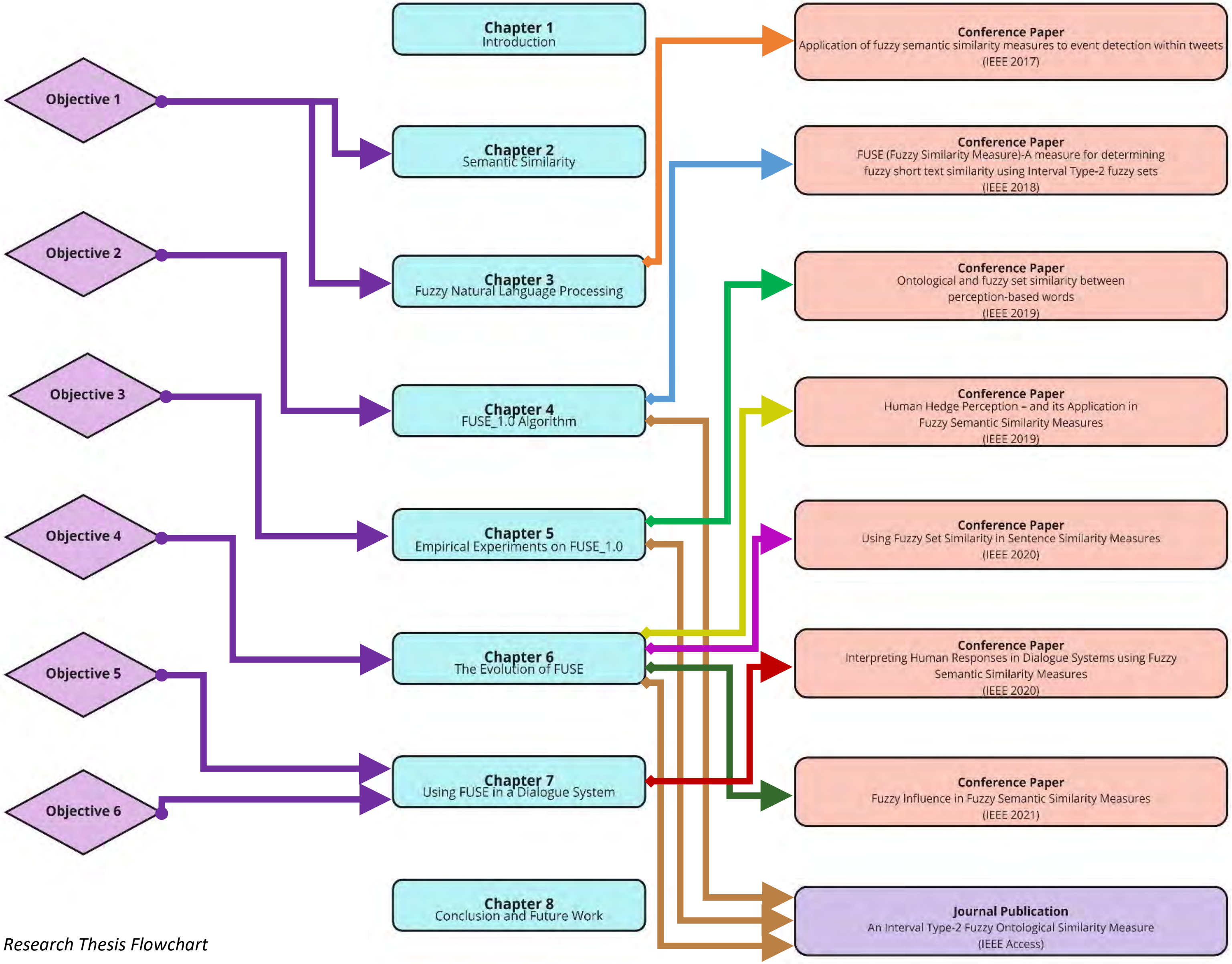

Objective 1
Objective 2
Objective 3
Objective 4
Objective 5
Objective 6
Chapter 1 Introduction
Chapter 2 Semantic Similarity
Chapter 3 Fuzzy Natural Language Processing
Chapter 4 FUSE_1.0 Algorithm
Chapter 5 Empirical Experiments on FUSE_1.0
Chapter 6 The Evolution of FUSE
Chapter 7 Using FUSE in a Dialogue System
Chapter 8 Conclusion and Future Work
Conference Paper Application of fuzzy semantic similarity measures to event detection within tweets (IEEE 2017)
Conference Paper FUSE (Fuzzy Similarity Measure)-A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets (IEEE 2018)
Conference Paper Ontological and fuzzy set similarity between perception-based words (IEEE 2019)
Conference Paper Human Hedge Perception – and its Application in Fuzzy Semantic Similarity Measures (IEEE 2019)
Conference Paper Using Fuzzy Set Similarity in Sentence Similarity Measures (IEEE 2020)
Conference Paper Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures (IEEE 2020)
Conference Paper Fuzzy Influence in Fuzzy Semantic Similarity Measures (IEEE 2021)
Journal Publication An Interval Type-2 Fuzzy Ontological Similarity Measure (IEEE Access)

Figure 3 - Research Thesis Flowchart

1.7 - Conclusion

This chapter provided a brief background on natural language processing (NLP). It highlighted the motivation behind the research proposed in this thesis before delving into the problems this research is looking to address. The aims and objectives set out for this research were broken down, and how the research will be tackled and what research questions it will try to answer shown.

The research methodology provided a brief summary of the methods used in the different phases of the research. The thesis overview presented the objectives set out in this research and how they are addressed in each chapter of this thesis. This is also presented with a list of publications that have resulted following the completed work in each chapter(s).

Chapter 2 will focus on semantic similarity and explain the history and concepts behind word and sentence similarity. Existing semantic similarity measures will be reviewed and compared in terms of the approaches that they use. The chapter will examine the challenges faced with sentence similarity, including word sense disambiguation before discussing word embedding models and their use within semantic similarity measures.

Finally, the challenges that arise from collecting human ratings for word or sentence similarity is addressed and the methods used to evaluate similarity measures, before bringing the chapter to a close by discussing some applications of sentence similarity.

CHAPTER 2

CHAPTER 2: SEMANTIC SIMILARITY

2.1 - Introduction

This chapter will first introduce word and sentence similarity measures and their use in semantic similarity, before moving on to providing an overview of four published sentence similarity measures and how they compare. Particular attention is based on an algorithm called STASIS (Li et al., 2006) which provides the foundations for the research presented in this thesis.

This chapter also examines some of the challenges with semantic similarity including word sense disambiguation. Word embedding models and their use in SSM is discussed and why this approach was not used for this research.

Finally, the problems and challenges associated with collecting human ratings for evaluating word and sentence similarity measures will be discussed. The chapter concludes by providing a brief description of some applications of sentence similarity.

2.2 - Word Similarity Measures

The study of semantic similarity between words has been a part of natural language processing and information retrieval for many years. Similarity between two words is often represented by the similarity between concepts associated with the two words. Similarity between words is influenced by the context in which those words are presented. For example, if the context is '*the outside covering of living object',* then *skin* and *bark* are more similar as they both cover outside parts of living object, i.e., *trees* and *humans* as an example; however, if the context was to change to '*body parts'* then *skin* and *hair* would be more similar than *skin* and *bark* (Li et al., 2003). Nevertheless, this becomes more complex as the number of words increase and turns into a sentence or short texts.

There are two main approaches to calculating word similarity. The first method known as a text-based approach relies on the use of a large corpus or word definition and uses statistical data to estimate a semantic similarity score using these sources (Sebti and Barfroush, 2008). In this approach, word relationships are often derived from their occurrence distribution in a corpus (Grefenstette, 1992). The second method referred to as the structure-based approach uses relations and the hierarchy of a thesaurus taxonomy such as WordNet (Miller, 1995). WordNet is an online lexical semantic database, developed at Princeton University by a group led by Miller (Miller, 1995) which will be explained in more detail later in this section. The distance between nodes, referred to as the depth or concept, specify the similarity measure. Path length also plays a contributing factor in calculating the similarity (Z. Wu and Palmer, 1994). Resnik introduced a new factor, referred to as Information Content (IC) in which the path length and depth were combined to give a semantic similarity measure in a taxonomy (Resnik, 1995). The notion of information content in any given concept i.e., sentence pair, is directly related to the frequency of the term in a given document collection. The frequencies of terms in any given taxonomy (such as WordNet) are estimated using noun frequencies in a large collection of texts (Resnik, 1995). A natural, time-honoured way to evaluate semantic similarity in a taxonomy is to measure the distance between the nodes corresponding to the items being compared, the shorter the path from one node to another, the more similar they are (Resnik, 1999). The idea behind semantic similarity information content metrics is that each concept includes a lot of information in WordNet. The more common information two concepts share, the more similar the concepts are (Resnik, 1995). In 1995 Resnik first proposed an information content-based similarity metric (Resnik, 1995). Resnik assumed that for a concept *c*,

$$IC(c) = -\log p(c)$$

Equation 1 (Source: Resnik, 1995)

Where *p(c)* is the probability of encountering an instance of concept *c* (Meng et al., 2014).

Jiang and Conrath (Jiang and Conrath, 1997) and Lin (D. Lin, 1998) also use the IC concept to calculate similarity in *is-a* hierarchies.

Jiang and Conrath presented an approach for measuring semantic similarity or distance between words and concepts in 1997 (Jiang and Conrath, 1997). The proposed measure is a combined approach that inherits the edge-based approach of the edge-counting scheme, which is enhanced by the node-based approach of the information content calculation. If the comparison of the concepts shares a lot of information, then the *IC* will be high and the semantic distance between the compared concepts will be smaller (Jiang and Conrath, 1997).

The edge-based approach for word similarity is a more natural and direct way of evaluating semantic similarity in a taxonomy. It estimates the distance (e.g., edge length) between nodes, which corresponds to the concepts/classes being compared. Given the multi-dimensional concept space, the conceptual distance can conveniently be measured by the geometric distance between the nodes representing the concepts. Noticeably, the shorter the path from one node to the other, the more similar they are (Jiang and Conrath, 1997).

Li et al., uses multiple information sources to calculate the semantic similarity of concepts. They proposed a metric based on the assumption that information sources are infinite to some extent, as humans would have compared word similarity with a finite interval between completely similar and nothing similar (Li et al., 2003). Naturally, the transformation between an infinite interval to a finite one is non-linear (Meng et al., 2014). Li et al., define local semantic density as a monotonically increasing function of *wsim*(w_1, w_2) (Li et al., 2003):

$$f_3(wsim) = \frac{e^{\lambda.wsim(w_1,w_2)} - e^{-\lambda.wsim(w_1,w_2)}}{e^{\lambda.wsim(w_1,w_2)} + e^{-\lambda.wsim(w_1,w_2)}}$$

Equation 2 (Source: Li et al., 2003)

Where λ > 0. If λ → ∞, then the information content of words in the semantic nets are not considered (Li et al., 2003; Meng et al., 2014).

One taxonomy which is often used in the field of semantic similarity is WordNet, which is used by the likes of Li et al. (Li et al., 2003) or Deerwester (Deerwester et al., 1990). WordNet (Miller, 1995) developed by Princeton University is a large lexical database in English. Words are grouped into sets referred to as cognitive synonyms (synsets), each expressing a distinct concept (Miller, 1995). Each synset also contains a brief definition referred to as a *gloss*, and one or more short sentences illustrating the use of the synset members. For example, looking at Figure 4 (Wubben, 2008), *CPU*, *keyboard* and *monitor* are synsets of the concept *computer.*

In addition to providing these groups of synonyms to represent a concept, WordNet also connects concepts using a variety of semantic relations (Miller, 1995). These semantic relations for nouns include:

- Hyponym/Hypemym (IS-A / HAS A)
- Meronym/Holonym (Part-of / Has-Part)

- Meronym/Holonym (Member-of / Has-Member),
- Meronym/Holonym (Substance-of / Has-Substance)

Each of these semantic relations is represented by pointers between word forms or between synsets. More than 116,000 pointers represent semantic relations between WordNet words and word senses (Miller, 1995).

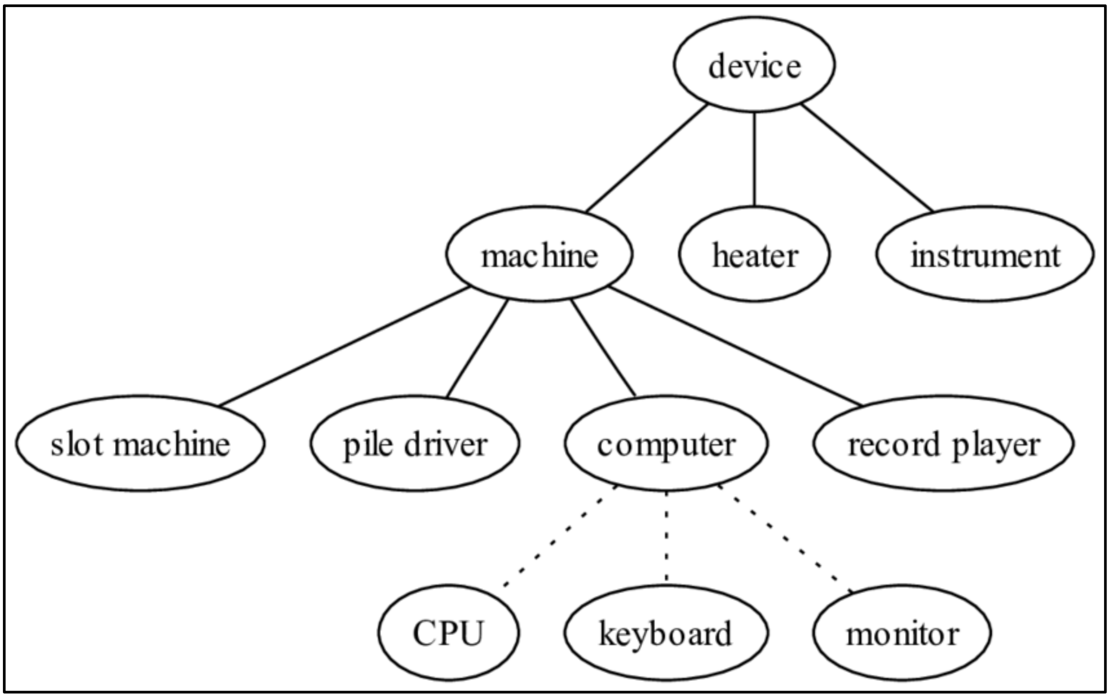


Figure 4 - WordNet Example (Source: Wubben, 2008)

2.3 - Semantic Similarity Measures - An Overview

Semantic similarity refers to similarity between two concepts in a taxonomy such as WordNet or CYC upper ontology (D. Lin, 1998). Semantic similarity is an important fundamental concept in AI and many other fields, since correct understanding of semantic information can lay a solid theoretical foundation for similarity calculation (Han et al., 2021). Examples of semantic similarity being used include word sense disambiguation (Resnik, 1999), automatic hypertext linking (Green, 1999), image retrieval (Smeulders et al., 2000), multimodal document retrieval (Srihari et al., 2000), paraphrasing identification using the Arabic language (Alian and Awajan, 2020), biomedical text mining (Lara-Clares et al., 2021), information retrieval (Po, 2020), document clustering (Mohammed et al., 2021) and healthcare applications (Zhang et al., 2021).

In NLP, understanding semantics correctly is crucial for understanding lexical diversity and uncertainty. This understanding of being able to identify the context of the information allows a more accurate semantic similarity calculation, which is why it has become a key factor in NLP (Han et al., 2021).

2.4 - Evaluation of Word and Sentence Similarity Measures

When comparing similarity for two sentences, words alone should not be looked at, but rather word order should also be considered. A sentence or short text is defined as having 10-20 words in length (J. O'Shea et al., 2013). This is best explained by using an example. Take the following two sentences, S_1 and S_2:

S_1: *A small fish in a big pond*

S_2: *A big fish in a small pond*

These two sentences contain the same words and the same number of words, furthermore, most words appear in the same order. The only difference is that *small* and *big* have swapped places in S_1 and S_2. It is clear for a human interpreter that these two sentences are only similar to some extent. The dissimilarity between them is the result of the difference in word order. Therefore, any efficient computational method for sentence similarity must take into account the impact of word order (Li et al., 2004).

Measuring semantic similarity of word pairs and sentence pairs is a general issue in linguistics, cognitive science, and artificial intelligence. It has been successfully applied in word sense disambiguation (Patwardhan et al., 2003), semantic annotation and summarisation (Sánchez et al., 2011; C.Y. Lin and Hovy, 2003), question and answering systems (Tapeh and Rahgozar, 2008), and information extraction (Atkinson et al., 2009). The measurement of word semantic similarity is also present in various software domains where Gomes (Gomes et al., 2006) presented an approach to software design using analogy, which is integrated in a Computer Aided Software Engineering (CASE) tool named REBUILDER, which comprised a Knowledge Base with several types of knowledge, including WordNet (Gomes et al., 2006).

Further examples of semantic similarity on sentences can be seen in bio-informatics domains such as The Gene Ontology (Lord et al., 2003), which is an annotation of gene products comprising of orthogonal taxonomies that hold terms describing the molecular function, biological process, and cellular component for a gene product (Lord et al., 2003). Therefore, a proper metric is crucial for improving the performance of the bulk of applications relying on semantic similarity.

2.5 - Types of Semantic Similarity Measures

There are many sentence measures available to date which are beyond the scope of this research. This section covers a brief historical overview which focuses on the fundamental approaches to semantic similarity but specifically the measures used for comparison in this research and the associated justification.

2.5.1 - LSA

Latent Semantic Analysis (LSA) is a fully automatic similarity measure used to compare words, sentences or passages. LSA does not use any natural language processing techniques or humanly constructed resources such as dictionaries, thesaurus, or lexical reference systems such as WordNet to compute semantic and syntactic relations between two utterances. Its only input which it uses are large amounts of text with an unsupervised learning technique (Dumais, 2004).

LSA uses four main steps for analysis which are explained below (Dumais, 2004):

1. **Term-Document Matrix:** This first step consists of a large collection of text that is represented as a term-document matrix. In this matrix, rows are individual words and columns are documents or smaller units such as passages or sentence. Individual cell entries contain the frequency with which a term occurs in a document. The order of words in the matrix is not important as "*bag of words*" representation is used.

2. **Transformed Term-Document Matrix:** In this step, the entries in the matrix are transformed and the best performance is observed depending on the frequencies cumulated in a sublinear fashion.

3. **Dimension Reduction:** In the third step, a reduced-rank Singular Value Decomposition (SVD) is performed on the matrix, in which the *k* largest singular values are retained, and the remainder set to 0. The resulting reduced-dimension SVD representation is the best k-dimensional approximation to the original matrix, in the least-squares sense. Each document and term are now represented as a k-dimensional vector in the space derived by the SVD.
4. **Retrieval in Reduced Space:** In this final stage, the similarities are computed among entities in the reduced-dimensional space, rather than in the original term-document matrix. Since both documents and terms are represented as vectors in the same space, document-document, term-term, and term-document similarities are all straightforward to compute.

LSA has also been applied to many problems related to information retrieval, including text classification, text clustering, and link analysis. Some examples include grading essays (Foltz, 1996), intelligent tutoring systems (Graesser et al., 2000) and evaluating flight landings using a simulator (Quesada, 2007). LSA appears to be especially useful for problems where input is noisy (such as speech input) and standard lexical matching techniques fail. LSA has also been used in the cognitive sciences to model aspects of human memory and cognition (Landauer et al., 1998). However, LSA does not take into consideration word order, syntactic relations or logic, or morphology, and thus, this is seen as a disadvantage as it is not grounded in human perception and intention (Landauer et al., 1998).

2.5.2 - STASIS

STASIS is a corpus-based similarity measure that measures the level of similarity between two utterances using an ontological approach based on a taxonomy of words (Li et al., 2003). STASIS calculates the distance between words in an ontology, using WordNet (Miller, 1995), a large lexical database that contains ontological relations between large numbers of entities as well as the distance of words to their closest subsumer.

Unlike LSA that uses SVD models, STASIS combines semantic and word order similarity taken from the words in the two utterances presented and WordNet is used to calculate the semantic similarity component. Information Content (IC) is taken from the Brown's Corpus

(Francis and Kucera, 1979) to calculate the semantic vector. STASIS also takes into account word order when calculating similarity, something that LSA does not do. STASIS was tested against the standard dataset created by Rubenstein and Goodenough (Rubenstein and Goodenough, 1965), and results showed a high correlation with human ratings (Li et al., 2003). O'Shea (J. O'Shea et al., 2013) created what is now referred to as gold standard datasets STSS-65 and STSS-131. These two datasets were first run with STASIS, and the correlation with human ratings was compared to LSA (J. O'Shea et al., 2013). Results showed that STASIS gave a higher correlation to human ratings than LSA. This is due to the ontological method used in STASIS, which successfully represents the inter-relatedness between a wide variety of words. This method gives the advantage over LSA which solely relies on corpus statistics to calculate similarity (Li et al., 2003). STASIS does not cater for human perception-based words (fuzzy words) such as *hot*, *cold*, *tall* or *short*, which are often significantly used in human dialogue.

2.5.3 - SEMILAR

SEMILAR, the SEMantic simILARity toolkit (Rus et al., 2013a), uses the word-to-word semantic similarity measures in the WordNet Similarity library (Patwardhan et al., 2003), as well as LSA (Landauer et al., 1998). SEMILAR uses two annotation protocols: greedy and optimal annotation. The greedy methods, pair a target word in one sentence with all the words in the other sentence and retains the matching word with the highest word-to-word similarity score to the target word, regardless of how other words match each other. The optimal matching strategy, is inspired from optimal matching methods, proposed for tasks where a set of items must be matched against another set, while optimizing the overall matching score and not individual scores.

While in greedy matching, the goal is for a target word to find a best matching word in the opposite sentence, in optimal matching, the goal is to match items such that an overall optimal matching is achieved (Rus et al., 2012). SEMILAR does not capture human perception-based words (fuzzy words) within short texts. SEMILAR was used to investigate and tune assessment algorithms for evaluating students' natural language input based on data from the DeepTutor computer tutor (Rus et al., 2013b). SEMILAR was also tested on several datasets to help with paraphrasing, entailment, and elaboration (Pedersen et al., 2004).

2.5.4 - Dandelion Semantic and Syntactic

Dandelion API is a commercial sentence similarity measure that uses a knowledge-based approach for short sentences between 5-20 words (SpazioDati, 2015) giving a rating of the similarity. Dandelion is a short sentence similarity measure which compares the semantic and syntactic similarity between two sentences and shows the semantic and syntactic results separately (Vysotska et al., 2019). The similarity rating provided by Dandelion API is displayed as two separate measures for semantic similarity and syntactic similarity and Dandelion API does not combine the two to give an overall sentence similarity. Dandelion API currently supports 7 languages (English, Italian, French, German, Portuguese, Spanish and Russian) (Lytvyn et al., 2019) but it does not cater for low resource languages such as Arabic, Urdu or Farsi.

Dandelion API has successfully been used for supervised multimodal search re-ranking technique using visual semantics (Bhuvan and Elayidom, 2020) which focused on webpage ranking and how the multi-media in a given page can impact its ranking. Results showed an improvement of the re-ranking of webpages using the proposed technique compared to state of the art retrieval models. Dandelion API has also been used as part of a knowledge-based system for automated assessment of short structured questions (Luchoomun et al., 2019). It focused on assessment in education, primarily comparing the meaning of two sentences, by taking into account both semantic and syntactic content and the corresponding grade to be returned. A dictionary was created for comparison of keywords from answers and the system was also given different variations of answers to allow it to learn. The system was tested with ten short, structured questions answered by 5 students, and the answers were marked by both the system and a tutor (manually). The system showed an improvement over results marked manually (Luchoomun et al., 2019).

2.6 - Comparison of Semantic Similarity Measures

The four SSM's identified in Section 2.5 (LSA, STASIS, SEMILAR and Dandelion API) use a variety of approaches, from ontological to corpus-based to knowledge-based often using datasets to evaluate their performance compared to correlations with human ratings. STASIS and LSA were tested on two gold standard datasets (STSS-65 and STSS-131) to compare correlation of results with human ratings (J. O'Shea et al., 2013). One weakness that was common in all four algorithms was the lack of acknowledgment for fuzzy words or utterances.

In this research, a fuzzy sentence or utterance is defined as a short text, which comprises of at least one fuzzy word. A fuzzy word is a word that has a subjective meaning and is characteristically used in everyday human natural language dialogue. Fuzzy words are often ambiguous in meaning, since they are based on an individual's perception (Adel et al., 2018).

Furthermore, Dandelion API is the only SSM from the mentioned above that is used on a commercial scale. LSA does not use a lexical approach and does not take syntactic or word order into consideration and is designed to work on paragraphs of text rather than sentences or utterances (10-20 words in length as defined in Section 2.4). While STASIS and SEMILAR do use a lexical approach and take into account semantic and syntactic similarities, STASIS relies on WordNet and Brown's Corpus and SEMILAR relies solely on WordNet for similarity measures. Therefore, it can be concluded that none of the SSM's consider the presence of fuzzy words.

2.7 - Problems and Challenges

Some of the challenges that arise from using a semantic similarity measure is to ensure the correct evaluation method is used to evaluate any results from experiments conducted. The proposed research in this thesis is the development and evaluation of a SSM that needs to consider the presence of fuzzy words (FSSM), due to the inability of the aforementioned SSM's in Section 2.6 to deal with the presence of perception-based words (fuzzy words) in sentences or utterances. A further challenge will be the use of human participants in any proposed experiments, as part of the development and evaluation of the proposed FSSM. In order to collect human ratings of word or sentence pairs, an approach which is often used in the semantic similarity community (J. O'Shea et al., 2013), human participants will have to be recruited which in itself is challenging as any participants recruited must present a valid sample to minimise any potential bias. Other challenges include the number of participants needed to make the experiment statistically significant. Along with the challenges mentioned, the participants must be native to the demographic region to improve accuracy. This section will examine some of the problems and challenges faced with the design and evaluation of a SSM and the collection of human ratings.

2.7.1 - Word Sense Disambiguation

For machines to understand the specific meaning of a word, they must process unstructured textual information and transform them into data structures to determine the underlying meaning. For example, take the two sentences *I work at the power plant* and *this plant needs watering*; both these sentences have the word *plant* in them, in the first sentence *plant* refers to an *industrial building* whereas in the second sentence *plant* refers to a *living organism*. This computational identification of meaning for words in context is referred to as Word Sense Disambiguation (WSD) (Navigli, 2009). WSD is the ability to computationally determine which sense of a word is activated by its use in a particular context (Navigli, 2009). Nevertheless, the manual creation of knowledge resources is an expensive and time-consuming effort (Ng, 1997), which must be repeated every time the disambiguation scenario changes. Other factors which can impact the creation of knowledge resources are human perception of words, which is further enhanced when trying to cover low resource languages such as Urdu (Kaleem, 2015) or Arabic (Aljameel et al., 2017).

2.7.2 - Word Embedding Models

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation, such as *happy* and *cheerful* both of which can be used to describe mood positively. Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network (Brownlee, 2017). Word2Vec (Mikolov et al., 2013) is a statistical method used for word embedding from a text corpus. Word Embedding Models (WEM) use artificial neural networks to learn a distributed representation of word co-occurrence information from a large corpus (Bengio et al., 2000), and have shown improvements in the performance of many Natural Language Processing (NLP) applications such as text classification (Bengio et al., 2000) and sentiment analysis (Tang et al., 2014). One of the drawbacks of using WEM's is that words which have seemingly opposite meanings may have a high score (for instance, *satisfied* and *unsatisfied*). This is because they are frequently used in similar contexts, therefore they will have similar vectors. Consequently, this means WEM’s are not well suited for sentence similarity measures in terms of considering the semantic relationships between words (Little et al., 2020). Therefore, this approach was not used in the research presented in this thesis.

2.7.3 - Collecting Human Ratings

Traditionally, word and sentence semantic similarity measures are evaluated through collecting human ratings in controlled experiments. O’Shea developed a methodology where sentence pairs where taken (J. O’Shea et al., 2013) and 64 participants were asked to assess their similarity on a scale of [0-4]. In order to ensure samples collected from participants will be valid for use in any experiment, the sample size is also important. O’Shea stated that 32 participants constitute a sample size which will provide statistically significant results (J. O’Shea et al., 2013).

Since the words being rated in the proposed research of this thesis are fuzzy words in English, it was important that participants were native English speakers, this was to remove the risk of a participant having a hugely different notion of a meaning of a fuzzy word based on English being a second language (Chandran, 2013). Although age, education and gender were not deciding factors in accepting a participant, all participants were asked to be over the age of 16. It creates a good balance of results when participants are from a good mix of genders, age groups and education backgrounds. Although demographic location was not an entry factor, all participants in the proposed research for this thesis were from the Northwest region of England, United Kingdom.

To create his two gold standard datasets (STSS-65 and STSS-131) O’Shea used a sample size of 32 participants (J. O’Shea et al., 2013) and thus this research will also focus on using 32 participants were possible, to collect valid sample ratings for experiments. In the proposed research for this thesis, a valid sample rating refers to results that are correct and effective based on the requirements for the experiment, it does not mean removing results that may be deemed as inconvenient or difficult to assess. 32 participants where needed for each experiment requiring human participants to constitute a sample size which will provide statistically significant results (J. O’Shea et al., 2013).

Participants for the proposed research in this thesis were asked to rate words on a scale of [0, 10], based on the same scale used in Wordsim-353 datasets (Finkelstein et al., 2001). They were allowed to go up to one decimal place for more accuracy. O'Shea did use a scale of [0, 4] (J. O'Shea et al., 2013) but this research will be using a scale of [0, 10] as this offers the participants a wider scale and thus more flexibility in rating the words.

2.8 - Applications of Sentence Similarity Measures

The are several applications where sentence similarity can be used such as text mining, question answering, and text summarization (Achananuparp et al., 2008). Given a pair of sentences an effective sentence similarity measure should be able to determine whether the sentences are semantically similar or not, taking into account the variability of natural language expression. Some examples of the applications of SSM's will now be described.

RECAP developed by Metzler et al. was used to develop methods for tracking and analysing the flow of facts and concepts through a text corpus (Metzler et al., 2005). In this research, Metzler used SSM's to identify passages or sentences that shared concepts and facts. This research presented a range of outcomes, such as methods for reuse detection, which measures the quality of reuse detection. This method demonstrates that reuse detection is both meaningful and feasible (Metzler et al., 2005). A vital factor to note from this research is that a scoring technique that can effectively identify similar documents at one threshold, but this similarity might not be effective for a different similarity threshold. Therefore, it is important to consider an appropriate similarity threshold based on the application, in order to achieve the best optimal scoring technique (Metzler et al., 2005).

Metzler et al. carry out further investigation into the effectiveness of lexical matching, language model, and hybrid measures, in order to compute the similarity between two short queries (Metzler et al., 2007). In his research, Metzler investigated the measure of the similarity between short segments of text from an information retrieval perspective, by looking at various types of text representations, including surface, stemmed, and expanded. There were several similarity measures used, including lexical matching and probabilistic measures, based on the language models estimated from unexpanded and expanded representations. Results showed that lexical matching is good in finding semantically identical matches and the probabilistic methods are better at finding interesting topics related matches (Metzler et al., 2007).

Balasubramanian et al. compared the performance of nine separate retrieval techniques in their research, where they looked at sentence retrieval techniques based on some measure of similarity to a query (Balasubramanian et al., 2007). Results showed that the performance of sentence retrieval is dependent on the retrieval technique used.

Liu and Wang proposed a means of calculating similarity between short texts and sentences without using an external corpus of literature and only relying on WordNet, common-sense knowledge base and human intuition (H. Liu and Wang, 2013). Results showed that this method provided a positive similarity measure and gave some degree of flexibility to allow the user to make comparisons without any additional dictionary or corpus information needed (H. Liu and Wang, 2013).

Spiccia et al. used a sentence similarity measure called Semantic Word Error Rate (SWER), which is based on the use of Levenshtein distance and LSA (Spiccia et al., 2016). This measure considers word order and the degree of similarity of words with different meanings. SWER was tested against Microsoft Research Paraphrase Corpus (MSR-PC) and results gave a higher accuracy when compared to LSA (Spiccia et al., 2016).

Measuring the semantic similarity between word or sentence pairs has long been one of the most challenging tasks in the field of natural language processing (J. Wang and Dong, 2020). This section aimed to highlight some of the applications of sentence similarity measures.

2.9 - Conclusion

This chapter discussed the notion of word and sentence similarity measures and how they are used and evaluated in the field of semantic similarity. A brief description was given of four existing published SSM's and how they compare with one another.

Some of the problems and challenges faced with sentence similarity were discussed including a brief explanation of word sense disambiguation, before moving into word embedding models and discussing the reasons why this approach was not used for this research. The problems and challenges associated with collecting human ratings as part of this research has also been highlighted. Finally, some applications of SSM’s have been given.

Chapter 3 will focus on reviewing the field of Fuzzy Natural Language Processing and the impact of Type-1 and Type-2 fuzzy sets on semantic similarity. It will also give an overview on the only known and published fuzzy sentence similarity measure FAST (Fuzzy Algorithm for Similarity Testing) and its limitations.

CHAPTER 3

CHAPTER 3: FUZZY NATURAL LANGUAGE PROCESSING

3.1 - Introduction

This chapter will introduce the concepts of fuzzy natural language processing in relation to the research presented in this thesis. It reviews the history behind computing with words before discussing the nature of fuzzy words and their potential applications.

The chapter will highlight related work with regards to the role of Type-1 and Type-2 fuzzy sets in the field of Computing with Words (CWW), introducing the concepts of footprints of uncertainty, for human perception-based words and how this plays a role in capturing the uncertainty of words.

The chapter will review the state of the art in fuzzy semantic similarity measures, which are limited, and identify the problems associated with the only published fuzzy sentence similarity measure. Finally, it will explore the challenges faced with modelling such measures using Type-1 and Type-2 sets and capturing the uncertainty of words. This chapter provides the background knowledge needed for the development of the new proposed fuzzy semantic similarity measure used in this research.

3.2 - Computing with Words

Computing with words (CWW) first originated with Zadeh's 1996 article in which he states: "*CWW is a methodology in which the objects of computation are words and propositions drawn from a natural language*" (Zadeh, 1996). In CWW, words are modelled using fuzzy sets (FS). Zadeh calls this "*precisiation*", and Mendel calls this "*encoding*" (Zadeh, 1996; Hao and Mendel, 2015). In essence, CWW is a methodology for reasoning, computing and decision-making with information described in natural language. The three main principles to CWW according to Zadeh are as follows (J. M. Mendel et al., 2010):

1. Much of human knowledge is described in natural language.
2. Words are less precise than numbers; words are used when the number is not known.
3. Precision carries a cost. If there is a tolerance for imprecision, it can be exploited using words in place of numbers.

Mendel et al. further proposes the following guidelines which they state must be passed in order for an application or system to be called CWW (J. M. Mendel et al., 2010):

1. A word must lead to a membership function, rather than a membership function leading to a word.
2. Numbers alone may not activate the CWW engine.
3. The output from CWW must be at least a word and not just a number.

He also suggests a fourth option:

4. Because words mean different things to different people, they should be modelled using at least Interval Type-2 Fuzzy Sets.

CWW relates to developing intelligent systems that are able to receive as input, words, perceptions, and propositions drawn from natural language, which can then produce a decision or output based on these words. CWW involves different needed components which are as follows (J. M. Mendel et al., 2010):

- Developing the mechanisms that can handle the uncertainties existing with natural language.
- Dealing with problems associated with change to the meaning of words according to context, region and culture.
- Developing reasoning mechanisms that can deal with words, perceptions, and propositions and result in words that address a problem in a similar way a human would address it.

CWW becomes a necessary tool when the available information is perception based or not precise enough to use numbers, referred to as Perceptual Computing (Per-C) (Gupta and Muhuri, 2019). Per-C is the case of most real-world applications involving humans. CWW adds to conventional modes of computing, the capability to compute with interpreted words, and propositions drawn from natural language (J. M. Mendel et al., 2010).

Gupta and Muhuri (Gupta and Muhuri, 2018) uses CWW in heart monitoring through perceptual computing, in order to assess the medical condition of a person suffering from heart failure. This research processes user feedback in terms of 'words' and generates recommendations about the medical attention needed to be given to the patient. Gupta and Muhuri (Gupta and Muhuri, 2019) further make use of CWW to conduct a student strategy evaluation and compare the different CWW approaches. Tešić et al used a CWW-based inference engine with a Deep Convolutional Neural Networks (DCNN) to detect and provide alerts concerning several maritime activities, including boat circling (Tešić et al., 2020). Rahmanian et al. (Rahmanian et al., 2021) propose a novel peer online assessment method for oral presentation using Per-C, where the numerical score for the overall assessment of a student in the presentation is compared and ranked using linguistic evaluation. Srivastava and Mondal (Srivastava and Mondal, 2022) have designed and developed an intelligent information system that runs on the knowledge base of Hesitant Fuzzy Weighting Linguistic Term Set (HFWLTS) for computing with words. This system had also been developed with reference to decision-making phenomena, in order to provide a new platform for decision-makers to conclude their decisions, with proper scaling of linguistic terms and weighting criteria.

3.3 - Similarity & Compatibility in Fuzzy Set Theory: Assessment & Applications

The first step in using fuzzy logic for CWW is to construct fuzzy sets to model words. Due to the linguistic uncertainty surrounding CWW, Mendel et al. stated using a Type-1 set to model a word is scientifically incorrect (J. M. Mendel et al., 2010), because a word is uncertain whereas a Type-1 set is certain. Hence, Mendel concludes that one should use Interval Type-2 fuzzy models to model first-order word uncertainties (Bilgin et al., 2012).

Zadeh claims that fuzzy logic, equates to CWW. In CWW, numbers are replaced with words not only when reasoning, but also when solving calculations. For example, *Temperature* is a linguistic variable, if its values are linguistic rather than numerical, i.e., [*hot*, *not hot*, *very hot*, *quite hot*, *cold*, *very cold*, *not very cold*, etc], as opposed to [18, 19, 20, 21, …] Celsius. Zadeh's examples use fuzzy granules to model words. A fuzzy granule is actually the footprint of uncertainty (FOU) of an Interval Type-2 FS (John and Coupland, 2006). According to Mendel, Type-2 FS are more difficult to use and understand than Type-1 (J. M. Mendel et al., 2010). However, Type-2 FS allow the effects of uncertainties in rule-based fuzzy logic to be modelled and minimised (J. M. Mendel and John, 2002). Mendel and John state that there are at least four sources of uncertainties in Type-1 sets (J. M. Mendel and John, 2002). These are as follows:

(i) The meaning of the words used in the antecedents and consequents of rules can be uncertain. Words mean different things to different people, e.g., *sick* to some people means poorly, while *sick* is also slang for cool, awesome or amazing.

(ii) Consequents may have a histogram of values associated with them. Therefore, when rules are obtained from a group of experts, consequents will often be different for the same rule, i.e., the experts will not necessarily be in agreement.

(iii) Measurements that activate a Type-1 set may be noisy and therefore uncertain; because when those parameters are optimised using uncertain (noisy) training data, the parameters become uncertain.

(iv) The data that is used to tune the parameters of a Type-1 set may also be noisy; because very often it is such measurements that activate the fuzzy logic system.

When many people rate words in the context of similarity, it is still the subjective opinion of those individuals as to whether a word belongs to a particular set or not. When gathering similarity for ratings, a group of people (participants) tends to be used to make the data collection statistically viable and reflective of population sample; all this creates gaps and noise which refers to data that may not be accurate or incorrect. Traditionally Type-1 sets were used when gathering human ratings; however as stated by Mendel and John Type-1 sets are not able to directly model such uncertainties (J. M. Mendel and John, 2002), because their membership functions are totally crisp and two dimensional; and it is for this reason that Type-2 FS are able to model such uncertainties, because their membership functions are fuzzy and three dimensional (J. M. Mendel and John, 2002). By being three-dimensional, Type-2 FS provide additional degrees of freedom that make it possible to directly model uncertainties. However, according to Mendel and John, this also makes Type-2 FS difficult to understand and use, because (J. M. Mendel and John, 2002):

1. The three-dimensional nature of Type-2 FS makes them very difficult to draw.
2. There is no simple collection of well-defined mathematically precise terms that allow effective communication regarding Type-2 FS.
3. Derivations of the formulas for the *Union*, *Intersection*, and *Complement* of Type-2 FS all rely on using Zadeh's Extension Principle, which itself is a difficult concept to understand.
4. Using Type-2 FS is computationally more complicated than using Type-1 FS.

Due to the difficulty of Type-2 FS, Mendel et al. further proposed the use of Interval Type-2 FS (J. M. Mendel et al., 2006) which will be fully explained in Section 3.4.

3.4 - The Role of Type-1 and Type-2 Fuzzy Sets

The difference between Type-1 and Type-2 FS can best be described using graphical representation. Figure 5 (Source: [adapted from] J. M. Mendel and John, 2002) shows an example of Type-1 set (*A*), where the membership function ($\mu_A(x)$) is shown in red, where *X* is the universe of discourse, and its elements are denoted by (*x*). The x-axis represents the domain of the fuzzy set *A*, and the y-axis represents the membership function (μ_A) (J. M. Mendel and John, 2002).

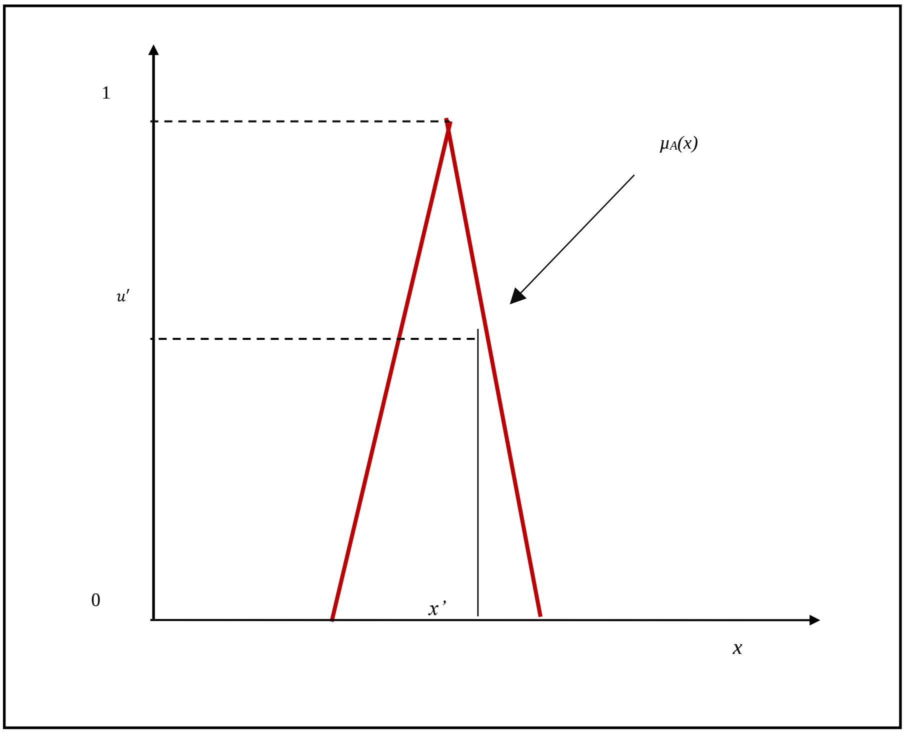


Figure 5 - Type-1 Membership Function (Source: [adapted from] J. M. Mendel and John, 2002)

If the red line was to be blurred to the left and right, like Figure 6 (Source: [adapted from] J. M. Mendel and John, 2002), then a Type-2 membership function is produced. Thus, for a specific value (*x'*), the membership function (μ_A), takes on different values (*u'*), which are indicated by the blurred area in Figure 6, which are not all weighted the same, thus membership grades can be assigned for all the points individually (J. M. Mendel and John, 2002).

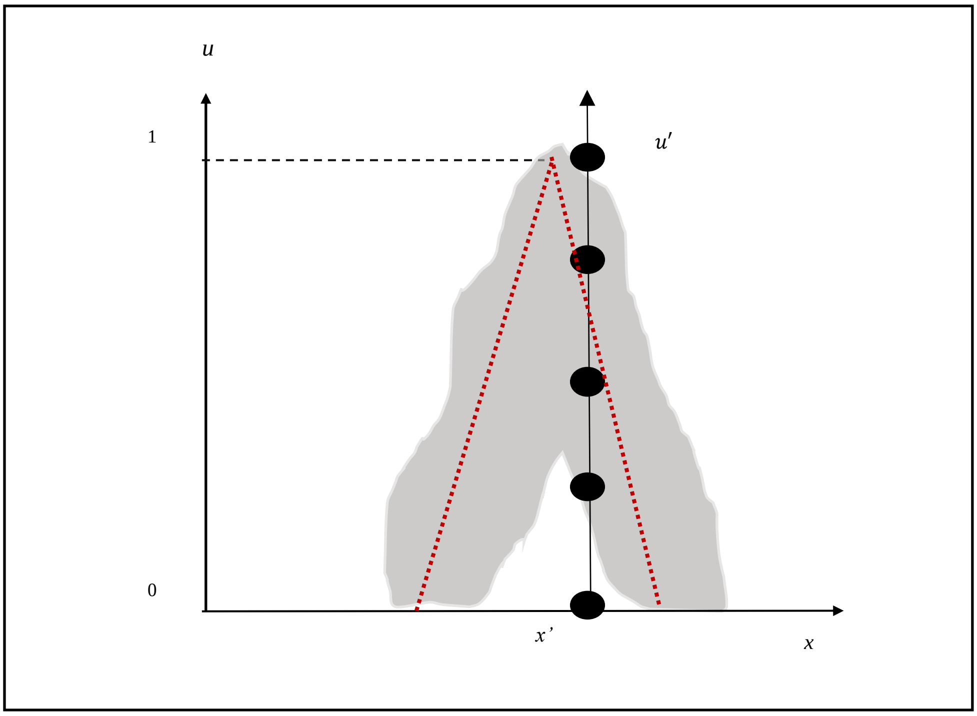


Figure 6 - Blurred Type-1 Membership Function (Source: [adapted from] J. M. Mendel and John, 2002)

By doing this for all ($x \in X$), a three dimensional Type-2 membership function is created, as shown in Figure 7 (Source: [adapted from] J. M. Mendel and John, 2002) which characterises a Type-2 fuzzy set. A normal Type-2 fuzzy set is three dimensional, where the third dimension is the value of the membership function, which is referred to as the *Footprint Of Uncertainty* (FOU) (Zadeh, 1975a). The shaded areas in Figure 7 represents the FOU, $\mu_{\tilde{A}}$ (x, u) for (x) and (u) discrete , where X = [1, 2, 3, 4, 5] and U = [0, 0.2, 0.4, 0.6, 0.8] (J. M. Mendel and John, 2002), where (X) is the primary domain, (J_x) is the secondary domain, ($\mu_{\tilde{A}}(x)$) is the secondary membership function at (x) and all secondary grades ($\mu_{\tilde{A}}$ (x, u)) $\in [0, 1]$.

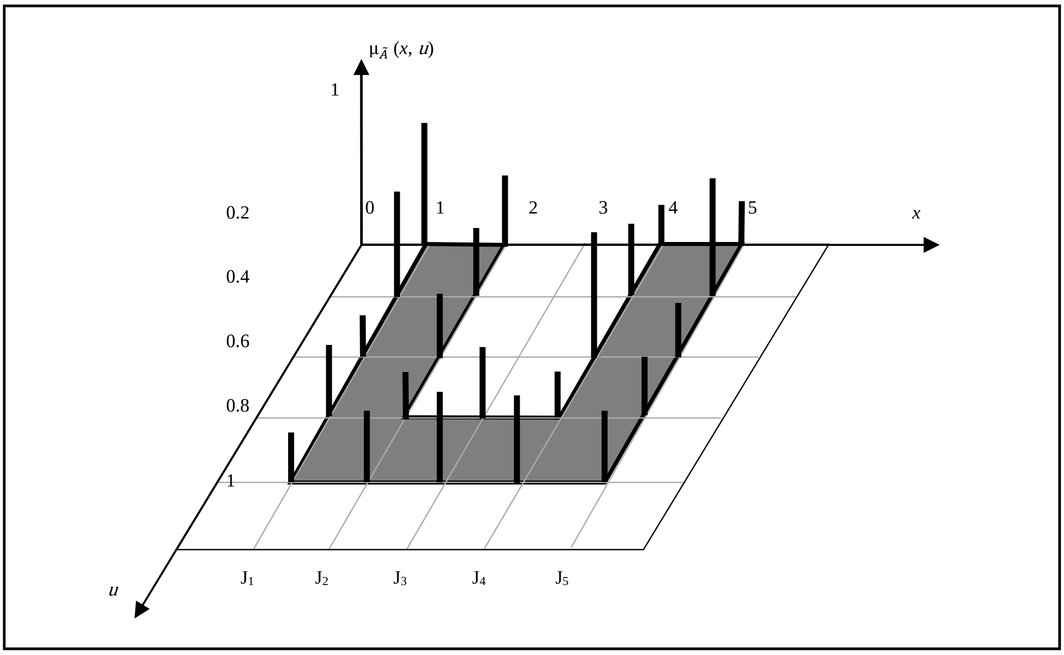


Figure 7 - 3D Example of Type-2 Membership Function (Source: [adapted from] J. M. Mendel and John, 2002)

A Type-2 FS ($\tilde{A}$)*,* is characterised by a Type-2 membership function $\mu_{\tilde{A}}(x, u)$ where (*x* ∈ *X*), and $u \in J_x \subseteq [0, 1]\}$ represents the primary membership of (*x*) such that:

$$\tilde{A} = \left\{ \left((x, u), \mu_{\tilde{A}}(x, u) \right) \middle| \, \forall_x \in X, \forall_\mu \in J_x \subseteq [0, 1] \right\}$$

Equation 3 (Source: J. M. Mendel and John, 2002)

Where, $0 \leq \mu_{\tilde{A}}(x, u) \leq 1$. Therefore, a Type-2 membership grade can be any subset in [0, 1], the primary membership, and corresponding to each primary membership there is a secondary membership, that defines the possibilities for the primary membership (Castillo and Melin, 2012; J. M. Mendel and John, 2002).

Since Type-2 FS are very complicated to use, this caused Interval Type-2 FS to be created. General Type-2 FS are computationally intensive because type-reduction is very intensive. Therefore, this can be simplified when the secondary membership functions (MFs) are interval sets (in this case, the secondary memberships are either zero or one) and this is refered to as Interval Type-2 FS and the recommended approach to be used by Mendel (J. M. Mendel et al., 2006) is denoted as:

$$\tilde{A} = \{((x, u), 1) |\ x\ \in X, u \in\ J_x, J_x \subseteq [0, 1]\}$$

Equation 4 (Source: J. M. Mendel et al., 2006)

Where (X) is the primary domain, (J_x) is the secondary domain, and all secondary grades ($\mu_{\tilde{A}}\ (x, u)$) is equal to 1.

General Type-2 FS are very similar to Type-1 sets, the major difference being the defuzzification process in Type-1 which is replaced by an output process in Type-2. This output process initially has type-reduction, followed by defuzzification. Type-reduction maps a Type-2 FS into a Type-1 set, and then the defuzzification maps that Type-1 set into a crisp number. This can be easily explained using Mendels's Fuzzy Logic System (FLS) diagram shown in Figure 8 (Source: [adapted from] J. M. Mendel, 2017):

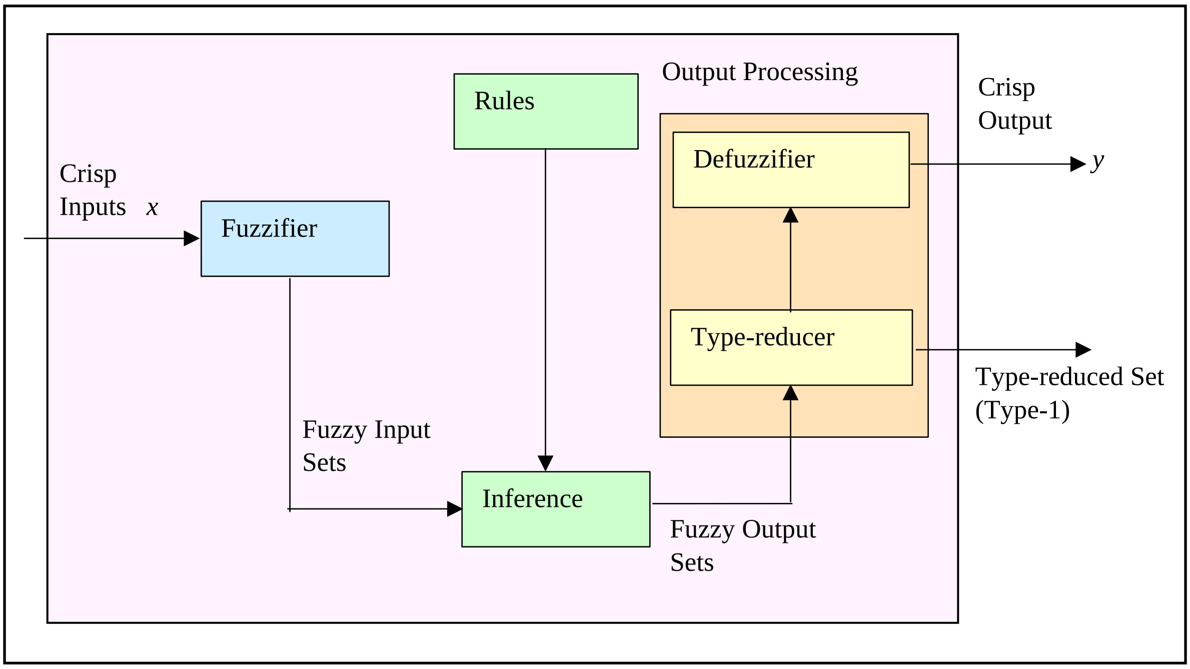


Figure 8 - Type-2 Fuzzy Logic System (Source: [adapted from] J. M. Mendel, 2017)

The FLS takes a crisp input and produces a crisp output. The FLS, maps crisp numbers into FSs. It activates rules that are in terms of linguistic variables (such as IF-THEN statements), which have FSs associated with them. The inputs to the FLS prior to fuzzification may be certain (e.g., perfect measurements) or uncertain (e.g., noisy measurements) (J. M. Mendel, 2007).

3.5 - Capturing the Uncertainty of Words

There are two well-known approaches in capturing the uncertainty of words according to Liu and Mendel (F. Liu and Mendel, 2008). These are the creation of a person membership function (MF) and the interval endpoint approach. In the person MF, a group of participants are asked to provide the FOU for a given word on a chosen scale. The FOU for each participant in this group captures what is known as the *intra-level* of uncertainty for that word, which is essentially the participants opinion of that word, this is explained further in Chapter 4. All of the uncertainties from the group of participants were collected and combined and fitted into an Interval Type-2 FS model. The main disadvantage of this approach is the participants in question must have knowledge of fuzzy sets, and this can limit the application of using this method. Furthermore, being able to collect participants in the first instance is also difficult, and a minimum of 32 participants is usually required to make the study statistically significant (J. O'Shea et al., 2013).

The interval endpoint approach asks a group of participants to give the two endpoints for a given word on a chosen scale. Once all endpoints were collected, the mean and standard deviation were calculated and mapped onto the Interval Type-2 FS model. The advantage this approach has over the person MF approach is participants do not require knowledge of FS. However, the disadvantage of this approach is closed-form mappings (the solution to a given problem in terms of functions and mathematical operations from a given generally accepted set (Karnik and Mendel, 2001)) are only available for symmetrical FOU's that are associated with data intervals whose two endpoint standard deviations are approximately equal. The actual interval endpoint data also shows that most words do not have equal endpoint standard deviations (F. Liu and Mendel, 2008); therefore, the shape of the FOU's must be chosen in advance.

Over time Mendel has introduced three main approaches to creating Type-2 fuzzy sets that were seen as evolutionary approaches to modelling words. These are the (i) Interval Approach (IA) (F. Liu and Mendel, 2008), the (ii) Enhanced Interval Approach (EIA) (D. Wu et al., 2011) and the (iii) Hao-Mendel Approach (HMA) (Hao and Mendel, 2015). Each will be explained briefly below.

(i) The Interval Approach (IA)

This approach uses the advantages of both the person MF and the interval endpoints approaches and is seen as one of the most important ways to model Interval Type-2 FS from data intervals (F. Liu and Mendel, 2008). It does this by collecting participants opinion of endpoints for chosen words without the prior FS knowledge needed and uses a simpler mapping approach for the FOU which does not require them to be symmetrical. The collected interval endpoints are mapped to a prespecified Type-1 person-membership function. The IA approach consists of two main parts known as the *Data Part* and the *Fuzzy Set (FS) part*. In the first part, the interval endpoints are pre-processed, and data statistics are computed. In the FS part, the computed data determines the model of the word which is either a left-shoulder, interior or right-shoulder FOU. The advantages of using the IA approach are that data collection from participants is easy to do, as they require no FS knowledge and the mapping of the FOU is straightforward. This means that it does not require prior assumption and it does not matter if the FOU is symmetrical or not. Its weakness, however, is that data is not adaptive and if more participant data is collected at a later date, the whole IA procedure must be repeated again. In this work (F. Liu and Mendel, 2008), the optimal number of participants required is not discussed and the researcher must select the group size based on their own discretion.

(ii) Enhanced Interval Approach (EIA)

To overcome the shortfalls of the IA approach, Mendel proposed the EIA approach. Similar to the IA approach, the EIA also consists of two parts, the data part and the fuzzy set part. The data part of the EIA approach has a more stringent approach, to overcome the limitations of IA and the FS part is more improved to compute the lower Type-1 membership functions. The Type-1 MF's are collected using the union to model the FOU. EIA also carries certain limitations such as the uniform distribution for a participant's data interval and lacks the measure to adjust the shape of the FS (only the trapezoidal FOU is discussed by Mendel) (Su et al., 2019).

(iii) Hao-Mendel Approach (HMA)

Following the limitations raised from both the IA and EIA approach, Mendel further introduced the HMA approach. Like the IA and EIA, HMA also has two distinguishing parts, the data part and the FS part. The data part of HMA is similar to that of EIA, however, there have been noticeable changes made to the FS part. The most notable difference with the FS, part to that of its predecessors, is the common overlap of participant data intervals is interpreted to indicate agreement by all of the subjects for that overlap, thus a membership grade of 1 is assigned to the common overlap (Hao and Mendel, 2015). The HMA also uses a more simplified approach to FOU than the EIA and requires fewer probability assumptions about the intervals than the IA or EIA approaches.

Taking into consideration the three proposed approaches, the method used for the research in this thesis is the HMA approach due to its simplified approach of managing the FOU's and needing less probability assumptions about the intervals than the IA or EIA approaches.

3.6 - Fuzzy Sentence Similarity Measures

As outlined in the introduction to this research (Chapter 1), a common problem in the field of semantic sentence similarity is the inability of semantic similarity measures (SSM) to accurately represent perception based (fuzzy) words that are commonly used in natural language. Prior to commencing this research, an algorithm known as FAST (Fuzzy Algorithm for Similarity Testing), which is the only known Fuzzy Semantic Similarity Measure (FSSM), that takes fuzzy words into account, was explored. FAST (Chandran et al., 2013) is an ontology-based similarity measure that uses concepts of fuzzy and CWW to allow for the accurate representation of fuzzy based words. Through human experimentation, fuzzy sets were created for six categories of words based on their levels of association with concepts using Type-1 fuzzy sets. These fuzzy sets were then defuzzified and the results used to create new ontological relations between the words. The next section will provide a brief overview of FAST.

3.6.1 - FAST (Fuzzy Algorithm for Similarity Testing)

As mentioned in the previous section, FAST (Chandran et al., 2013) is an ontology-based similarity measure that uses concepts of fuzzy and computing with words to allow for the representation of fuzzy based words. Non-fuzzy words are calculated using Wordnet (Miller, 1995). FAST is able to measure the effect that a limited number of fuzzy words in a short text have on the overall levels of semantic sentence similarity (Chandran et al., 2013). FAST was evaluated on two datasets, each containing 30 sentence pairs that used different fuzzy words but were similar in meaning. Existing sentence similarity datasets did not contain fuzzy words, thus in order to test FAST, Chandran created two fuzzy datasets, SWFD (Single Word Fuzzy Dataset) and MWFD (Multi Word Fuzzy Dataset) (Chandran, 2013).

The SWFD dataset contained 30 sentence pairs, and each sentence contained a fuzzy word taken from the 6 fuzzy categories of FAST. The origin of the 30 sentence pairs was the gold standard STSS-131 dataset (J. O'Shea et al., 2013). The MWFD dataset contained 30 sentence pairs, and each sentenced contained two or more fuzzy words taken from the 6 fuzzy categories of FAST. The sentence pairs in the MWFD dataset were extracted from fuzzy sentences in a large corpus and the fuzzy words in those sentences were replaced by suitable fuzzy words in one of the 6 fuzzy categories of FAST (Chandran et al., 2013). Human ratings were obtained for the sentence pairs in both SWFD and MWFD dataset and results compared to FAST, STASIS (Li et al., 2003) and LSA (Dumais, 2004). The results showed FAST had some improvement in sentence similarity compared to LSA and STASIS (that do not cater for fuzzy words) (Chandran et al., 2013). However, the use of Type-1 sets causes a weakness for FAST, since these words are not a true representation of each category. This is because fuzzy words are Type-2 and not Type-1 (J. M. Mendel, 2007); therefore, obtaining values using Type-1 sets adversely affects the accuracy of the words in each category by the method used to quantify fuzzy words. The FAST algorithm also has a very limited vocabulary of only 196 fuzzy words, which limits the coverage of fuzzy words in the English language. Furthermore, FAST does not cater for linguistic hedges (such as *very* or *slightly*) or any negation words or phrases (such as *not*), they are simply passed to WordNet and a measure for each word is obtained from WordNet.

3.7 - Challenges in Evaluating Fuzzy Semantic Similarity Measures

One of the most challenging tasks in creating a dataset to test any similarity measure is correctly representing the words and phrases typically used in conversation by humans. This also rings true in this instance, and one of the most important challenges in this research is creating datasets suitable for evaluating the proposed FSSM, FUSE (FUzzy Similarity mEasure). Evaluating traditional SSM is difficult as limited word and sentence datasets that have captured human ratings using sound methodological approaches exist. O'Shea highlights this point well and further states that even selecting random sentences is no guarantee of quality (J. O'Shea et al., 2013). When O'Shea created his gold standard datasets STSS-65 and STSS-131, sentences were carefully selected that would represent some property of the English language and have a diverse representation of grammatical, syntactic, and semantic properties of the English language. He also tried to choose utterances that would occur in everyday human communication, speaking and internet chats and forums (J. O'Shea et al., 2013). The gold standard datasets, STSS-65 and STSS-131, created by O'Shea (J. O'Shea et al., 2013) contain very few fuzzy words, therefore they can be used to test the FUSE algorithm to measure its performance when faced with datasets containing little to no fuzzy words.

The existing fuzzy datasets created by Chandran (Chandran, 2013), SWFD and MWFD, were limited in the number of fuzzy words they contained, as they were limited to a subset of fuzzy words used with the FAST algorithm.

Additionally, it presents certain degrees of challenges to be able to recruit 32 participants to fit the accepted sample size (J. O'Shea et al., 2013) for each rating, and recruitment must be done carefully to ensure participants are all native English speakers from the same region to avoid ratings being too far out from each other (J. O'Shea et al., 2013). Any datasets created for the evaluation of the FUSE algorithm must take into consideration these points mentioned, in order to have a good representation of results and also cover a wide range of fuzzy words used in day-to-day human conversation.

3.8 - Conclusion

This chapter covered the concepts of fuzzy natural language processing in relation to the research presented in this thesis. The role of CWW is also explained before discussing the nature of fuzzy words and their potential applications.

The chapter overviewed the theory of fuzzy sets and the role of Type-1 and Type-2 fuzzy sets in the field of CWW, and why Type-1 sets are not a good representation of fuzzy words. The footprint of uncertainty (FOU) is described, and the different approaches introduced for the analysis of the FOU, before capturing the uncertainty of words, to provide an overview of the published FSSM, FAST and its limitations.

Based on the findings in this chapter, Chapter 4 will discuss one of the key contributions to this research, the creation of the FUSE algorithm. This algorithm is an ontology-based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception-based words. The FUSE algorithm will be formulated using the elements of Mendel's work.

CHAPTER 4

CHAPTER 4: DEVELOPMENT OF FUSE_1.0

4.1 - Introduction

This chapter describes one of the major contributions of the work presented in this thesis - the creation of the FUSE (FUzzy Similarity mEasure) algorithm. FUSE is an ontology-based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception-based words. The aim of this chapter is to present the first prototype of FUSE referred to as FUSE_1.0 and highlight its key contributions.

This chapter will contribute towards answering the first research question:

RQ1. Investigate the feasibility of utilising Type-2 Fuzzy Sets and their representation of an individual's perception of fuzzy words and evaluate the suitability of the resulting fuzzy word models for incorporation into a Fuzzy Semantic Similarity Measure (FSSM).

This chapter utilises the research methodology presented in Chapter 1 to discuss the stages that were used to create the FUSE_1.0 algorithm. This includes the process of collecting human ratings for the words in the fuzzy dictionary and analyse the methodology behind the algorithm, before elaborating on the ethical decisions needed for the development of the algorithm.

The FUSE_1.0 algorithm will be developed in two core phases. Phase 1 will investigate and model fuzzy words using Interval Type-2 Fuzzy Sets, which will formulate a new fuzzy dictionary. Phase 2 takes the categories of fuzzy words created in Phase 1, to create a series of fuzzy word ontologies which are used in the calculation of fuzzy semantic similarity measures between two user utterances, for the development and implementation of the first version of the FUSE algorithm referred to as FUSE_1.0.

4.2 - Key Contributions of FUSE_1.0

The FUSE algorithm has been designed to model *intra-personal* (the uncertainty a person has about the word) and *inter-personal* (the uncertainty that a group of people have about the word) uncertainties, which are intrinsic to natural language. This is because the membership grade of an Interval Type-2 fuzzy set is an interval instead of a crisp number as in Type-1 sets (Adel et al., 2018).

The FUSE algorithm identifies fuzzy words in a human utterance and determines their similarity in the context of both the semantic and syntactic construction of the sentence. The main difference between the FUSE algorithm compared to FAST is:

- The FUSE algorithm contains a fuzzy dictionary which incorporates more fuzzy words than FAST.
- The FUSE algorithm was developed with a new fuzzy ontology that is able to deal with the words in the fuzzy dictionary.
- The FUSE algorithm contains defuzzified values modelled on Interval Type-2 membership, compared to Type-1 membership in FAST.

4.3 - FUSE_1.0 Overview

This section provides an overview of FUSE_1.0 architecture. Figure 9 is a component diagram of FUSE_1.0. It displays how two user utterances U_1 and U_2 are both fed into the FUSE algorithm and the overall sentence similarity rating that is achieved. In this work, a user utterance is defined as a a short text comprising of 25 words or less. All the words in U_1 and U_2 are placed into a bag of words before the similarity of the word-token pairs are calculated. If a word is present in the fuzzy dictionary of the FUSE algorithm, then the fuzzy rating for that word will be used to determine word similarity. This is only valid if the words per sentence pair belong to the same fuzzy category of the fuzzy dictionary. The fuzzy dictionary of the FUSE algorithm is a collection of fuzzy words split into categories, with deffuzified similarity ratings, using the Interval Type-2 FS approach (Section 4.4.4). However, if this is not the case, then it will use WordNet (Miller, 1995) to obtain the rating of the word similarity; if the word is also not in WordNet, then a similarity rating of 0 will be returned. From the word order being computed and finally using the fuzzy ontology created in the FUSE algorithm, the semantic and syntactic values are calculated before the algorithm gives a final sentence similarity rating between the two utterances U_1 and U_2 .

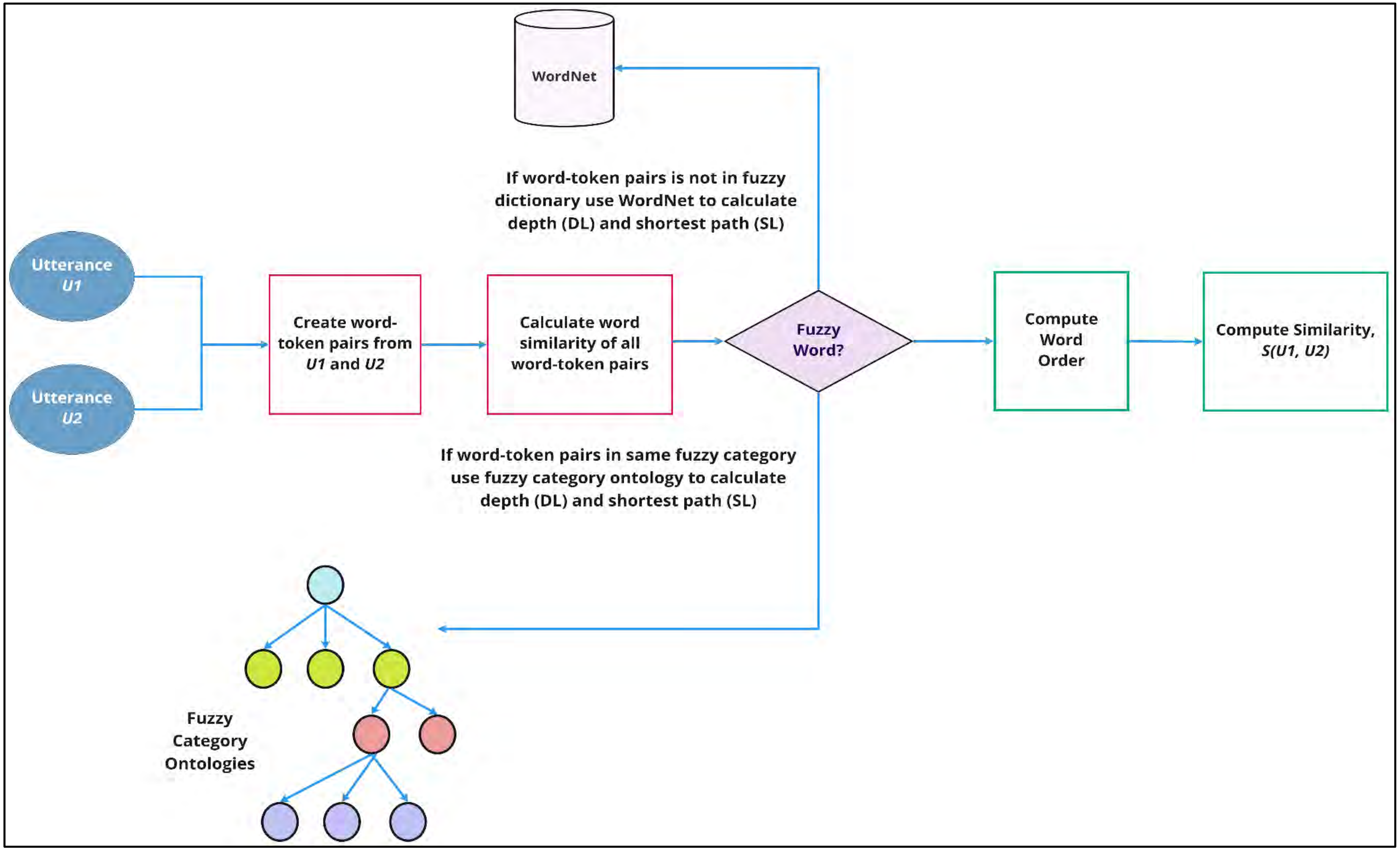

WordNet
If word-token pairs is not in fuzzy dictionary use WordNet to calculate depth (DL) and shortest path (SL)
Utterance *U1*
Utterance *U2*
Create word-token pairs from *U1* and *U2*
Calculate word similarity of all word-token pairs
Fuzzy Word?
Compute Word Order
Compute Similarity, *S(U1, U2)*
If word-token pairs in same fuzzy category use fuzzy category ontology to calculate depth (DL) and shortest path (SL)
Fuzzy Category Ontologies

Figure 9 - Component Diagram for FUSE

4.4 - Phase 1 of FUSE_1.0

In order to develop the FUSE_1.0 algorithm, six fuzzy categories were adapted from previous published work that utilised the Type-1 similarity measure known as FAST (Chandran, 2013). The category *Goodness* in FAST was renamed to *Worth* in FUSE_1.0. A key weakness of FAST, as mentioned in Section 3.6.1, was that coverage of fuzzy words in FAST was limited, with only 196 words in total for all six categories. Therefore, initial work was undertaken to increase the words in these six categories using a new methodology with Interval Type-2 FS. In order to expand the words in each category, the Oxford English Dictionary (Oxford English Dictionary, 2021) was used. All one-word synonyms for each existing word in the six categories of FAST was collected, an example of this is synonyms for '*Old*', one-word synonyms were identified such as '*Mature*', '*Aged*' and '*Senior*'; any two-word synonyms such as '*Grey-haired*' were disregarded. This initial process increased the total number of words in the six categories to 309 words, giving a 60.07% increase over FAST. Table 1 shows the number of words in the six categories of FUSE_1.0 and the percentage increase of words per category compared to FAST.

Categories	Words Per Category in FAST	Words Per Category in FUSE_1.0	Percentage Increase over FAST
Size/Distance	45	91	50.54%
Temperature	31	36	13.88%
Age	32	42	23.80%
Frequency	26	48	45.83%
Level of Membership	21	31	32.25%
Worth	41	61	32.78%

Table 1 - Word Expansion of FUSE_1.0

Initially, FUSE_1.0 started with six categories and this was later expanded to nine categories in Phase 3; this expansion is fully discussed in Chapter 6. Once the synonyms were collected, the next stage was to obtain human ratings of each word in the fuzzy categories of FUSE_1.0.

The original ratings for words in FAST had used the Type-1 approach and as Mendel and Wu have mentioned, this is an incorrect way of modelling fuzzy words, as Type-1 membership functions are totally crisp and two dimensional (J. Mendel and Wu, 2010) and thus not suited to model the uncertainty related to fuzzy words.

4.4.1 - Obtaining Human Ratings of Words

Once the synonyms had been collected as part of Phase 1, fuzzy ratings were required and were normalised on a scale of [-1, 1] for each word in the fuzzy dictionary. To obtain this rating, human participants needed to be used to collect ratings for each word in all six fuzzy categories to result in defuzzified values for each fuzzy word, e.g., the normalised defuzzied value of *Frozen* is now assigned as (-1). The methodology for collecting these ratings adopts the Hao-Mendel Approach (HMA) as discussed in Section 3.5, which takes a humans collective subjective ratings of words. This further models the respresentation using Interval Type-2 fuzzy sets which are then defuzzified to give a rating per word on a scale of [-1, 1]. The benefits of using the HMA approach is fully described in Section 3.5 and the methodology of how it was applied to obtain human ratings can be found in Section 4.4.3.

Word similarity experiments typically use a point rating scale with descriptions of the endpoints of the scale, for example, *no similarity of meaning* to *perfect synonymy* (J. O'Shea et al., 2013). O'Shea reported using a 4-point scale (between 0 and 4) (J. O'Shea et al., 2013) for his STSS-65 and STSS-131 datasets, similar to the scale used by Li for STASIS (Li et al., 2006), however O'Shea reported the restriction of this scale. Chandran used a 10-point scale (between 0 and 10) to measure the ratings of words for FAST (F. Liu and Mendel, 2008; Chandran, 2013) and Mendel also used a 10-point scale (between 0 and 10) to determine the Footprint of Uncertainty (FOU) for his codebook (F. Liu and Mendel, 2008; Chandran, 2013). Thus, it was decided to use a 10-point scale of (between 0 and 10), 0 being *no similarity of meaning* and 10 being *perfect synonymy* when asking the human participants to rate each word. An accuracy of 1 decimal place was also allowed to offer further flexibility. 32 participants were needed for each category of words to be able to collect a sufficient sample size to allow the results to be statistically significant as reported by O'Shea (J. O'Shea et al., 2013).

To collect the human ratings for the words, native English speakers from the Northwest region of England, United Kingdom were used. This was to ensure that words did not have meanings that were too far apart, lessening the risk of distorting the results. O'Shea (J. O'Shea et al., 2013) further iterates that regional dialect might also interfere with the ratings of words given by participants in an experiment. This was also taken into consideration when collecting human ratings and participants taking part in this experiment were all native English speakers from the Northwest region of England, United Kingdom. Participants were recruited through poster campaigns, advertised on university campus, and posted on social media platforms. The requirements for the participants were discussed in Section 2.7.3.

To analyse the results, the statistical measure used to collect these ratings must also be considered. The Pearson's correlation coefficient (*r-value*), a long-established measure of agreement, is used in semantic similarity as a linear relationship between the two variables. This relationship will be compared and applied as the statistical measure in this work to evaluate FUSE_1.0 (Adel et al., 2018).

The Intra-Class Correlation Coefficient (ICC) in a study represents the extent to which the data collected in the study is correct and a good representation of the variables measured.

Cicchetti gives the following guidelines for the interpretation of the ICC, referred to as Inter-Rater Agreement measures, also known as the *a-value* (Cicchetti, 1994):

- *a-value* < 0.40 - Poor.
- 0.40 >= *a-value* <= 0.59 - Fair.
- 0.60 >= *a-value* <= 0.74 - Good.
- 0.75 >= *a-value* <= 1.00 - Excellent.

The *a-value* is an important factor, as it shows the extent of the data that is collected, and the representation of the variables measured. The aim is to achieve an E*xcellent* rating to maximise reliability of the human ratings of the short text pairs, with the similarity rating returned by the FUSE_1.0 algorithm (Adel et al., 2021).

A statistical test (*p-value*) is a way to evaluate the evidence from the data provided against a hypothesis (Ross, 2004). This hypothesis is called the *null hypothesis* and is often referred to as H_0. The (*p-value*) for each dataset shows if the hypothesis (H_0) can be accepted or rejected.

A (*p-value*) less than 0.05 (typically ≤ 0.05) is statistically significant for a confidence level of 95% and indicates strong evidence in support for the research hypothesis H_0.

4.4.2 - Ethical Considerations

In this phase of the experiments, a Participant Information Sheet and Consent Form were designed and given to each participant. No personal information is recorded from participants, and responses cannot be traced back to them. This phase of the research was given an ethical approval by Manchester Metropolitan Universities Science and Engineering Research Ethics and Governance Committee to proceed.

4.4.3 - Methodology for Human Ratings

The methodology used to obtain human ratings for words is based on the Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets (Hao and Mendel, 2015) as described in Section 3.5. In the HMA approach, 50 intervals were used to obtain the FOU for a word. To do this Hao and Mendel asked one participant to rate words on a scale of (*L*-*R*), (*L*) being left and (*R*) being right, giving the left (*x*,*y*) and right (*x*,*y*) endpoints on a given scale. Using this one rating, Mendel generated 100 random numbers (L_1, L_2, … , L_{50}; R_1, R_2, … , R_{50}) and used these to generate 50 endpoint interval pairs [(L_1, R_1), (L_2, R_2), … ,(L_{50}, R_{50})]. In the HMA approach, Hao and Mendel used one participant to obtain 50 intervals to reduce the time required to collect ratings.

The research proposed in this thesis, does not use the one-person approach, rather it uses 32 participants, to create a richer array of human results from 32 different people. In the research presented in this thesis, 32 participants were needed per category to provide ratings, and each category had in excess of 32 participants, so even after removing noise and outliers, each category was still left with feasible results from 32 participants.

To obtain the human ratings for each word in the six categories, questionnaires were set up for each category of words and participants were invited to take part. Each category had an introduction page giving them more information about the expectations of their results. An example is shown below taken from the category *Temperature*:

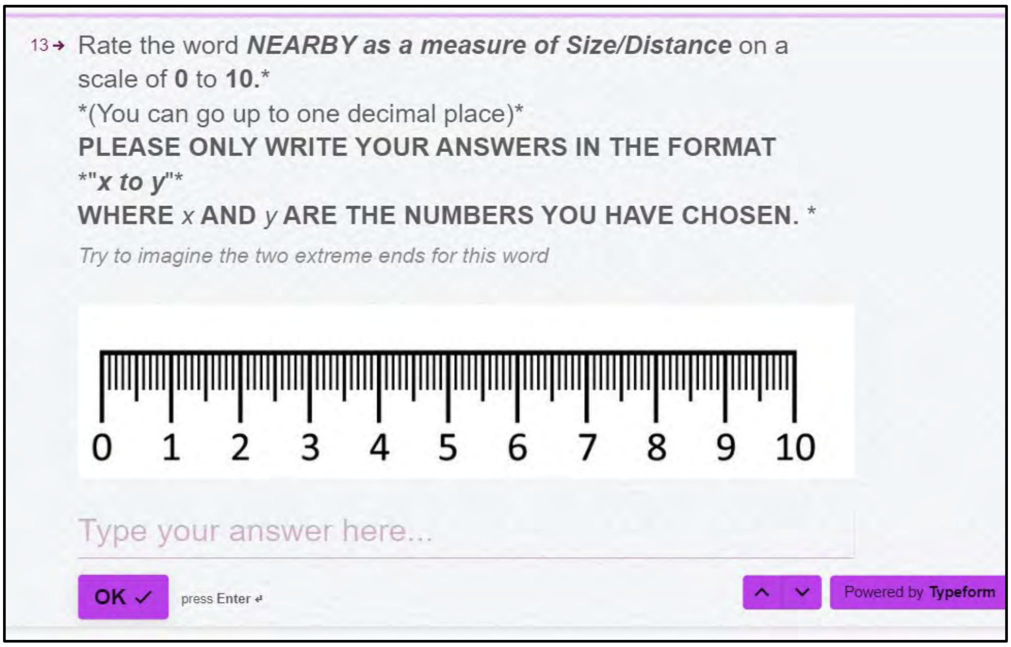
This experiment consists of 36 words belonging to the category TEMPERATURE.

I am going to give you a scale of 0 to 10. For each word that you are given, try to imagine the two extreme ends for this word. I want you to take this word and tell me where it would start, and where it would finish on this scale. You can use one decimal place (e.g., 3.2) for finer precision.

PLEASE ONLY WRITE YOUR ANSWERS IN THE FORMAT "x to y" WHERE x AND y ARE THE NUMBERS YOU HAVE CHOSEN.

For example, the word BABY which belongs to the category AGE. In my opinion I would say that on a scale of 0 to 10, Baby is between 1 to 1.5.

In order not to exhaust the participants and potentially affect the quality of the results, each participant was asked to only rate one category in one sitting. An example of how the question was presented per category is shown in Figure 10 along with a ruler image to give some visual representation of the scale of [0, 10] to ensure users understood what range, start point and end point meant.



13→ Rate the word ***NEARBY as a measure of Size/Distance*** on a scale of **0** to **10.***
(You can go up to one decimal place)
PLEASE ONLY WRITE YOUR ANSWERS IN THE FORMAT
*"***x to y***"*
WHERE *x* AND *y* ARE THE NUMBERS YOU HAVE CHOSEN. *
Try to imagine the two extreme ends for this word

Type your answer here...

Figure 10 - Example of Questionnaire with 0-10 Ruler Image for Categories

4.4.3.1 - Category Data Collection and Cleaning

Once all responses were collected the removal of noise could begin. Using Hao and Mendel's statistics and probability theory, the following steps below were adapted to remove noise (Hao and Mendel, 2015):

1. **Remove bad data** - in this step all nonsensical results were removed; in this case, it was any results that fell outside the [0, 10] range requested.
2. **Remove outliers** - using the Box and Whisker tests (Walpole et al., 1993), outliers are removed simultaneously from the results and the results were left with the data intervals that fell within an acceptable two-sided tolerance limit. Only the data intervals that are within an acceptable two-sided tolerance limit were kept. According to Hao and Mendel, a tolerance interval is a statistical interval within which, with some confidence level 100 (1- γ)%, confidence that the given limits contain at least the proportion (1-α) of the measurements (Hao and Mendel, 2015).
3. **Remove data intervals** that have no overlap or too little overlap with other data intervals. This is due to the fact that while Mendel stated *words mean different things to different people*, he had also argued that *words should mean similar things to different people* (F. Liu and Mendel, 2008). Therefore, if most participants rated a word between the intervals of [2-4] and a few rated the same word between the intervals of [6-7], then the latter would be considered no overlap with the other results and it will be removed (J. Mendel and Wu, 2010).

4.4.3.2 - Representing Category Words as Discrete FOU's

On completion of the three steps in Section 4.4.3.1, each category contained 32 clean data were *m* ≤ *n*, where *n* is the original data ranges collected by all participants and *m* is the data intervals after conducting the above three steps, where *m* = 32 clean value ranges per category for all six fuzzy categories.

The cleaned data was now ready for modeling; each category was analysed word by word. This was achieved by finding the upper FOU and lower FOU for each word and from this, the COG (Centre of Gravity) was calculated as defined in Equation 5*:*

$$COG = \frac{\left(\left(\frac{a+\ b}{2}\right) + \left(\frac{c+\ d}{2}\right)\right)}{2}$$

Equation 5

Where *a* = upper left FOU, *b* = lower left FOU, *c* = lower right FOU and *d* = upper right FOU.

A triangular norm (often referred to as *T-norm*) is a binary operation T : $[0, 1]^2$ → [0, 1] which is commutative, associative, non-decreasing in both variables and 1 is its neutral element (Mesiarová-Zemánková and Ahmad, 2010). In the field of mathematics, *T-norm* is respected as an operation in the interval [0, 1], which is always utilized in fuzzy logic. A *T-norm* can be extended to be a conjunction in fuzzy logic and an intersection in fuzzy set theory, such as a Minimum T-norm (*T-norm*$_{(min)}$) and a Product T-norm (*T-norm*$_{(prod)}$), both of which are involved in the operations on Type-2 fuzzy sets (Galindo, 2008).

Table 2 and Table 3 show defuzzified examples for the words '*Tiny*' and '*Gigantic*' respectively from the category '*Size/Distance*' on a scale of [0, 10]. The values are calculated using the triangular membership function. '*x*' is the scale of [0, 10], '*lower*' represents the lower boundaries, and '*upper*' represents the upper boundaries. '*T-norm*$_{(prod)}$' is the multiplication of lower and upper, and '*T-norm*$_{(min)}$' is the minimum boundary from the lower or upper.

Figure 11 and Figure 12 show the Type-1 defuzzified graphical representation of the word '*Tiny*' and the word '*Gigantic*' respectively in the category '*Size/Distance*' that has resulted from the triangular membership calculation. The values of 'T-*norm*$_{(min)}$' have been used to plot the graphs in Figure 11 and Figure 12 shown with the symbol × representing each datapoint on the graphs.

X	Lower	Upper	T-norm$_{(prod)}$	T-norm$_{(min)}$
0	0.00	0.00	0.00	0.00
1	0.24	0.38	0.09	0.24
2	0.71	0.76	0.54	0.71
3	0.83	0.85	0.70	0.83
4	0.37	0.47	0.17	0.37
5	0.00	0.08	0.00	0.00
6	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00

Table 2 - Scaled word for 'Tiny'

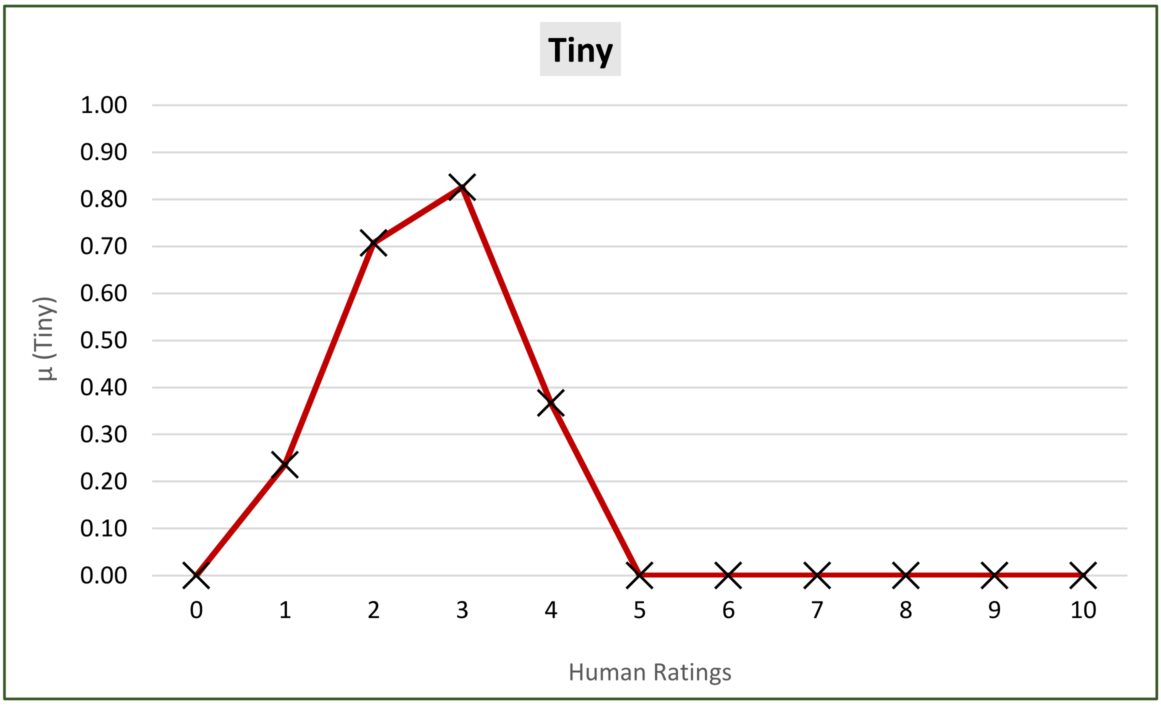


Figure 11 - Defuzzified figure for 'Tiny'

X	Lower	Upper	T-norm$_{(prod)}$	T-norm$_{(min)}$
0	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00
8	0.00	0.58	0.00	0.00
9	0.76	0.78	0.60	0.76
10	0.00	0.00	0.00	0.00

Table 3 - Scaled word for 'Gigantic'

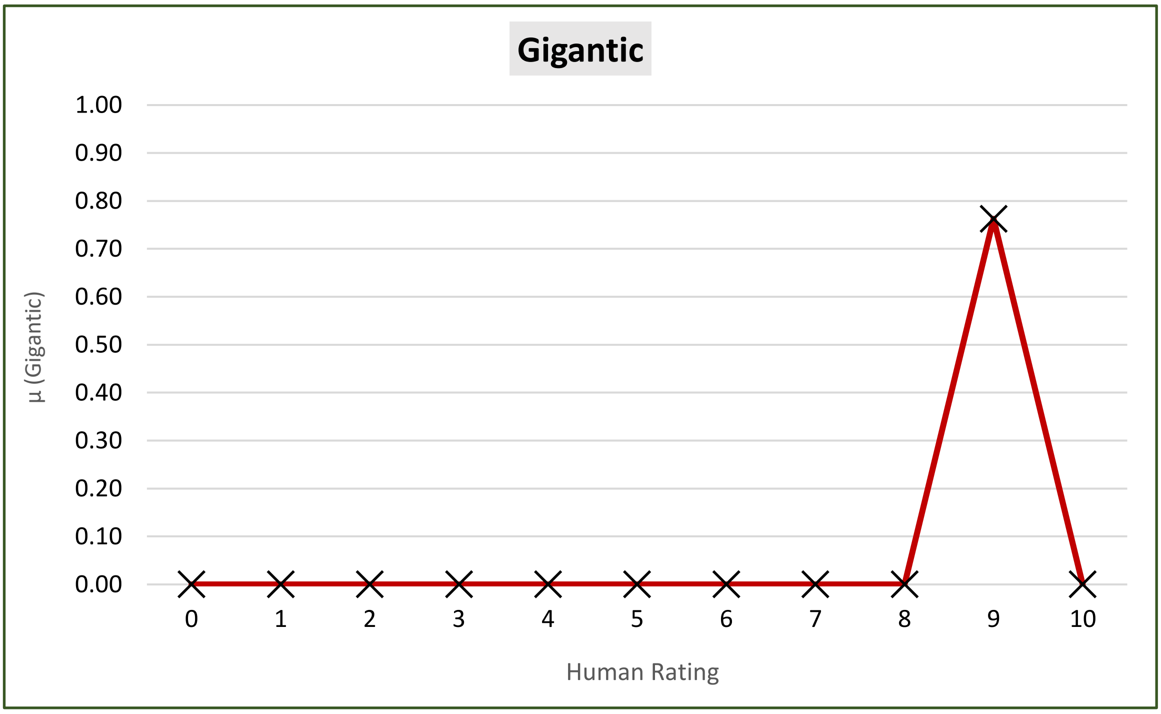


Figure 12 - Defuzzified figure for 'Gigantic'

All the results (*y*) were then normalised to the scale of [-1, 1] to ensure data can be utilized the same way across all the defuzzified values in each fuzzy category ontology as shown in Equation 6:

$$y = a + \frac{(x - A)(b - a)}{B - A}$$

Equation 6

Where *A* = smallest number in the proposed fuzzy category, *B* = largest number in the proposed fuzzy category, *a* = minimum normalised value (-1), *b* = maximum normalised value (+1) and *x* = value we want to scale (in this case the COG).

This now meant that every category contained words with values on a scale of [-1, 1]. This scale was selected to allow representation of defuzzified values in each fuzzy category ontology, which is required to obtain measurements in FUSE_1.0 which will be fully explained in Section 4.4.4.

The initial fuzzy dictionary created for FUSE_1.0 for all six categories containing the defuzzified values is shown in Tables 4 - 9. One of the aims of this research was to create a Fuzzy Dictionary in order to expand categories of words (Section 1.3). This was done partially in the first phase by expanding the number of fuzzy words in the six initial categories for FUSE_1.0, and this has also been expanded with a further three categories as part of the evolution of the FUSE algorithm (discussed in Chapter 6). The full fuzzy dictionary for the nine categories of the FUSE algorithm is available in Appendix C which can be used by other researchers in the field of NLP.

1 - SIZE/DISTANCE					
MICROSCOPIC	-1	ALONGSIDE	-0.27976	CONSIDERABLE	0.309524
MINUSCULE	-0.88095	ADJACENT	-0.26191	LOADS	0.333333
DINKY	-0.86905	ORDINARY	-0.22619	THICK	0.333333
TEENY	-0.85714	MEDIUM	-0.20238	FAR	0.363095
TITCHY	-0.7381	PROXIMATE	-0.20238	SIZEABLE	0.392857
LITTLE	-0.70833	EQUIDISTANT	-0.14286	LARGE	0.482143
SMALL	-0.70833	TIDY	-0.14286	PRINCELY	0.482143
WEE	-0.70833	USUAL	-0.1131	BOUNDLESS	0.535714

INSIGNIFICANT	-0.70238	AWAY	-0.10119	DISTANT	0.541667
PETITE	-0.64286	NORMAL	-0.10119	WHACKING	0.541667
DIMINUTIVE	-0.58333	PROXIMAL	-0.05357	SUBSTANTIAL	0.60119
NEAREST	-0.58333	REGULAR	-0.05357	BIG	0.660714
PIDDLING	-0.58333	STANDARD	-0.05357	GREAT	0.660714
TINY	-0.55952	BONNY	-0.02381	FARAWAY	0.666667
MINUTE	-0.55357	MEDIAL	0.011905	HEFTY	0.678571
SHORT	-0.52381	AVERAGE	0.029762	LONG	0.684211
UNIMPORTANT	-0.52381	MEAN	0.029762	JUMBO	0.720238
PALTRY	-0.51191	ACCESSIBLE	0.035714	EPIC	0.75
TRIVIAL	-0.5	HALFWAY	0.035714	MASSIVE	0.75
NEAR	-0.47619	ISOLATED	0.047619	OVERSIZED	0.754386
MESIAL	-0.44048	CENTRAL	0.065476	IMMENSE	0.754386
CONJOINING	-0.43452	GOODLY	0.065476	GIANT	0.809524
BESIDE	-0.41071	MIDWAY	0.065476	HUGE	0.827381
ADJOINING	-0.38095	MIDPOINT	0.066667	ENORMOUS	0.833333
THIN	-0.36364	CENTRE	0.066667	MEGA	0.839286
TOKEN	-0.35714	MEDIAN	0.083333	COLOSSUS	0.869048
NEARBY	-0.35119	MIDDLE	0.083333	GIGANTIC	0.892857
QUALITY	-0.35119	MID	0.089286	MAMMOTH	0.894
MOMENT	-0.32143	REMOTE	0.178571	GARGANTUAN	1
NORM	-0.29167	METHODICAL	0.184524		
CLOSE	-0.28571	ABUNDANT	0.214286		

Table 4 - Size/Distance Fuzzy Dictionary

2 – TEMPERATURE					
FROZEN	-1	BRACING	-0.31488	SPICY	0.550173
SUB-ZERO	-1	NIPPY	-0.28028	BAKING	0.619377
ARCTIC	-0.93772	TEPID	-0.24568	HOT	0.619377
FREEZING	-0.89619	MILD	-0.23875	SWEATY	0.688581

ICY	-0.7301	BODY-TEMPERATURE	0	SCALDING	0.750865
FROSTY	-0.70934	FRIGID	0.100346	HEATED	0.757785
CHILLY	-0.6955	BALMY	0.134948	STEAMING	0.757785
BRISK	-0.6263	TEMPERATE	0.204152	SWELTERING	0.792388
COLD	-0.57786	LUKEWARM	0.231834	ROASTING	0.861592
BITTER	-0.55709	WARM	0.480969	BOILING	0.889273
BITING	-0.45329	HUMID	0.550173	SCORCHING	0.930796
COOL	-0.45329	PERSPIRING	0.550173	BURNING	1

Table 5 - Temperature Fuzzy Dictionary

3 - AGE					
BABY	-1	IMMATURE	-0.333333	OLDER	0.789855
NEW	-0.963768	CHILDLIKE	-0.33333	EXPERIENCED	0.8260869
LATEST	-0.93939	PREPUBESCENT	-0.29078	OLD	0.8478260
BABYISH	-0.891304	TEENAGE	-0.144927	MATURE	0.8623188
CHILDISH	-0.804347	MIDDLEAGED	0.049645	PRIMITIVE	0.8695652
EARLIEST	-0.789855	FULL-GROWN	0.06383	SENIOR	0.8913043
INFANTILE	-0.789855	GROWNUP	0.078014	PRIMAL	0.8985507
VULNERABLE	-0.768115	PRIMORDIAL	0.0797101	ELDERLY	0.9275362
UNDERAGE	-0.659420	PREHISTORIC	0.33333	ARCHAIC	0.9347826
RECENT	-0.623188	JUVENILE	0.4565217	ANTIQUE	0.9710144
CHILD	-0.586956	AGED	0.6449275	PENSIONABLE	0.9710144
YOUNG	-0.586956	PRIMEVAL	0.7028985	ANCIENT	1
ADOLESCENT	-0.514492	ADULT	0.7173913		
YOUTHFUL	-0.514492	ANTIQUATED	0.7898550		
PUBESCENT	-0.442028	DECREPIT	0.7898550		

Table 6 - Age Fuzzy Dictionary

4 - FREQUENCY					
NEVER	-0.68	UNCOMMONLY	-0.165	ORDINARILY	0.4

HARDLY	-0.425	ON-OCCASION	-0.14035	FREQUENTLY	0.405
BARELY	-0.4	USUALLY	-0.005	OFTEN	0.405
SOMEWHAT	-0.4	HABITUALLY	0	REPEATEDLY	0.405
SCARCELY	-0.39	FAIRLY	0.085	CONSTANTLY	0.425
SELDOM	-0.365	INVARIABLY	0.135	CONTINUOUSLY	0.425
FAINTLY	-0.35	EXCEPTIONALLY	0.15	DAILY	0.425
NARROWLY	-0.335	MODERATELY	0.15	INEVITABLY	0.425
RARELY	-0.33	REGULARLY	0.25	GENERALLY	0.45
INFREQUENTLY	-0.325	ESPECIALLY	0.3	NORMALLY	0.45
SLIGHTLY	-0.325	PERIODICALLY	0.3	CONTINUALLY	0.5
NOTABLY	-0.3	COMMONLY	0.325	ROUTINELY	0.5
UNPREDICTABLY	-0.255	CUSTOMARILY	0.35	ALWAYS	0.575
CONVENTIONALLY	-0.245	NATURALLY	0.35	EXTREMELY	0.625
UNUSUALLY	-0.23	TYPICALLY	0.35	PERSISTENTLY	0.645
OCCASIONALLY	-0.2	CONSISTENTLY	0.4		

Table 7 - Frequency Fuzzy Dictionary

5 - LEVEL OF MEMBERSHIP					
BARELY	-1	ADEQUATE	-0.088	USUALLY	0.4
HARDLY	-0.968	ENOUGH	0.12	ALMOST	0.44
LITTLE	-0.92	RATHER	0.12	SUFFICIENT	0.44
SCARCELY	-0.88	HALFWAY	0.128	MAINLY	0.64
BIT	-0.76	MIDDLING	0.184	SERIOUSLY	0.672
SCRAPING	-0.76	SUITABLE	0.2	SUBSTANTIALLY	0.712
FRACTIONALLY	-0.648	AVERAGE	0.24	SIGNIFICANTLY	0.72
SLIGHTLY	-0.64	APPROPRIATE	0.36	LARGELY	0.76
PARTIALLY	-0.48	MOSTLY	0.36	GREATLY	1
JUST	-0.216	AMPLE	0.4	SUITABLE	0.2
SOMEWHAT	-0.16	GENERALLY	0.4		

Table 8 - Level of Membership Fuzzy Dictionary

6 - WORTH					
APPALLING	-1	UNDESIRABLE	-0.68965	PLEASANT	0.2068965
DIRE	-1	NASTY	-0.66667	DELIGHTFUL	0.3793103
DREADFUL	-1	INADEQUATE	-0.65517	ENJOYABLE	0.4137931
HORRENDOUS	-1	SUBSTANDARD	-0.58620	GOOD	0.4827586
INSUFFERABLE	-1	FINE	-0.41379	GREAT	0.5448275
INTOLERABLE	-1	MEDIOCRE	-0.41379	SUBLIME	0.5517241
USELESS	-0.95862	OK	-0.27586	LOVELY	0.5862068
UNSATISFACTORY	-0.93103	REASONABLE	-0.20689	WONDERFUL	0.6896551
UNBEARABLE	-0.91724	SUITABLE	-0.20689	SPLENDID	0.7172413
POOR	-0.89655	ACCEPTABLE	-0.13793	BRILLIANT	0.7241379
UNACCEPTABLE	-0.87586	FAIR	-0.137931	FANTASTIC	0.7379310
BAD	-0.83448	ADEQUATE	-0.068965	AMAZING	0.7931034
DISAPPOINTING	-0.82758	PERMISSIBLE	-0.068965	TREMENDOUS	0.8275862
TERRIBLE	-0.82758	ALRIGHT	-0.048275	ASTONISHING	0.8620689
AWFUL	-0.79310	MIDDLING	-0.034482	SUPERB	0.8965517
PATHETIC	-0.79310	SATISFACTORY	0	EXCELLENT	0.9310344
ROTTEN	-0.75862	NORMAL	0.0344827	MAGNIFICENT	0.9379310
UNPLEASANT	-0.75862	ORDINARY	0.0344827	MARVELLOUS	0.9655172
DISSATISFYING	-0.72413	PASSABLE	0.0344827	GLORIOUS	1
TEDIOUS	-0.69655	AVERAGE	0.1034482		
BORING	-0.68965	NICE	0.2068965		

Table 9 - Worth Fuzzy Dictionary

4.4.4 - Creating Fuzzy Word Category Ontologies

To show how words in a category are introduced on a scale of [-1, 1], it was necessary to construct a series of fuzzy word ontologies - with each ontology representing a fuzzy word category. One of the benefits of using an ontology is that it will incorporate the means of determining appropriate senses, allowing the program to evaluate the contexts in which words are used (Miller, 1995). The fuzzy word ontologies in the FUSE algorithm will be used as a complement to WordNet (Miller, 1995). These ratings will only be considered, when there are defuzzified ratings for fuzzy words present in the fuzzy dictionary of FUSE. For any word that is not present in the fuzzy dictionary of FUSE, the algorithm will use WordNet to calculate path length and depth of the Lowest (or Least) Common Subsumer (LCS) (See Figure 9 in Section 4.3). As mentioned in Section 2.2 *"An ontology consists of a hierarchical description of important classes (or concepts) in a particular domain, along with the description of the properties (of the instances) of each concept"* (Fullér, 2010). A fuzzy ontology as explained by Fuller (Fullér, 2010) is a quintuple (fivefold) represented by Equation 7 as:

$$F = < I, C, T, N, X >$$

Equation 7 (Source: Fullér, 2010)

Where *I* is the set of individual (objects) also referred to as the instances of the concepts, *C* is a set of fuzzy concepts where each concept is a fuzzy set on the domain of instances and the set of entities of the fuzzy ontology is $E = C \cup I$. *T* denotes the fuzzy taxonomy relations among the set of concepts *C* where the concepts are organised into sub-(super-)concept tree structures. The taxonomy relationship $T\ (i, j)$ indicates that the child *j* is a conceptual specification of the parent *i* with a certain degree. *N* denotes the set of non-taxonomy fuzzy associative relationships that relate entities across tree structures such as (*Naming Relationships*, describing the names of concepts, *Locating Relationships*, describing the relative location of concepts and *Functional Relationships*, describing the functions of concepts), finally *X* is the set of axioms in a proper logical language, i.e., predicates that constrain the meaning of concepts, individuals, relationships, and functions.

By taking this factor into considerations, each fuzzy word category is treated as a concept. Words within each concept are treated as instances. Each concept has a taxonomy that arranges the words as a binary tree so that the root node always takes the value 0. The defuzzified value of words are equally placed into nodes in intervals of ± 0.2, which was an empirically determined threshold.

As discussed in Section 2.5.2, STASIS is a corpus-based similarity measure that measures the level of similarity between two utterances using an ontological approach based on a taxonomy of words (Li et al., 2003). As mentioned in Section 3.6, FAST (Chandran et al., 2013) is an ontology-based similarity measure that uses concepts of fuzzy and CWW to allow for the accurate representation of fuzzy based words. The FUSE algorithm is inspired by STASIS and

FAST and has improved the approaches used in STASIS and FAST, to allow calculation of the path length and depth of the Lowest (or Least) Common Subsumer (LCS). LCS is the most specific common ancestor of two concepts computed in a given ontology. Semantically, it represents the commonality of the pair of concepts (Batet and Sánchez, 2015). In this research, it is calculated for specific fuzzy words which could not be achieved using traditional resources such as WordNet (Miller, 1995), as discussed in Section 2.2, due to lack of coverage of fuzzy words. Figure 13, shows the words in the category '*Temperature*' represented in an ontological structure. The numbers next to each word represent the defuzzified value of that word obtained from the human rating experiment described in Section 4.4.3. Each partition contains words up to a certain fixed value, with the negative values on the left side and the positive values on the right side; this allows path length to be calculated. The ontological structures for all six of the FUSE_1.0 fuzzy categories can be seen in Appendix A.

In FAST (Chandran, 2013), Chandran used a simple way of distinguishing the defuzzified values of fuzzy words using 4 nodes. Figure 14 shows the ontological structure used in FAST (Chandran, 2013). As can be seen, the negative values are only accommodated by two sub-groups (*veryneg* and *neg*) which are shown with the orange arrows, and the positive values also only contain two subgroups (*pos* and *verypos*) indicated by blue arrows. This limited grouping can cause great detail to be missed since the intervals are ± 0.4 as well as the average value being placed between -0.2 and 0.2, as opposed to 0 shown with the green arrow.

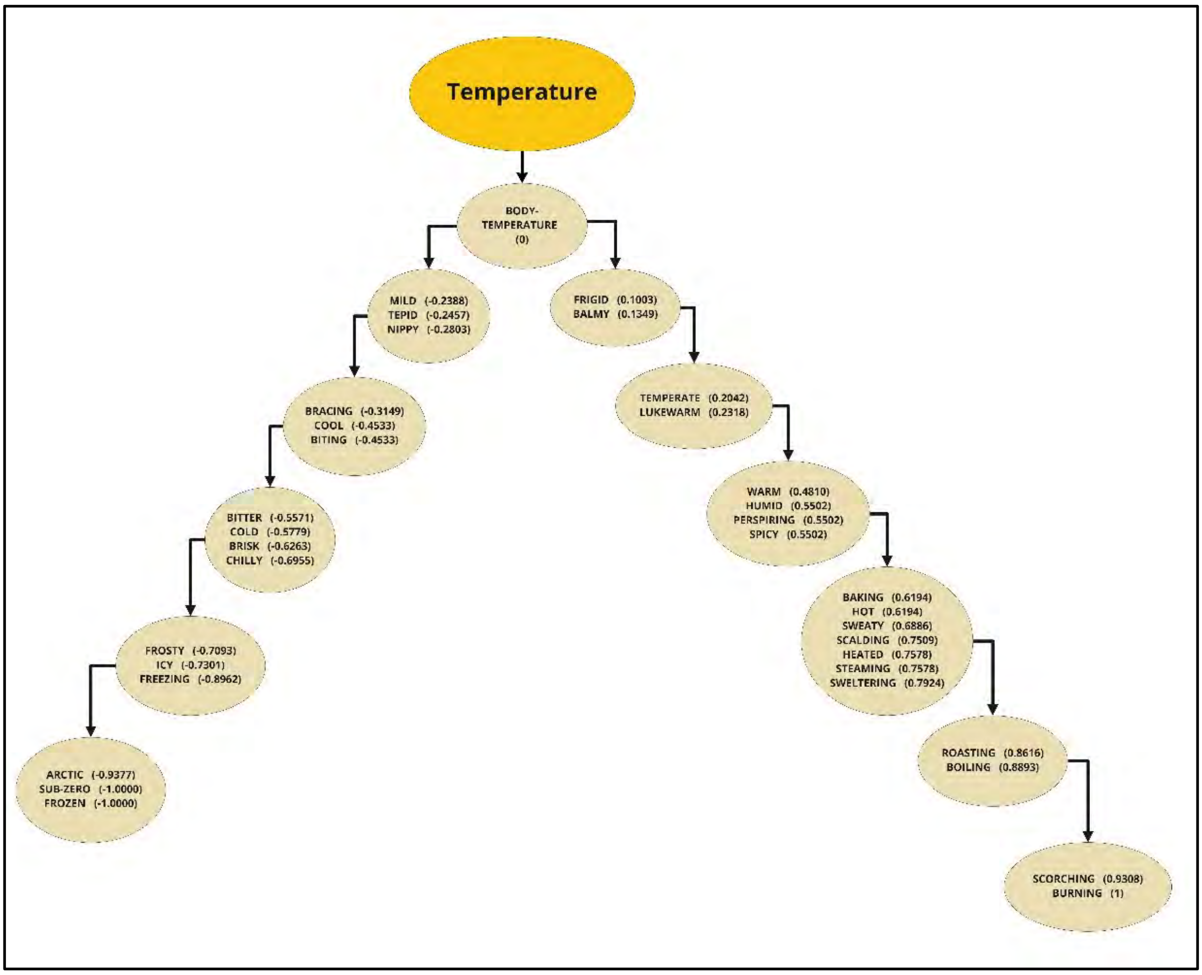
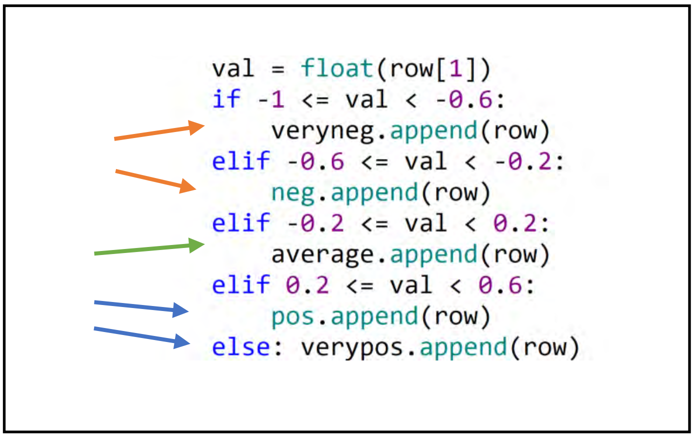


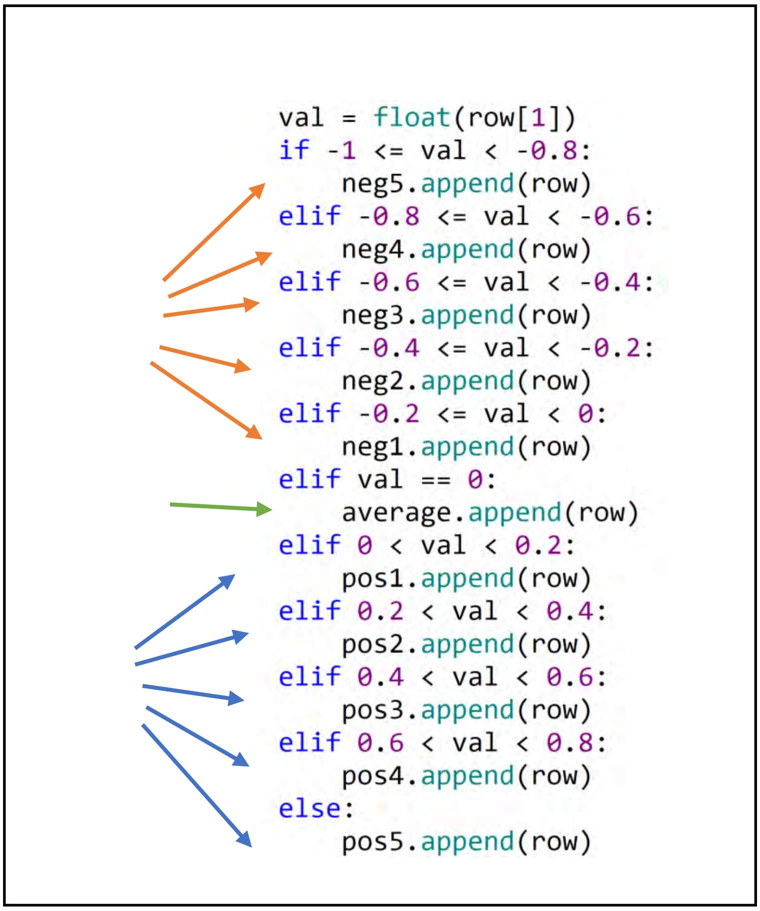
Figure 13 - Temperature Ontology Structure



```
val = float(row[1])
if -1 <= val < -0.6:
    veryneg.append(row)
elif -0.6 <= val < -0.2:
    neg.append(row)
elif -0.2 <= val < 0.2:
    average.append(row)
elif 0.2 <= val < 0.6:
    pos.append(row)
else: verypos.append(row)
```

Figure 14 - FAST Ontology

In the new proposed ontology, the FUSE algorithm uses more concepts with greater detail using intervals of ±0.2 and the average being placed at 0. Figure 15 shows the proposed ontological structure of the FUSE algorithm. As can be seen from the ±0.2 interval values (from *if… elif… else* condition) shown with the orange, green and blue arrows which will be used throughout all the versions of FUSE. Hence, the split has more details for the negative sub-group [comprising of 5 negative nodes shown using the orange arrow (*neg5*, *neg4*, *neg3*, *neg2*, *neg1*) and likewise for the positive sub-group comprising of 5 positive nodes shown using the blue arrow (*pos5*, *pos4*, *pos3*, *pos2*, *pos1*) with the centralized node being given an average value of 0, shown using the green arrow (*average*) (Adel et al., 2018).



```
val = float(row[1])
if -1 <= val < -0.8:
    neg5.append(row)
elif -0.8 <= val < -0.6:
    neg4.append(row)
elif -0.6 <= val < -0.4:
    neg3.append(row)
elif -0.4 <= val < -0.2:
    neg2.append(row)
elif -0.2 <= val < 0:
    neg1.append(row)
elif val == 0:
    average.append(row)
elif 0 < val < 0.2:
    pos1.append(row)
elif 0.2 < val < 0.4:
    pos2.append(row)
elif 0.4 < val < 0.6:
    pos3.append(row)
elif 0.6 < val < 0.8:
    pos4.append(row)
else:
    pos5.append(row)
```

Figure 15 - FUSE Ontology

FUSE_1.0 utilizes a semantic similarity measure which contains a word similarity measure referred to as STASIS (Li et al., 2003), when computing word similarity between nouns and verbs. When FUSE_1.0 encounters perception-based words within an utterance, word similarity is calculated through determining the path length, *l*, and the length of the shortest path, *d*, from the associated fuzzy category ontology. Ontologies were created for all six categories for FUSE_1.0 and the two structures shown in Figure 14 and Figure 15 where both tested with a sample dataset to perceive which approach is able to provide the highest correlation to human ratings. To test the two structures, the SWFD dataset designed by Chandran (Chandran et al., 2013) was used. SWFD contained 30 sentence pairs containing one fuzzy word, which was used for experimentation of FAST. Each sentence pair contains an associated average human rating score, that had been captured empirically through human experimentation (Chandran et al., 2013). Table 10 shows the 30 sentence pairs with the AHR for each sentence pair followed by the results obtained for each structure when run with FUSE_1.0.

Sentence Pairs	Sentences	AHR	Structure 1 (FAST Ontology)	Structure 2 (FUSE Ontology)
SP 1	When I was going out to meet my friends there was a short delay at the train station. The train operator announced to the passengers on the train that there would be a massive delay.	0.38	0.77	0.74
SP 2	I bought a small child's guitar a few days ago, do you like it? The old weapon choice reflects the personality of the carrier.	0.00	0.62	0.62
SP 3	You must realize that you will definitely be severely punished if you play with the alarm. He will absolutely be harshly punished for setting the fire alarm off.	0.73	0.67	0.67
SP 4	I will make you laugh so very hard that your sides ache and split. He will absolutely be harshly punished for setting the fire alarm off.	0.80	0.74	0.74
SP 5	Sometimes in a large crowd accidents may happen, which can cause life threatening injuries. There was a small heap of rubble left by the builders outside my house this morning.	0.13	0.55	0.55

SP 6	**I offer my sincere condolences to the parents of John Smith, who was unfortunately murdered.** **I extend my upmost sympathy to John Smith's parents, following his murder.**	0.87	0.77	0.77
SP 7	**If you continuously use these products, I guarantee you will look very young.** **I assure you that, by using these products over a long period of time, you will appear almost youthful.**	0.71	0.83	0.80
SP 8	**I always like to have a tiny slice of lemon in my drink, especially if it's coke.** **I like to put a large wedge of lemon in my drinks, especially cola.**	0.67	0.91	0.89
SP 9	**The key always never works, can you give me another?** **I dislike the word quay, it confuses me every time, I always think of the thing for locks, there's another one.**	0.10	0.68	0.67
SP 10	**Though it took many hours travel on the extremely long journey, we finally reached our house safely.** **We got home safely in the end, though it was a mammoth journey.**	0.82	0.67	0.67
SP 11	**The man presented a minuscule diamond to the woman and asked her to marry him.** **A man called Dave gave his fiancée an enormous diamond ring for their engagement.**	0.50	0.55	0.54
SP 12	**Does this soggy sponge look dry to you?** **Does pleasant music help you to relax or does it distract you too much?**	0.05	0.48	0.47
SP 13	**The tiny ghost appeared from nowhere and frightened the old man.** **The diminutive ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.**	0.33	0.57	0.59

SP 14	**Global warming is what everyone is really worrying about greatly today.** **Global warming is what everyone is mildly worrying about today.**	0.64	0.90	0.88
SP 15	**Midday is 12 o'clock in the midpoint of the day.** **Midday is 12 o'clock in the centre of the day.**	0.91	1.00	1.00
SP 16	**The first thing I do in a morning is make myself a lukewarm cup of coffee.** **The first thing I do in the morning is have a cup of hot black coffee.**	0.68	0.90	0.89
SP 17	**Just because I am middle aged, people shouldn't think I'm a responsible grown-up, but they do.** **Because I am the eldest one, I should be more responsible.**	0.32	0.30	0.30
SP 18	**This is a terrible noise level for a new car, I expected it to be of good quality.** **That's a very good car, on the other hand mine is great.**	0.21	0.51	0.49
SP 19	**Meet me on the huge hill behind the church in half an hour.** **Join me on the small hill at the back of the church in 30 minutes.**	0.68	0.78	0.76
SP 20	**It gives me immense pleasure to announce the winner of this year's beauty pageant.** **It's a great pleasure to tell you who has won our annual beauty parade.**	0.90	0.76	0.76

SP 21	**There is no point in trying hard to cover up what you said, we all know.** You shouldn't be burying what you feel.	0.35	0.59	0.59
SP 22	**Will I have to drive a great distance to get to the nearest petrol station?** Is it a long way for me to drive to the next gas station?	0.89	0.86	0.84
SP 23	**You have a very familiar face; do I know you from somewhere nearby?** You have a very familiar face; do I know you from somewhere where I used to live far away.	0.70	0.92	0.91
SP 24	**I have invited a great number of different people to my party so it should be interesting.** A small number of invitations were given out to a variety of people inviting them down the pub.	0.38	0.71	0.71
SP 25	**I am sorry but I can't go out as I have loads of work to do.** I've a gargantuan heap of things to finish so I can't go out I'm afraid.	0.89	0.60	0.59
SP 26	**Get that wet dog off my latest sofa.** Get that wet dog off my barely new sofa.	0.76	0.83	0.81
SP 27	**Will you drink a glass of excellent wine while you eat?** Would you like to drink this wonderful wine with your meal?	0.89	0.77	0.79

SP 28	**Can you get up that relatively small tree and rescue my cat, otherwise it might jump?** **Could you climb up the tall tree and save my cat from jumping please?**	0.69	0.86	0.86
SP 29	**Large Boats come in all shapes but they all do the same thing.** **Oversized Chairs can be comfy and not comfy, depending on the chair.**	0.13	0.39	0.39
SP 30	**I am so hungry I could eat a whole big horse plus desert.** **I could have eaten another massive meal, I'm still starving.**	0.66	0.55	0.57

Table 10 - SWFD Dataset

Table 11 shows the correlation of results for the SWFD with the AHR for each structure run with FUSE_1.0. As can be seen, Structure 2 gave a higher correlation of results and thus determined this structure to be used for the FUSE algorithm.

Dataset	**Structure 1** **(FAST Ontology)**	**Structure 2** **(FUSE Ontology)**
SWFD	0.6404	0.6491

Table 11 - SWFD Correlation with AHR

4.5 - Phase 2 of FUSE_1.0

4.5.1 - Designing The FUSE_1.0 Algorithm

This section formally defines the FUSE_1.0 algorithm. Given two fuzzy utterances, U_1 and U_2, compute their similarity $S(U_1, U_2)$. The FUSE_1.0 algorithm builds upon the original STASIS approach (Li et al., 2006), where the semantic similarity vectors and the word order similarity vectors for both the utterances are computed. These vectors are constructed using information about the word pairs and their associated depth and shortest path length in the

WordNet dictionary (Miller, 1995). The extra information about the fuzzy words were included, and when applicable, the lowest common subsumer depth and shortest path length are computed using the FUSE_1.0 approach (Adel et al., 2018). The information content measurements for the Brown Corpus (Francis and Kucera, 1979) are included. Combining all this information, allows the similarity between the two utterances to be computed.

w_i is denoted as a single word in either of the utterances for $i \in I$, some indexing set. Let $U = U_1 \cup U_2$ be the set of all distinct words appearing in U_1 or U_2. Following Li's approach (Li et al., 2006) T := {adjective, adverb, conjunction, determiner, noun, numeral, particle, pronoun, verb}, was set to be the set of all the possible tags to be assigned to each word w_i via the map

$$\tau : U_i \longrightarrow U_i \times T \text{ such that:}$$

$$\omega_i \coloneqq \tau(w_i) = (w_i, t)$$

Equation 8

This information is obtained from WordNet (Miller, 1995) and Brown's Corpus (Francis and Kucera, 1979). W_1 and W_2 are defined to be the sets of all the word-token pairs (w_i, t) from $U_1 \times T$ and $U_2 \times T$ respectively. The first stage of computation is shown in Algorithm 1, which populates these sets. Let $\omega_{i,j} \in W_1 \times W_2$ be a pair of word pairs ω_i and ω_j, i.e., $\omega_{i,j} \coloneqq (\omega_i , \omega_j)$. The set of all pairs of word-token couples are denoted by Ω.

The function $f\colon W_1 \times W_2 \longrightarrow \{0{,}1\}$ on the elements $\omega_{i,j} \in \Omega$, are defined via:

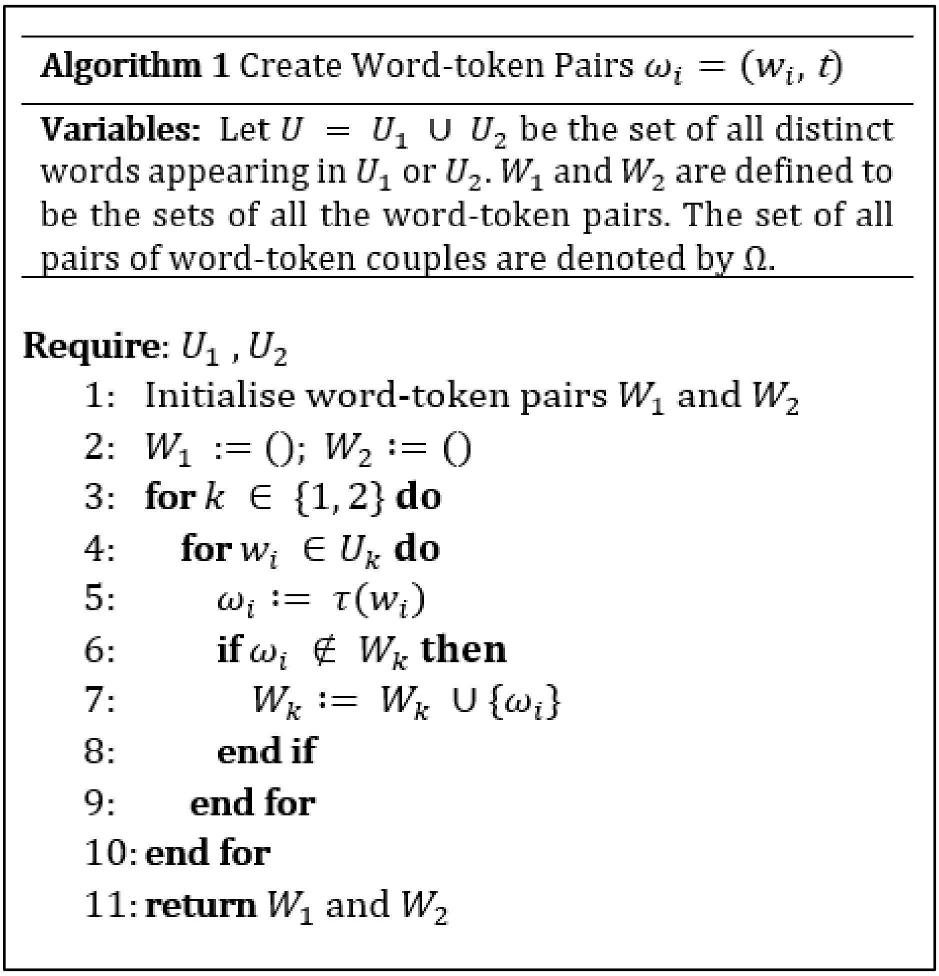
$$f(\omega_1, \omega_2) = \begin{cases} 1 & \text{if both are } \omega_1 \text{ and } \omega_2 \text{ are fuzzy words} \\ 0 & \text{otherwise} \end{cases}$$

Equation 9

Let C denote the set of fuzzy categories, where $C\colon=$ {size/distance, temperature, age, frequency, worth, level of membership, brightness, strength, speed}. The co-membership in a fuzzy category is determined by the function $c\colon W1 \times W2 \longrightarrow \{0{,}1\}$ such that:

$$c(\omega_1, \omega_2) \;= \begin{cases} 1 & \text{if } \omega_1 \text{ and } \omega_2 \text{ are in the same fuzzy category } C \\ 0 & \text{otherwise} \end{cases}$$

Equation 10



Algorithm 1 Create Word-token Pairs $\omega_i = (w_i, t)$

Variables: Let $U = U_1 \cup U_2$ be the set of all distinct words appearing in U_1 or U_2. W_1 and W_2 are defined to be the sets of all the word-token pairs. The set of all pairs of word-token couples are denoted by Ω.

```
Require: $U_1$ , $U_2$
  Initialise word-token pairs $W_1$ and $W_2$
  $W_1 := ()$; $W_2 := ()$
  for $k \in \{1, 2\}$ do
    for $w_i \in U_k$ do
      $\omega_i := \tau(w_i)$
      if $\omega_i \notin W_k$ then
        $W_k := W_k \cup \{\omega_i\}$
      end if
    end for
  end for
  return $W_1$ and $W_2$
```

Algorithm 1 - Create Word-Token Pairs

If two words are not in the same fuzzy category or are not both fuzzy words, it calculates the depth and shortest path length from the values obtained from WordNet:

The depth of the word pair is computed via:

$$SD\colon \Omega \longrightarrow (0,1) \text{ such that}$$

$$\omega_{i,j} \longmapsto d_{i,j}$$

Equation 11

The path length via:

$$SL\colon \Omega \longrightarrow (0,1) \text{ such that}$$

$$\omega_{i,j} \longmapsto l_{i,j}$$

Equation 12

Word similarity $wordSim$ via

$$wordSim: \Omega \times \mathbf{R} \times \mathbf{R} \longrightarrow$$

$$wordSim(\omega_{i,j}, d_{i,j}, l_{i,j}) \longmapsto e^{-\alpha l} \cdot tanh\,(\beta d)$$

Equation 13

Where $d := d_{i,j}, l := l_{i,j}$ and the parameters α and β need to be empirically determined (which will be fully explained in Section 5.2.1).

However, if two fuzzy words come from the same fuzzy category $c \in C$, the lowest common subsumer depth and the shortest path length can be computed within this ontology. These are denoted by FD and FL, the fuzzy analogues to SD and SL, coming from the FUSE_1.0 ontology. These attributes are used in Algorithm 2 to compute the matrix of similarities of the word pairs $\omega_{i,j}$. Finally, in Algorithm 3, for each of the utterances U_k, it computes the semantic similarity vector s_k and the word order (syntactic) similarity vector r_k. The angular distances between these determine the level of the similarity, and thus:

1. The semantic similarity S_s is computed as the cosine of the angle γ between the vectors s_1 and s_2:

$$S_s := \frac{s_1 \cdot s_2}{||s_1||\,||s_2||} = cos(\gamma)$$

Equation 14

2. The word order (syntactic) similarity S_r is computed in terms of tan of half the angle μ between the word order vectors r_1 and r_2:

$$S_r := 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} = 1 - tan(\tfrac{1}{2}\mu)$$

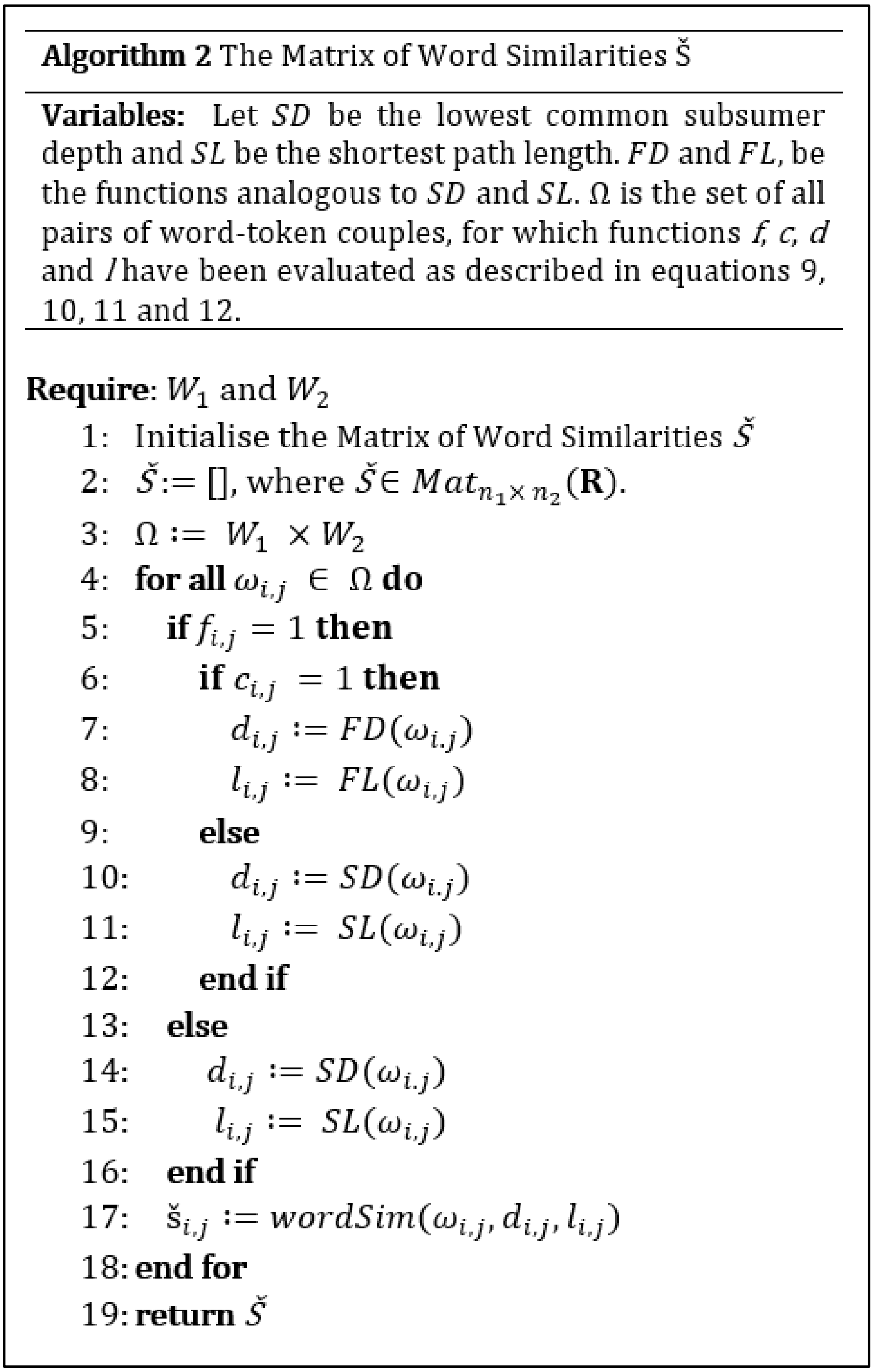
Equation 15

3. The similarity of the two utterances S is determined to be a linear combination of S_s and S_r:

$$S(U_1, U_1) := \delta cos(\gamma) + (1 - \delta)tan(\frac{1}{2}\mu),$$

Equation 16

where $0 < \delta \leq 1$ decides the relative contributions of semantic and word order information to the overall similarity computation.

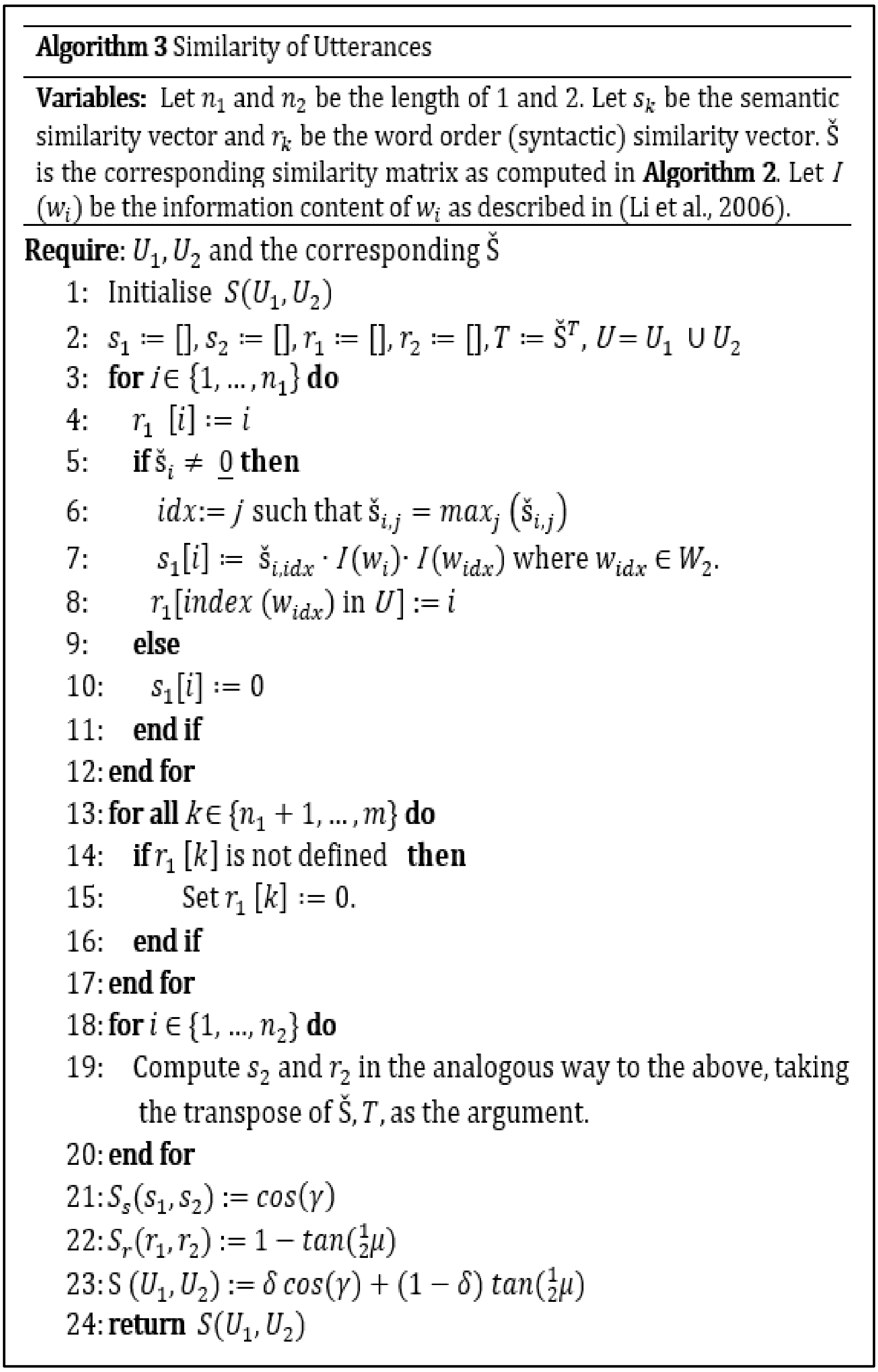


Algorithm 2 The Matrix of Word Similarities Š

Variables: Let SD be the lowest common subsumer depth and SL be the shortest path length. FD and FL, be the functions analogous to SD and SL. Ω is the set of all pairs of word-token couples, for which functions f, c, d and l have been evaluated as described in equations 9, 10, 11 and 12.

```
Require: $W_1$ and $W_2$
  Initialise the Matrix of Word Similarities $\check{S}$
  $\check{S} := []$, where $\check{S} \in Mat_{n_1 \times n_2}(\mathbf{R})$.
  $\Omega := W_1 \times W_2$
  for all $\omega_{i,j} \in \Omega$ do
    if $f_{i,j} = 1$ then
      if $c_{i,j} = 1$ then
        $d_{i,j} := FD(\omega_{i.j})$
        $l_{i,j} := FL(\omega_{i,j})$
      else
        $d_{i,j} := SD(\omega_{i.j})$
        $l_{i,j} := SL(\omega_{i,j})$
      end if
    else
      $d_{i,j} := SD(\omega_{i.j})$
      $l_{i,j} := SL(\omega_{i,j})$
    end if
    $š_{i,j} := wordSim(\omega_{i,j}, d_{i,j}, l_{i,j})$
  end for
  return $\check{S}$
```

Algorithm 2 - Matrix of Word Similarities Š



Algorithm 3 Similarity of Utterances

Variables: Let n_1 and n_2 be the length of 1 and 2. Let s_k be the semantic similarity vector and r_k be the word order (syntactic) similarity vector. Š is the corresponding similarity matrix as computed in **Algorithm 2**. Let I (w_i) be the information content of w_i as described in (Li et al., 2006).

```
Require: $U_1, U_2$ and the corresponding Š
    Initialise $S(U_1, U_2)$
    $s_1 := [], s_2 := [], r_1 := [], r_2 := [], T := Š^T$, $U = U_1 \cup U_2$
    for $i \in \{1, \ldots, n_1\}$ do
        $r_1 \ [i] := i$
        if $š_i \neq \underline{0}$ then
            $idx := j$ such that $š_{i,j} = max_j \left(š_{i,j}\right)$
            $s_1[i] := š_{i,idx} \cdot I(w_i) \cdot I(w_{idx})$ where $w_{idx} \in W_2$.
            $r_1[index \ (w_{idx})$ in $U] := i$
        else
            $s_1[i] := 0$
        end if
    end for
    for all $k \in \{n_1 + 1, \ldots, m\}$ do
        if $r_1 \ [k]$ is not defined then
            Set $r_1 \ [k] := 0$.
        end if
    end for
    for $i \in \{1, \ldots, n_2\}$ do
        Compute $s_2$ and $r_2$ in the analogous way to the above, taking
        the transpose of Š, $T$, as the argument.
    end for
    $S_s(s_1, s_2) := cos(\gamma)$
    $S_r(r_1, r_2) := 1 - tan(\frac{1}{2}\mu)$
    S $(U_1, U_2) := \delta \ cos(\gamma) + (1 - \delta) \ tan(\frac{1}{2}\mu)$
    return $S(U_1, U_2)$
```

Algorithm 3 - Similarity of Utterances (U1, U2)

4.5.2 - Illustrative Example of FUSE_1.0 with Sample Sentence Pair

To illustrate how the overall sentence similarity is computed for a pair of fuzzy utterances in FUSE_1.0, a detailed description is provided below for two sample utterances:

U_1: it looks like a warm day

U_2: the weather is chilly today

A joint word set of *U* is formed where (Algorithm 1)
U = {It looks like a warm day today the weather is chilly}

The semantic vectors for [U_1 and U_2] are derived from *U* using WordNet (Miller, 1995) and the fuzzy dictionary in FUSE_1.0 as shown in Table 12 (Algorithm 2).

	it	looks	like	a	warm	day	the	weather	is	chilly	today
It	1										
looks		1	0.4565								
like			1								
A				1							
warm					1			0.5331		0.5452	
day						1					0.9806
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Š	1	1	1	1	1	1	0	0.5331	0	0.5452	0.9806
Weight	*I*(it) *I*(it)	*I*(looks) *I*(looks)	*I*(like) *I*(like)	*I*(a) *I*(a)	*I*(warm) *I*(warm)	*I*(day) *I*(day)	*I*(the) *I*(the)	*I*(warm) *I*(weather)	*I*(is) *I*(is)	*I*(warm) *I*(chilly)	*I*(day) *I*(today)

Table 12 - Process for Deriving the Semantic Vector

The first row in Table 12 lists words in the joint word set *U* and the first column in Table 12 lists words in U_1.

For each word in *U*, if the same word exists in U_1, the cell at the cross point is set to 1 or set to their similarity value, otherwise, the cell at the cross point of the most similar word is set to their similarity value or 0, dependent on whether the highest similarity value exceeds the pre-set threshold of 0.2.

The second to last row $\check{S}$, is the lexical vector obtained by selecting the largest value in each column (Algorithm 3).

The last row lists the corresponding information content for weighting the significant weights of each word obtained from the Brown's corpus (Francis and Kucera, 1979). The word order vectors r_1 and r_2 for the example fuzzy sentence pair U_1 and U_2 are as follows:

r_1 = {1 2 3 4 5 6}

r_2 = {7 8 9 10 11}

The Semantic similarity for sentence pair U_1 and U_2 is (0.5529) and the Syntactic similarity is (0.7213). Finally, the total similarity between sentence pair U_1 and U_2 = (0.15 * 0.7213) + (0.85 * 0.5529) = 0.5782, using α = 0.15 and β = 0.85 which are empirically derived values (Algorithm 3) (fully explained in Section 5.2.1).

4.6 - Conclusion

This chapter has introduced the first version of the FUSE algorithm, referred to as FUSE_1.0, and highlighted the research methodology used to create this algorithm. FUSE_1.0 is based on STASIS (Li et al., 2006) and FAST (Chandran, 2013) algorithms, and developed over two initial phases. Phase 1 involved the expansion of six categories of fuzzy words and experiments on gathering human ratings, the modelling of fuzzy words using Interval Type-2 fuzzy sets approach and the design of six fuzzy ontologies. Phase 2 covered designing the FUSE_1.0 algorithm using the findings of Phase 1 with an illustrative example of how sentence similarity is calculated using FUSE_1.0.

Chapter 5 will explain the experiments and evaluation conducted on the FUSE_1.0 algorithm, using several published datasets, in order to test its performance, with the results being compared with STASIS (where fuzzy word similarity is not considered) and FAST (only published fuzzy sentence similarity that uses Type-1 fuzzy sets).

CHAPTER 5

CHAPTER 5: EMPIRICAL EXPERIMENTS ON FUSE_1.0

5.1 - Introduction

Chapter 4 presented the FUSE algorithm and the fuzzy dictionary which formed the first version of FUSE referred to as FUSE_1.0.

This chapter will describe a series of experiments conducted on FUSE_1.0, which firstly seek to validate the collection and modelling of fuzzy words by humans, through application within FUSE 1.0. Secondly, FUSE_1.0 is evaluated using a range of published datasets and the results are compared with both the traditional semantic similarity measure, STASIS (Li et al., 2006) and the fuzzy semantic similarity measure, FAST (Chandran et al., 2013). The result of these experiments will contribute towards the first research question presented in Section 1.4:

RQ1. Investigate the feasibility of utilising Type-2 Fuzzy Sets and their representation of an individual's perception of fuzzy words and evaluate the suitability of the resulting fuzzy word models for incorporation into a Fuzzy Semantic Similarity Measure (FSSM).

The work in this chapter was published in **FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets.** N. Adel, K. Crockett, A. Crispin, D. Chandran, JP. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Brazil, 2018. DOI: 10.1109/FUZZ-IEEE.2018.8491641

5.2 - Experimental Methodology

To evaluate the modelling of fuzzy words in the fuzzy dictionary of FUSE_1.0 using Interval Type-2 FS (Tables 4 - 9 in section 4.4.3.2), it was important to test the FUSE_1.0 algorithm and determine the improvements compared to both STASIS and FAST. To achieve this, datasets were needed to determine the performance of using FUSE_1.0 with the sentence pairs in each dataset, with datasets containing both fuzzy words and non-fuzzy words.

Three published datasets (MWFD (Chandran, 2013), STSS-65 (J. O'Shea et al., 2013) and STSS-131 (J. O'Shea et al., 2013)) were used for this experiment. The sentence similarity ratings obtained from FUSE_1.0 were compared to those obtained from human ratings to find out the improvements of the expansion of categories and the ontology representation of FUSE_1.0 over both STASIS (Li et al., 2006) and FAST (Chandran et al., 2013).

A description of each dataset is provided in Section 5.2.1. The hypothesis which will be tested in this experiment is defined as:

H_0: Will the expansion of the fuzzy dictionary for the existing six categories in FUSE_1.0 give a higher overall correlation with human ratings, when compared to other Semantic Similarity Measures such as STASIS and FAST.

Each published dataset (explained in section 5.2.2) contains pairs of sentences and an associated average human rating score that had been captured empirically through human experimentation. Each dataset was run through FUSE_1.0 and the similarity of each sentence pair was calculated. Likewise, each dataset was also run with STASIS and FAST. The results for each algorithm were then compared with the average human rating for each of the published datasets. To analyse the results, the Pearson's correlation coefficient (Kent State University, 2013) will be used to calculate the correlation of the quantitative data collected, compared with the Average Human Rating (AHR), which is fully explained in Section 5.3.

5.2.1 - Determining Syntactic & Semantic Weights

Prior to any experiments being conducted with FUSE_1.0, it was essential to determine the α (syntactic similarity) and β (semantic similarity) values of the FUSE algorithm mentioned in Algorithm 1 in Section 4.5.1. To do this, the α (syntactic similarity) and β (semantic similarity) values of STASIS (Li et al., 2003) was examined, and different values for each similarity value were empirically tested to determine which value produces the optimum result.

In FUSE_1.0, an optimal result is a sentence similarity rating that is close to the AHR. The SWFD dataset designed by Chandran (Chandran et al., 2013) contained 30 sentence pairs containing one fuzzy word, which was used for experimentation of FAST. A sentence pair from this dataset was selected at random shown in Table 13. This sentence pair contains an associated average human rating score, that had been captured empirically through human experimentation by Chandran (Chandran et al., 2013).

Sentence 1	Sentence 2	AHR
I always like to have a *tiny* slice of lemon in my drink, especially if it’s coke.	**I like to put a *large* wedge of lemon in my drinks, especially cola.**	0.67

Table 13 - SWFD Sample Sentence

The α and β values of FUSE_1.0 were changed in increments of 0.05, with total weights equaling one, and the test sentence pair was run, and results recorded. Table 14 shows the results of this experiment, and it can be seen that the $\alpha = 0.15$ and $\beta = 0.85$ values produced the closet results to the AHR with the use of FUSE_1.0. These values returned a similarity rating of 0.6739, which is the closest to the AHR value from Table 13 which is 0.67. Thus, the values of α (syntactic similarity) and β (semantic similarity) for the FUSE algorithm were determined.

Syntactic Weight (α)	Semantic Weight (β)	Syntactic Similarity	Semantic Similarity	Overall Similarity
0.00	1.00	0.4497	0.7135	0.7135
0.05	0.95	0.4497	0.7135	0.7003
0.10	0.90	0.4497	0.7135	0.6871
0.15	0.85	0.4497	0.7135	0.6739
0.20	0.80	0.4497	0.7135	0.6607
0.25	0.75	0.4497	0.7135	0.6475
0.30	0.70	0.4497	0.7135	0.6343
0.35	0.65	0.4497	0.7135	0.6211
0.40	0.60	0.4497	0.7135	0.6079
0.45	0.55	0.4497	0.7135	0.5948
0.50	0.50	0.4497	0.7135	0.5816
0.55	0.45	0.4497	0.7135	0.5684
0.60	0.40	0.4497	0.7135	0.5552
0.65	0.35	0.4497	0.7135	0.5420
0.70	0.30	0.4497	0.7135	0.5288
0.75	0.25	0.4497	0.7135	0.5156
0.80	0.20	0.4497	0.7135	0.5024
0.85	0.15	0.4497	0.7135	0.4892
0.90	0.10	0.4497	0.7135	0.4760
0.95	0.05	0.4497	0.7135	0.4629
1.00	0.00	0.4497	0.7135	0.4497

Table 14 - Result of α and β experiment for FUSE_1.0

5.2.2 - Datasets

To test H_0, three published datasets were used. These consisted of:

- Multi Word Fuzzy Dataset [MWFD] (Chandran, 2013);
- 65 Sentence Pair Dataset[STSS-65] (J. O'Shea et al., 2013);
- 131 Sentence Pair Dataset [STSS-131] (J. O'Shea et al., 2013).

MWFD consists of 30 sentence pairs that have been purposely placed with at least two fuzzy words in each sentence. Sentences were taken from the Gutenberg Corpus (Gomes et al., 2006) and random fuzzy words from the six categories of FAST were substituted in each sentence to create this dataset (Chandran, 2013). The two gold standard datasets STSS-65 and STSS-131 (J. O'Shea et al., 2013) contained 65 and 131 short text sentence pairs respectively and did not contain any purposely placed fuzzy words.

To create the two datasets of STSS-65 and STSS-131 O'Shea (J. O'Shea, 2010) used 32 participants, and asked participants to write two sentences (between 10 to 20 words) in natural language dialogue derived from 16 stimulus words, thus creating sentences in an unbiased manner. 32 different participants were then asked to rate sentence pairs on a scale of [0, 4] in terms of similarity (J. O'Shea, 2010). All three datasets have an associated average human rating score for each sentence pair that had been derived from empirical experiments by Chandran (Chandran, 2013) and O'Shea (J. O'Shea, 2010).

5.3 - Analysis Methods

Each of the three datasets was first executed with the FUSE_1.0 algorithm, and then with both STASIS (Li et al., 2006) and FAST (Chandran et al., 2013), and the ratings from each sentence pair were recorded. The Pearson's correlation coefficient (Kent State University, 2013) is used to calculate the correlation of the collected quantitative data. Each collected dataset is then compared with the Average Human Rating (AHR). The Pearson's correlation (*r-value*) allows statistical evidence to be calculated, and produces a linear relationship between two variables *x* and *y,* and the result can be computed as shown in Equation 17 (Kent State University, 2013):

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)} \,.\, \sqrt{\mathrm{var}(y)}}$$

Equation 17 (Source: Kent State University, 2013)

Where r_{xy} is the correlation coefficient, cov(*x*, *y*) is the sample covariance of *x* and *y*; var(*x*) is the sample variance of *x*; and var(*y*) is the sample variance of *y*.

By calculating the *r-value* for each of the three datasets per algorithm versus the AHR, allowed an effective comparison, that takes place with all three proposed algorithms. The *r-value* should be normalised between [-1, 1], as the value [-1] shows a perfectly negative linear relationship, which will mean a decreasing relationship, [0] shows no relationship, and [1] shows a perfectly positive linear relationship which will mean an increasing relationship. The magnitude of the value (how close it is to -1 or 1) will indicate the strength of the correlation (Kent State University, 2013; Bird et al., 2009).

5.4 - Results and Discussion

Appendix B shows the similarity results per sentence pair for the three datasets MWFD (Table B-1), STSS-65 (Table B-2) and STSS-131 (Table B-3), when run on STASIS, FAST and FUSE_1.0 and well as the AHR.

Table 15 shows the *r-value* calculated from the three datasets MWFD, STSS-65 and STSS-131 tested against the three algorithms STASIS, FAST and FUSE_1.0, versus the AHR from the results in Appendix B. Each of the rows represent the three datasets being used and each column represents the three algorithms being tested. The test results from each dataset/algorithm were presented in the column represented by the *r-value* for each algorithm.

Based on the results shown in Table 15, FUSE_1.0 produced a higher correlation (*r* = 0.7682) in the **MWFD** dataset with human ratings, compared with both STASIS (*r* = 0.7452) and FAST (*r* = 0.7305).

Algorithms Datasets	STASIS *(r-value)*	FAST *(r-value)*	FUSE_1.0 *(r-value)*
MWFD	0.7452	0.7305	0.7682
STSS-65	0.6813	0.6908	0.6910
STSS-131	0.5208	0.5163	0.5180

Table 15 - r-value Results

Additionally, FUSE_1.0 gave a higher correlation (*r* = 0.6910) in the dataset **STSS-65** than both STASIS (*r* = 0.6813) and FAST (*r* = 0.6908). Similar performance in the dataset **STSS-131**, which FUSE_1.0 produced a higher correlation (*r* = 0.5180) than FAST (*r* = 0.5163), except it was marginally smaller than STASIS (*r* = 0.5208). This slight lack of performance is mainly due to the large number of crisp words in the STSS-131 dataset and the presence of a low number of fuzzy words. Since the FUSE algorithm is a *fuzzy* similarity measure, therefore it relies on the presence of fuzzy words in a sentence to calculate similarity.

A graphical representation of these results is shown in Figure 16. It can be seen from the graph that FUSE_1.0 gave a higher correlation, against both STASIS and FAST for the datasets MWFD and STSS-65; this shows that higher representation of fuzzy words in the fuzzy dictionary of

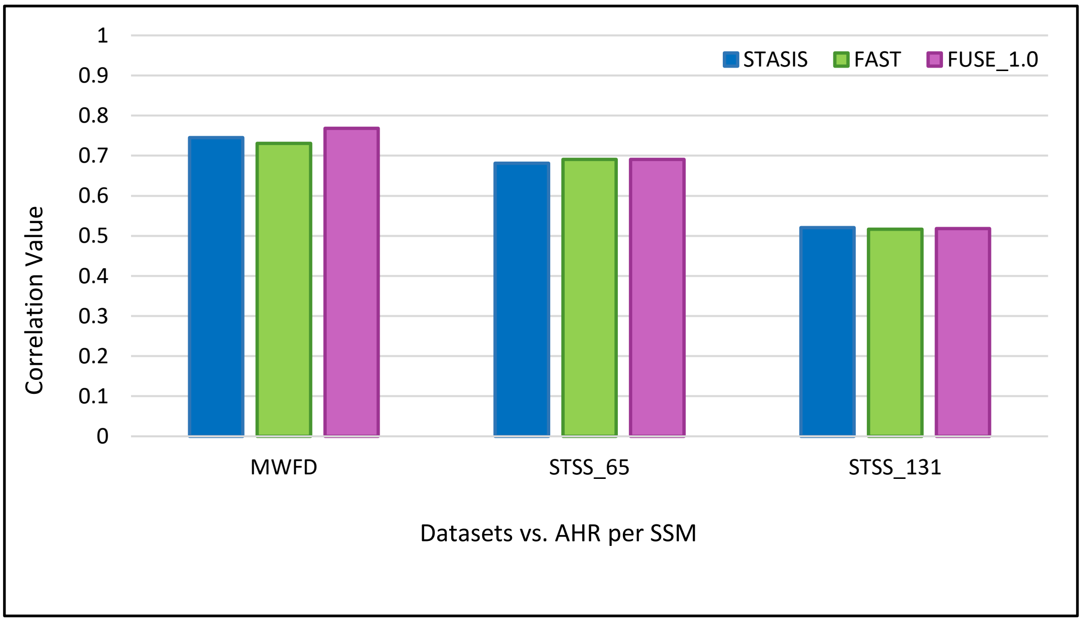


Figure 16 - Correlation Results for Datasets vs. AHR per SSM

the FUSE_1.0 algorithm plays a positive role in increasing the correlation to that of human ratings.

Referring back to the H_0 in this experiment (Section 5.2), Table 16 shows that the *p-value* for each dataset tested is less than 0.05 (p < .001) for a confidence level of 95% for all three algorithms, thus providing support for the research hypothesis.

Referring to the detail of the results in Table 16, the three algorithms (STASIS, FAST and FUSE_1.0) are analysed with some example sentence pairs of instances, where STASIS and FUSE_1.0 performed better or worst. The goal of similarity obtained from each algorithm is not about how high the result value produced from the three algorithms are, but rather how close is the result compared to the rating given by the AHR. Full results for each dataset are available in Appendix B.

Algorithms Datasets	STASIS *(p-value)*	FAST *(p-value)*	FUSE_1.0 *(p-value)*
MWFD	*p* < .001	*p* < .001	*p* < .001
STSS-65	*p* < .001	*p* < .001	*p* < .001
STSS-131	*p* < .001	*p* < .001	*p* < .001

Table 16 - p-value Results

5.4.1 - Results ruled in favour of STASIS

Table 17 shows three examples representing each of the three tested datasets (MWFD, STSS-65 and STSS-131), where STASIS produced a closer rating to the AHR than both FAST and FUSE_1.0 with the following scenario:

- Where there is a fuzzy word present in a sentence pair (it is shown in the table with *italics/underlined*).
- In SP 9, from the MWFD dataset, the fuzzy words are (*massive* and *little*) belonging to the Size/Distance category and (*mediocre* and *poor*) belonging to the Worth category.

- Although SP 60 from the STSS-65 dataset was not a fuzzy dataset, it can be seen that there are words in this sentence pair that fall into the fuzzy category (*child* and *young*) belonging to the Age category.
- SP 72 from the STSS-131 dataset does not have any fuzzy words present.

One possible explanation of STASIS outperforming both FAST and FUSE_1.0 for the two sentence pairs that did contain fuzzy words (MWFD and STSS-65) is the difference in the Natural Language Took Kit (NLTK) (Loper and Bird, 2002) versions from when STASIS was originally developed, and the versions used with FAST and FUSE_1.0. NLTK covers symbolic and statistical natural language processing and is interfaced to annotated corpora (Loper and Bird, 2002). Furthermore, STASIS outperformed FAST and FUSE_1.0 for SP 72 from the STSS-131 dataset due to there being no fuzzy words present in this sentence. FUSE_1.0 is a fuzzy SSM and thus it relies on the presence of fuzzy words in a sentence pair for optimal performance.

Sentence Pairs	**Sentences**	**AHR**	**STASIS**	**FAST**	**FUSE_1.0**
SP 9 (MWFD)	**Have *massive* mercy on the *mediocre* men** **Have a *little* mercy on the *poor* men**	0.4873	**0.7940**	0.8074	0.8428
SP 60 (STSS-65)	**A boy is a *child* who will grow up to be a man.** **A lad is a *young* man or boy.**	0.5800	**0.7248**	0.7420	0.7593
SP 72 (STSS-131)	**You shouldn’t be covering what you really feel.** **There is no point in covering up what you said, we all know.**	0.5525	**0.5776**	0.5793	0.5793

Table 17 - STASIS Good Performing SP Examples

5.4.2 - Results ruled in favour of FUSE_1.0

Table 18 shows three examples representing each of the three datasets tested (MWFD, STSS-65 and STSS-131), where FUSE_1.0 produced a better rating closer to the AHR than STASIS and FAST, with the following scenario:

- Where there is a fuzzy word present in a sentence pair it is shown in the table with *italics/underlined*.
- In SP 12, which is from the MWFD dataset, the fuzzy words are (*almost* and *rather*) belonging to the Level of Membership category.
- In SP 55, which is taken from the STSS-65 dataset, although this dataset was not a fuzzy dataset, there are still words in this sentence pair that fall into the fuzzy category (*specially* and *often*) belonging to the Frequency category.
- SP 67 is taken from the STSS-131 and like STSS_65, although this dataset is not a fuzzy dataset, there are words in this sentence pair that fall into the fuzzy category (*seriously*) which is repeated in both sentences belonging to the Frequency category.

From the results shown in Table 18, it is proven that modelling the words using the Interval Type-2 FS approach plays an important factor in bringing the similarity closer to the AHR. It is further apparent that FAST is also a FSSM that caters for the presence of fuzzy words in a sentence pair. However due to the modelling of fuzzy words using the Type-1 approach in FAST, it can be seen that the ratings are not as close to the AHR compared to FUSE_1.0.

Sentence Pairs	Sentences	AHR	STASIS	FAST	FUSE_1.0
SP 12 (MWFD)	**And he laughed *almost* dreadfully** **And he laughed *rather* unpleasantly**	0.7127	0.4997	0.6269	**0.6284**
SP 55 (STSS-65)	**An autograph is the signature of someone famous which is *specially* written for a fan to keep.** **Your signature is your name, written in your own characteristic way, *often* at the end of a document to indicate that you wrote the document or that you agree with what it says.**	0.4050	0.7649	0.7649	**0.7579**
SP 67 (STSS-131)	**I advise you to treat this matter very *seriously* as it is vital.** **You must take this most *seriously*, it will affect you.**	0.8450	0.7271	0.7271	**0.7587**

Table 18 - FUSE_1.0 Good Performing SP Examples

5.5 - Conclusion

This chapter has described a series of the experiments conducted with the FUSE_1.0 algorithm. Results have shown a higher correlation value compared with human ratings (AHR) than the other two semantic similarity algorithms, STASIS and FAST for the MWFD dataset (r = 0.7682) and the STSS-65 dataset (r = 0.6910). For the STSS-131 dataset, FUSE_1.0 gave a higher correlation (r = 0.5180) than FAST, and it was marginally lower than STASIS. This is mainly due to a large number of crisp words present in the STSS-131 dataset and the occurrence of very few fuzzy words. Since the FUSE algorithm in general is a *fuzzy* similarity measure and thus, it relies on the presence of fuzzy words in a sentence to calculate similarity.

Referring to the H_0 proposed in Section 5.2, the improvement FUSE_1.0 had over STASIS and FAST for the three datasets of MWFD, STSS-65 and STSS-131 is down to several factors. Firstly, the coverage of fuzzy words is far greater in FUSE_1.0, due to the increase of fuzzy words in the fuzzy dictionary as explained in Section 4.4.3.2. Secondly, a new set of fuzzy ontologies has been developed for these categories in FUSE_1.0 (Appendix A). Finally, the ability to represent uncertainty using Interval Type-2 fuzzy sets, as opposed to Type-1 has been shown to contribute towards a higher correlation between FUSE_1.0 and the average human ratings.

However, as mentioned in Section 5.4, there is a degree of subjectivity in gathering human ratings. This was discovered through the interpretation of the *r-value* for the STSS-131 dataset, where there is a greater presence of non-fuzzy words, the *r-value* is reduced. This is because the FUSE_1.0 algorithm is designed to cater for the presence of fuzzy words in sentences. Thus, the smaller the presence of these types of words, the lower the *r-value* will be, as the ratings in the fuzzy dictionary are not used, but instead ratings are gathered from WordNet.

Chapter 6 will examine a series of improvements with the FUSE_1.0 algorithm. Firstly, by introducing the concept of linguistic hedges into the algorithm. The FUSE algorithm will then be further developed through different versions by expanding the six fuzzy categories of FUSE_1.0 to nine fuzzy categories (FUSE_2.0). Negation operators such as '*not'* are explored in FUSE_3.0 and finally a fuzzy influence factor is introduced in FUSE_4.0, which will investigate the presence of fuzzy words in a pair of sentences, which do not belong to the same fuzzy category, as is the existence case with FUSE_1.0.

CHAPTER 6

CHAPTER 6: THE EVOLUTION OF FUSE

6.1 - Introduction

Chapter 5 discussed the experiments conducted with FUSE_1.0 and the results had shown an improvement on the correlations with the average human rating (AHR) compared to both STASIS, a traditional SSM, and FAST, a FSSM. However, on analysis of the experimental results, a number of weaknesses were identified in FUSE_1.0. This chapter will seek to address these weaknesses.

To overcome the weaknesses of FUSE_1.0, four modifications will be made to the FUSE_1.0 algorithm, over three versions:

i) FUSE_1.0 will be modified to address the presence of linguistic hedges (such as '*very*') in fuzzy utterances and introduce a Fuzzy Hedge category.

ii) FUSE_2.0 will address the limitations of fuzzy words in the six categories. This will look to expand the current fuzzy categories by introducing three new fuzzy categories, and the subsequent fuzzy words collection in each new fuzzy category.

iii) FUSE_3.0 will address the presence of logical negation in fuzzy utterances (such as '*not*') and explore how utterance can be assessed in a FSSM.

iv) FUSE_4.0 will address the presence of the impact of fuzzy words in different categories in fuzzy utterances, with the introduction of the Fuzzy Influence Factor. Up to this stage, the FUSE algorithm only used fuzzy measures if the fuzzy words in a sentence pair belonged to the same fuzzy category. However, the fuzzy influence factor explores the concept of fuzzy words present in a sentence pair, but not belonging to the same fuzzy category.

The above evolution of the FUSE algorithm will further contribute towards research question ***RQ1. Investigate the feasibility of utilising Type-2 Fuzzy Sets and their representation of an individual's perception of fuzzy words and evaluate the suitability of the resulting fuzzy word models for incorporation into a Fuzzy Semantic Similarity Measure (FSSM).***

This is aiming to incorporate as many characteristics as possible of a sentence structure (such as the presence of negation), rather than rely on similarity ratings from WordNet.

6.2 - Linguistic Hedges Overview

The FUSE_1.0 algorithm ignored the presence of hedge words and simply used WordNet (Miller, 1995) to obtain a value for any hedge words, since hedges were never incorporated into the FUSE_1.0 algorithm. Therefore, to bring the correlations of the AHR higher, the presence of hedges were explored in sentence structures and how they could be incorporated into the FUSE_1.0 algorithm.

A linguistic variable carries with it the concept of fuzzy set qualifiers, called hedges. A hedge is a marker of uncertainty in language. Hedges are terms that modify the shape of FS. They include adverbs such as *very*, *somewhat*, *quite*, *more or less* and *slightly* (Negnevitsky, 2005). Linguistic variables represent crisp information in a form and precision appropriate for the problem. Linguistic variables associate a linguistic condition with a crisp variable. A crisp variable is the kind of variable that is used in most computer programs, often referred to as an absolute value. A linguistic variable, on the other hand, has a proportional nature, where in all of the software implementations of linguistic variables, they are represented by fractional values in the range of 0 to 1 (Banks, 2003).

Hedges can modify verbs, adjectives, adverbs, or even whole sentences. They are used as:

- All-purpose modifiers, such as *very*, *quite* or *extremely*
- Truth-values, such as *quite true* or *mostly false*
- Probabilities, such as *likely* or *not very likely*
- Quantifiers, such as *most*, *several* or *few*
- Possibilities, such as *almost impossible* or *quite possible*.

Hedges act as operations themselves. For instance, *very* performs concentration and creates a new subset. For example, taking a simple example (*short lady*), hedges could be applied such as (*very short lady*, *somewhat short lady*, *slightly short lady*). Therefore, the set of (*slightly short lady*) is broader than the set of (*short lady*) (Negnevitsky, 2005).

Figure 17 (Source: [adapted from] Negnevitsky, 2005) gives an application of hedges using the *short lady* example above inspired by (Negnevitsky, 2005). The fuzzy words *short*, *average* and *tall* are modified mathematically by the hedge word *very*. Considering the example of a lady who is 150 cm tall is a member of the short women set with a degree of membership of

0.5. However, she is also a member of the set of very short women with a degree of 0.2, which is reasonable.

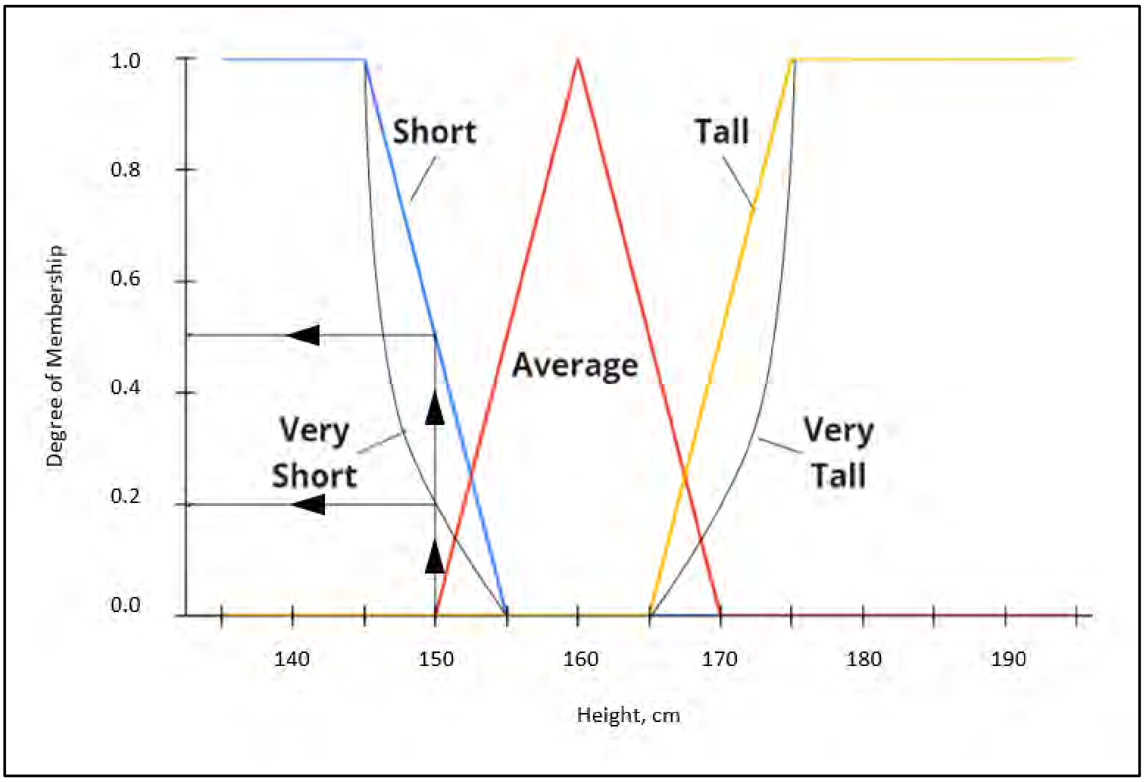


Figure 17 - Fuzzy Sets with Very Hedge Example (Source: [adapted from] Negnevitsky, 2005)

Hedges are useful as operations, but they can also break down continuums into fuzzy intervals. For example, the following hedges could be used to describe temperature: *very cold, moderately cold, slightly cold, neutral, slightly hot, moderately hot* and *very hot*. Obviously, these fuzzy sets overlap. Hedges help to reflect human thinking, since people usually cannot distinguish between *slightly hot* and *moderately hot* (Negnevitsky, 2005). According to Zadeh (Zadeh, 1975b), a linguistic variable is a variable of which the values are words or sentences in a natural or artificial language, e.g., Age. It would be a linguistic variable if its values were linguistic rather than numerical, so if Age = [*young*, *not so young*, *very young*, … , *old*, *not very old*, *not very young*], as opposed to Age = [20, 21, 22, … 60, 61, …].

A linguistic variable is characterised by a quintuple (*L*, *T*(*L*), *U*, *G*, *M*) where:

L is the name of the linguistic variable, *T*(*L*) is the term set of *L* (collection of linguistic values), *U* is the universe of discourse, *G* is a *syntactic rule* which generates the terms in *T*(*L*), *M* is a *semantic rule* which associates with each linguistic value *X,* its meaning *M*(*X*) where *M*(*X*) denotes a fuzzy subset of *U*.

Referring to the previous example of *petite lady*, using the *very* operation, will narrow a fuzzy set down, and so it will reduce the degree of membership of fuzzy elements (Negnevitsky, 2005). This can be calculated using a mathematical square as follows:

$$\mu_A^{very}(x) = [\mu_A(x)]^2$$

Equation 18 (Source: Negnevitsky, 2005)

Where *X* is the universe of discourse, *A* is a crisp subset of *X*, and $\mu_A(x)$ is the membership function of element x in the subset *A.* Thus, if someone had a 0.84 membership in the set of *petite lady*, then they will have a 0.7056 membership in the set of *very petite lady*. Applying this formulation to *very* means that the membership function is intensified. Table 19 is a mathematical representation of some of the common hedge words taken from (Negnevitsky, 2005).

Hedge Word	Mathematical Equation
A little	$[\mu_A(x)]^{1.3}$
Slightly	$[\mu_A(x)]^{1.7}$
Very	$[\mu_A(x)]^2$
Extremely	$[\mu_A(x)]^3$
Very very	$[\mu_A(x)]^4$
More or less	$\sqrt{[\mu_A(x)]}$
Somewhat	$\sqrt{[\mu_A(x)]}$
Indeed	$2[\mu_A(x)]^2$ if $0 \leq \mu_A \leq 0.5$ $1 - 2[1 - \mu_A(x)]^2$ if $0.5 \leq \mu_A \leq 1$

Table 19 - Mathematical Representation of Hedges (Source: Negnevitsky, 2005)

The measure values shown in Table 19 are to be applied to fuzzy control problems, but the values do not have the desired effect when applied to natural language problems. This is best explained using an example. The fuzzy word *hot* from the fuzzy category *Temperature* of FUSE_1.0 where [hot = 0.619377] on a scale of [0 , 1]; if we are to use the measure given in Table 19 to calculate the word *very hot* then (*very hot* → $[0.6193]^2 = 0.3836$) on a scale of [0 , 1]. A human will naturally understand the word *very hot* means it has a higher value than *hot*. However, mathematically it shows a lower value for *very hot* (Zadeh, 1999; Negnevitsky, 2005)*,* which means the measure values in Table 19 cannot be used correctly when applied to natural language. Therefore, another approach is needed to tackle the use of hedge words in natural language, when using the FUSE_1.0 algorithm.

The work in this thesis relating to Linguistic Hedges (Section 6.2) was published in **Human Hedge Perception – and its Application in Fuzzy Semantic Similarity Measures**. N. Adel, K. Crockett, A. Crispin, JP. Carvalho, D. Chandran 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). New Orleans, USA, 2019. DOI: 978-1-5386-1728-1/19/

6.2.1 - Capturing Human Perception of Hedges - A Study

The aim of this study is to investigate the effect of inclusion of hedge modifiers within the similarity calculation of fuzzy semantic similarity measures. The hypothesis for this experiment is:

H_0: Will the inclusion of hedges improve the precision of the similarity measurement in FUSE_1.0 through obtaining a higher collaboration of similarity with human ratings.

The experiment was conducted in two parts to investigate this hypothesis:

The first part required obtaining human perceptions, to see what the intensity of a hedge word is on a fuzzy word, in order to create a fuzzy ontology for hedges. Fuzzy intensity in this research refers to the perceptive numerical measure a word is given, be that measure positive (such as *very satisfactory*), or negative (such as *below adequate*) as rated by a human.

The second part entails testing the collected ratings for the hedge words on a dataset and investigating the comparison to STASIS (Li et al., 2003), a traditional SSM which solely relies on WordNet for its measures without catering for any fuzzy words or hedge words.

To complete Part 1 of the experiments, a set of the most common hedges were identified in natural language (Negnevitsky, 2005), where these words were not already present in any of the six fuzzy categories of FUSE_1.0. Let the fuzzy subset Hedges = [Below, Approximately, Neighbouring, Roughly, About, Around, Quite, Indeed, Definitely, Positively, Very, Above]. Six fuzzy words were selected from the six original categories proposed in FUSE_1.0 as follows: [Adequate (*Level of Membership*), Satisfactory (*Worth*), Middle-Aged (*Age*), Mild (*Temperature*), Fair (*Frequency*), Average (*Size/Distance*)].

These fuzzy words were chosen by selecting the word with the value closest to 0 in each category on a scale of [-1, 1]. Once human perceptions were captured, they could be used to construct Interval Type-2 models, similar to those used in FUSE_1.0 and used to derive a hedge ontology. The ontology would be used to determine the path length and depth between words as part of the word component similarity measures in FUSE_1.0. The path length and depth of hedge words are relative to their position in the hedge ontology, where each hedge category is treated as a concept. Each concept is constructed using a taxonomy (binary tree) where the root node always takes the value 0.

Defuzzified hedge words are then placed into tree nodes at intervals of ± 0.2. From the hedge taxonomy, the path length and depth of the Lowest Common Subsumer can be determined for hedge words in a category. This would allow the defuzzified hedge value to influence its associated defuzzified fuzzy word values, in terms of intensity, be this positively in that the sentence similarity value increased or negatively in that the sentence similarity value decreased.

To complete Part 2 of the aforementioned experiments, the required hedge values would need to be tested on a dataset, to determine if it helped the FUSE_1.0 algorithm to obtain a higher correlation with AHR. The results were compared to STASIS (Li et al., 2003), a traditional SSM which solely relies on WordNet for its measures and does not cater for fuzzy words or hedge words.

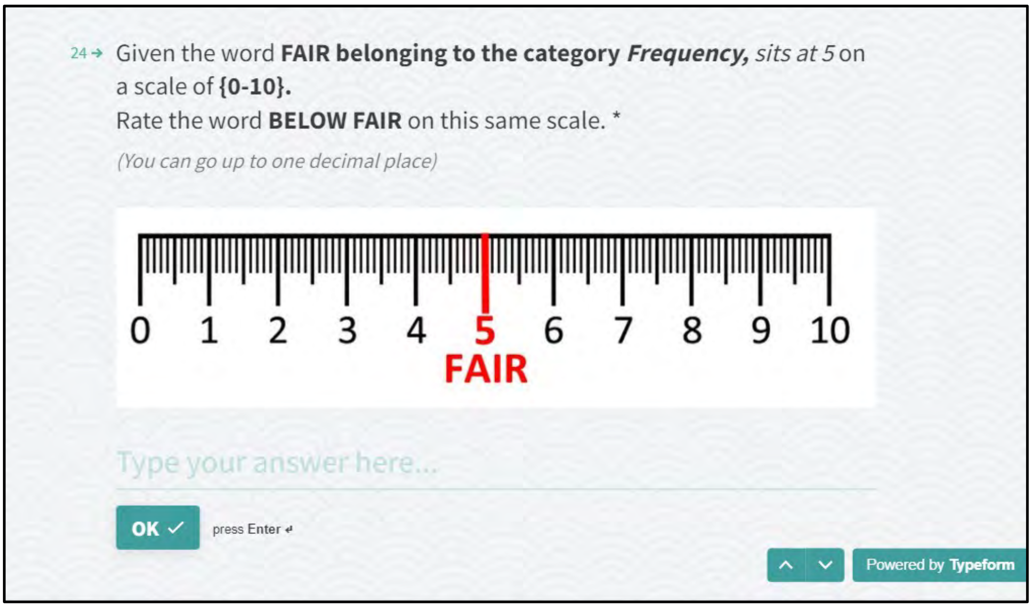
6.2.2 - Hedge Intensity Experiment - Part 1

To determine the intensity of hedges when applied to fuzzy words, 32 participants consented to take part in this study, all were native English speakers above the age of 18 from the Northwest region of England, United Kingdom. In total, there were 12 hedge words that were not already present in the FUSE_1.0 fuzzy dictionary that had mathematical definitions. When the mathematical value of a hedge word, (such as *very*) as defined in Equation 18 was applied to a fuzzy word, it did not represent the mathematical model that was linguistically represented. Therefore, a different approach was needed to cater for hedge words. As explained in Section 6.2, the hedge word *very* has a mathematical measure as shown in Equation 18, where x is the fuzzy value. Therefore, taking the word *Hot* = 0.6193, and computing the phrase *very hot* = $(0.6193)^2$ = 0.3836, calculated the mathematical value of *very hot* to be smaller than the mathematical value of *hot*, whereas linguistically *very hot* has a more positive intensity than *hot*. Therefore, a different approach to measuring the intensity was required that required the perceptions of humans.

To achieve this approach to measure the intensity from the human participants, the subset of 12 hedge words, were each added prior to the fuzzy words selected. One word from each of the 6 categories represented in FUSE_1.0. The middle word in each category with the value closest to zero was selected, and a random hedge word was added to the beginning of each of these six words (provided it still made sense). Participants were given a description of the task first, which included a simple linguistic definition of a hedge and a fuzzy word. An extract from the experiment description is as follows:

> *"The aim of this experiment is to help contribute towards computer systems that will understand the English language. This experiment is about HEDGES. Hedges are terms that modify the shape of a sentence. They include adverbs such as very, somewhat, quite, more or less and slightly. In this experiment, I am going to give you 6 words belonging to 6 categories. A category in this instance is just the name given for a group of words that fall under a similar meaning. For instance, for the category TEMPERATURE, it will contain words such as [hot, cold, mild, boiling, scorching, freezing…]. I am going to give you a scale of 0 to 10. Each word sits in the middle of this scale (5). I am going to pair each word with some hedge words and would like you to tell me where these new words would sit on this scale. You can use one decimal place (e.g. 3.2) for finer precision."*

An image of a ruler was used as a visual aid to make understanding the word placement visually easier as shown in Figure 19. The chosen word from each category was always located at mark 5 on the ruler and highlighted in red. The participants were then asked to rate the new hedge word when applied to the fuzzy word on this ruler, between a scale of [0, 10] with 1 decimal place permitted for accuracy. This extra decimal place is to keep in line with the previous scale used for the collection of fuzzy word ratings explained in section 4.4.3. An example of a word used in Figure 18 experiment is the hedge word *Below*. Taking the fuzzy word *Fair*, belonging to the category *Frequency*, one participant felt that the word *Below Fair* would be represented by a value of 3.5 as shown in Figure 19. Their opinion was that the hedge, *Below*, negatively reduced the intensity of the category word *Fair*.



24 → Given the word **FAIR belonging to the category *Frequency,*** *sits at 5* on a scale of **{0-10}.**
Rate the word **BELOW FAIR** on this same scale. *
(You can go up to one decimal place)

Figure 18 - Scale for Hedge Intensity Experiment

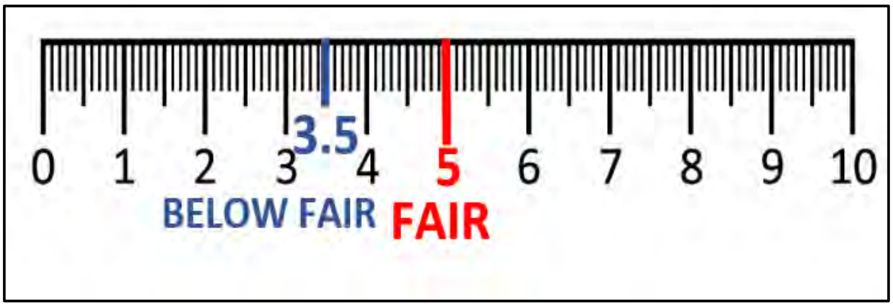


Figure 19 - Sample Answer for Below Fair

The aim of the hedge intensity experiment was to try and mimic the perceptions of humans using natural language, despite them not actually thinking about words on a scale. On obtaining all human measurements, the average value for each hedge word was calculated and this was normalised on a scale of [-1, 1] to create a hedge ontology as shown in Figure 20. This was done to match the same scale and ontological structure as the words in the fuzzy dictionary used within FUSE_1.0, and the fuzzy dictionary for the Hedge category is shown in Table 20.

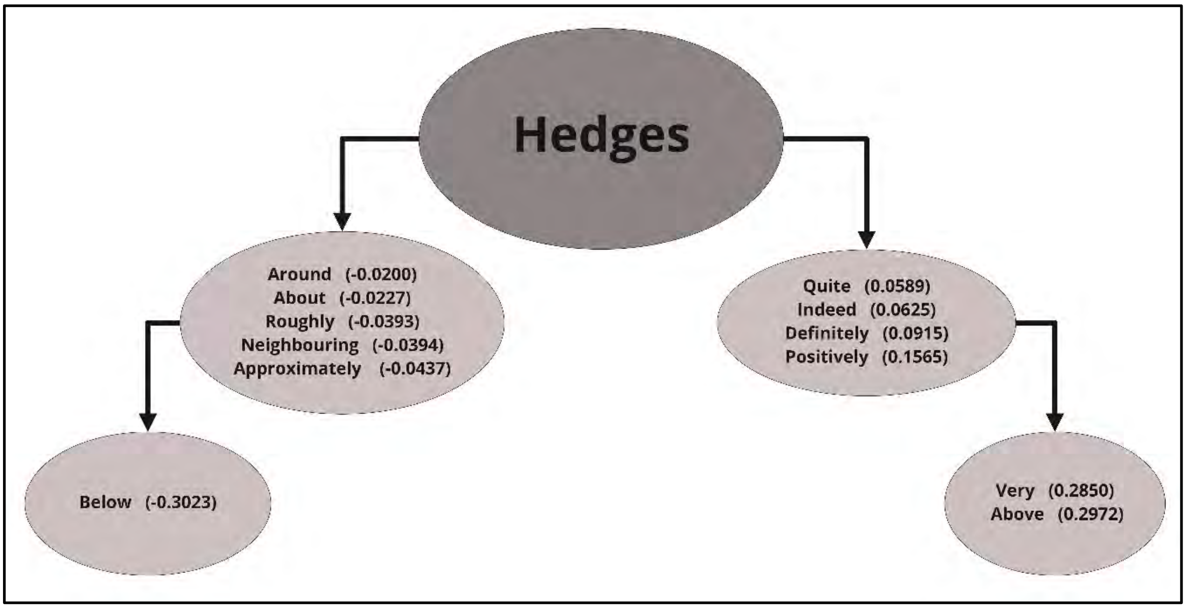


Figure 20 - Hedge Ontology Structure

HEDGES	
Word name	**Defuzzified Value**
BELOW	-0.3023
APPROXIMATELY	-0.0437
NEIGHBOURING	-0.0394
ROUGHLY	-0.0393
ABOUT	-0.0227
AROUND	-0.0200
QUITE	0.0589

INDEED	0.0625
DEFINITELY	0.0915
POSITIVELY	0.1565
VERY	0.2850
ABOVE	0.2971

Table 20 - Hedge Values

6.2.3 - Hedge Intensity Experiment - Part 2

To assess the intensity of hedges in the natural language context, it was necessary to compute the sentence similarity between pairs of sentences, which contained hedge words. Following analysis, it was established that the fuzzy sentence benchmark datasets, known as SWFD and MWFD (Chandran, 2013), did not contain a sufficient number of hedge words in order to conduct a rigorous evaluation.

Therefore, a dataset containing 16 sentence pairs containing hedge words was created. The methodology comprised of randomly extracting 16 sentences pairs from the MWFD (Chandran, 2013) ranging from high to low similarity based on human ratings (Chandran, 2013). For each fuzzy word in the Hedge Sentence Pair (HSP) dataset, a hedge word was assigned prior to that fuzzy word with the help of English language experts to ensure the sentences still made sense, i.e., for HSP1 "*The little village of Resina is also situated approximately near the spot*", the hedge word *approximately* was added. Table 21 shows the full set of hedge sentence pairs from the HSP dataset. The words in red are the fuzzy words from the fuzzy dictionary of FUSE_1.0 and the words in blue are the hedge words added.

O'Shea et. al. (J. O'Shea et al., 2013) emphasized the importance of establishing rigorous methodology when obtaining human ratings of similarities between words and sentence pairs. This is especially true in relation to sample size, population distribution and the inclusion of calibration pairs, providing representation of the highest and lowest sentence similarity pairs within the dataset. Adopting this methodology, the experiment for Part 2 consisted of 16 participants who were all native English speakers above the age of 18 from a diverse range of backgrounds, but all from the Northwest region of England, United Kingdom. They were provided with the 16 HSP's dataset and were asked to rate each sentence pair on a scale of [0, 10], with 1 decimal place permitted for accuracy, based on how similar they were to each other. The scale of [0, 10] was adopted to be consistent with approaches used for earlier human ratings as described in Section 4.4.3.

Hedge Sentence Pairs	Sentence 1	Sentence 2
HSP 1	The little village of Resina is also situated approximately near the spot	He seems quite excellent man and I think him uncommonly pleasing
HSP 2	A little quickness of voice there is which definitely rather hurts the ear	The only living thing near was a very old bony grey donkey
HSP 3	It is as long again as approximately almost all we have had before	was scarcely less below warm than hers and whose mind -- Oh
HSP 4	A positively frosty youthful man	A indeed hot old man
HSP 5	A definitely thick juvenile man	A very little old man
HSP 6	Had you married you must have been definitely regularly acceptable	Had you married you must have been indeed always poor
HSP 7	So would roughly useless diminutive Harriet	So would indeed poor little Harriet
HSP 8	Have massive mercy on the above mediocre men	Have a little mercy on the below poor men
HSP 9	How positively marvellous middling Piccola must have been	How quite good poor Piccola must have been
HSP 10	Behold how definitely fine a matter an adjacent fire kindleth	Behold how approximately great a matter a little fire kindleth
HSP 11	We will not say how about small for fear of shocking the youthful ladies	We will not say how indeed near for fear of shocking the young ladies
HSP 12	What's the fine neighbouring pensionable man	What's the roughly good old man
HSP 13	And he laughed around almost dreadfully	And he laughed very rather unpleasantly
HSP 14	Yesterday's ruling is a positively great first step toward better coverage for	He said the court 's ruling was a very great first step toward better coverage

	poor Maine residents he said but there is more to be done	for poor Maine residents but that there was more to be done.
HSP 15	It is largely a quite sizeable story, said Turnbull smiling	It is roughly rather a long story, said Turnbull smiling
HSP 16	The eyes were full of a frosty and above frozen wrath a kind of utterly heartless hatred	The eyes were full of a frozen and around icy wrath a kind of utterly heartless hatred

Table 21 - HSP Dataset

6.2.4 - Results and Discussion

6.2.4.1 - Hedge Intensity Results

Table 22 shows the results of the Average Human Ratings (AHR) collected as a result of the Hedge Intensity Experiment in Part 1. The table shows the six words from the fuzzy dictionary categories (Fuzzy Words), and the chosen twelve hedge words (Hedge Words). It gives a (Total Average), which is the average of each hedge row, that is then normalised on a scale of [-1, 1] to match the rest of the values scaling in the fuzzy dictionary, ordered from low to high.

On examining the results, it can be seen that *Very Fair* is more positively intensified than Fair, and the results indicate this closely i.e., *Fair* = 0.085 and *Very Fair* = 0.285. The same applies to *Mild* = -0.2387 and V*ery Mild* = 0.285; thus, the hedge *Very* positively intensifies a fuzzy word between the ranges of [0.0462, … ,0.37]. An example of the effect of negative intensity is the hedge word *Below*, with *Below Fair* = -0.2173 and *Below Mild* = -0.5411, thus *Below* negatively intensifies a fuzzy word between the range of [-0.541, ... ,-0.2173].

Fuzzy Words Hedge Words	Adequate	Satisfactory	Middle-Aged	Mild	Fair	Average	Total Average	Scaled [-1 ,1]
Below	3.250	3.517	3.718	3.775	3.612	3.577	**3.489**	**-0.302**
Approximately	4.327	4.470	5.008	4.869	5.153	4.813	**4.781**	**-0.044**
Neighbouring	4.662	5.036	4.723	4.736	4.692	4.831	**4.803**	**-0.039**
Roughly	4.819	4.829	4.627	4.464	4.969	4.314	**4.804**	**-0.039**
About	5.033	4.864	5.124	4.623	5.055	4.856	**4.887**	**-0.023**
Around	4.840	4.763	4.990	4.789	4.624	4.824	**4.900**	**-0.020**
Quite	5.535	5.589	5.550	4.607	5.546	5.600	**5.294**	**0.059**
Indeed	5.813	5.660	5.933	4.887	6.220	5.160	**5.313**	**0.063**
Definitely	5.900	6.933	5.900	5.653	5.515	5.682	**5.457**	**0.092**
Positively	5.615	6.460	6.533	6.200	6.131	5.907	**5.782**	**0.157**
Very	6.856	7.053	6.913	5.156	6.806	6.625	**6.425**	**0.285**
Above	6.535	6.625	6.438	6.219	6.438	6.688	**6.485**	**0.297**

Table 22 - AHR for Hedge Intensities Results

6.2.4.2 - Hedge Sentence Pairs Results

Table 23 shows the average human ratings (AHR) obtained from the 16 participants who rated the HSP's as a result of the Hedge Intensity Experiment in Part 2. The 16 participants were different individuals from those who had taken part in the Hedge Intensity Experiment in Part 1 outlined in Section 6.2.2. Sentence similarity measurements are shown for FUSE_1.0 and for comparison the similarity is also shown for the measure STASIS (Li et al., 2003) which does not incorporate any fuzzy words or consider the presence of hedges. All values shown in Table 23 are on a scale of [0, 1].

Hedge Sentence Pairs	AHR	STASIS	FUSE_1.0
HSP 1	0.0313	0.2242	0.1936
HSP 2	0.0188	0.5353	0.6038
HSP 3	0.0375	0.3106	0.3246
HSP 4	0.4456	0.3333	0.6647
HSP 5	0.4556	0.6272	0.8617
HSP 6	0.5563	0.6672	0.9249
HSP 7	0.6144	0.6968	0.9620
HSP 8	0.6106	0.7384	0.8300
HSP 9	0.7531	0.8517	0.9068
HSP 10	0.7638	0.8784	0.9073
HSP 11	0.8138	0.9221	0.9747
HSP 12	0.7850	0.7627	0.9220
HSP 13	0.8850	0.4693	0.6570
HSP 14	0.9381	0.8888	0.8921

HSP 15	0.9406	0.9033	0.9242
HSP 16	0.9144	0.9963	0.9924

Table 23 - Comparison of AHR vs STASIS & FUSE_1.0

Looking at the *r-value* from Table 24, it is determined that the FUSE_1.0 algorithm gave a higher correlation with the AHR (*r* = 0.8028) compared to STASIS (*r* = 0.7959). This result demonstrates there is a positive effect on the correlation with human ratings, when taking into consideration the presence of hedges in sentence pairs. Although the correlation difference was not significant, the value still showed an improvement over STASIS, which proves that fuzzy hedge intensity does play an important role in sentence similarity. This small improvement can be attributed to several factors:

(i) The fact that only twelve hedge words were modelled

(ii) The coverage of the hedge words in the HSP dataset was limited

(iii) The number of human ratings was only 16 – (acceptable in the NLP community) but on the low end of the scale where 32 participants is typically recommended.

Algorithms **Dataset**	**STASIS** **(*r-value*)**	**FUSE_1.0** **(*r-value*)**
HSP	0.7959	0.8028

Table 24 - r-value Results - STASIS vs. FUSE_1.0

Table 25 which relates to HSP13 shows one example of a sentence pair taken from the HSP dataset (Table 21) with results for AHR, STASIS and FUSE_1.0 taken from Table 23 presented on a scale of [0, 1]. The hedges used in this example are around and very shown in blue. The fuzzy words in the sentence pairs are *almost* and *rather* belonging to the category *Level of Membership*, shown in red. STASIS ignores all fuzzy and hedge words and simply used WordNet for any measures and therefore similarity is low (STASIS = 0.4693), FUSE_1.0, on the other hand caters for both fuzzy words and hedge words. Therefore, it has a higher similarity rating (FUSE_1.0 = 0.6570) which is closer to the (AHR = 0. 885). The goal is for the returned value to be as close to that of the AHR. This has proven that fuzzy words and hedge words play an important role in the similarity rating of a short text. On the other hand, Table 26 which relates to HSP12 shows that (STASIS = 0.7627) has a closer rating to the (AHR = 0.7850) then (FUSE_1.0 = 0.9220) presented on a scale of [0, 1]. This is likely to be due to the human sample size being relatively small (J. O'Shea et al., 2013) and/or the variations of WordNet used in STASIS and FUSE_1.0, as WordNet is constantly being updated.

Hedge Sentence Pairs	Sentence 1	Sentence 2	AHR	STASIS	FUSE_1.0
HSP 13	And he laughed around almost dreadfully	And he laughed very rather unpleasantly	0.8850	0.4693	0.6570

Table 25 - Good Example of HSP

Hedge Sentence Pairs	Sentence 1	Sentence 2	AHR	STASIS	FUSE_1.0
HSP 12	What's the fine neighbouring pensionable man	What's the roughly good old man	0.7850	0.7627	0.9220

Table 26 - Bad Example of HSP

Looking at the Intra-Class Correlation Coefficient (ICC) in Table 27, for each of the algorithms it can be seen that for STASIS (*a* = 0.865) and for FUSE_1.0 (*a* = 0.867). Cicchetti gives the following guidelines for the interpretation of the ICC, referred to as Inter-Rater Agreement measures, also known as the *a-value* (Cicchetti, 1994):

- *a-value* < 0.40 - Poor.
- 0.40 >= *a-value* <= 0.59 - Fair.
- 0.60 >= *a-value* <= 0.74 - Good.
- 0.75 >= *a-value* <= 1.00 - Excellent.

Based on Cicchetti's guidelines, it can be concluded that the Intra-Class Correlation Coefficient is deemed as *Excellent* for both datasets.

Intra-Class Correlation Coefficient Matrix		
	STASIS **(*a-value*)**	**FUSE_1.0** **(*a-value*)**
HSP Dataset	0.865	0.867

Table 27 - ICC for STASIS vs. FUSE_1.0

6.2.5 - Recommendations for the Use of Hedges

This study looked at the application of linguistic hedges within fuzzy semantic similarity measures. The first part involved obtaining human intensity ratings for a small selection of hedges using fuzzy words. These hedges were then modelled using Interval Type-2 fuzzy sets for inclusion in the FUSE_1.0 fuzzy dictionary. The second part involved the creation of 16 hedge sentence pairs using the modelled hedges from Part 1 where 16 participants rated their similarity.

Although there was minor improvement on the similarity measurement correlation between average human ratings (AHR) and the fuzzy measure FUSE_1.0, it was not significant. This is mainly due to the number of hedges modelled and the number of participants involved in rating the hedge sentence pairs. However even with this small sample, linguistically modelled hedges have a positive effect on sentence similarity.

Looking back at the hypothesis for this experiment:

H_0: Will the inclusion of hedges improve the precision of the similarity measurement in FUSE_1.0 through obtaining a higher collaboration of similarity with human ratings.

It can be concluded that hedges do play a part in obtaining a higher collaboration of similarity with human ratings. Further work is required using a larger participant sample and more testing of the hedge category on larger datasets and comparison with other SSM's to compare correlations with AHR.

6.3 - Category Expansion of the Fuzzy Dictionary

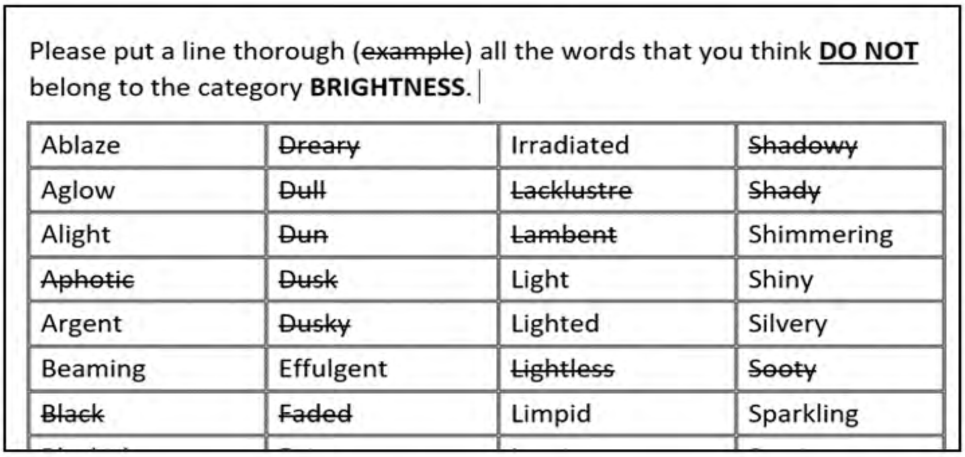
FUSE_1.0 consisted of six fuzzy categories (Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership) that made up the fuzzy dictionary as discussed in Chapter 4. Experiments conducted with FUSE_1.0 were explained in Chapter 5, and the results showed improvement over other SSM's. To further improve the coverage of fuzzy words, three new categories were introduced (Brightness, Strength, Speed) inspired by the work of Zadeh (Zadeh, 1999), where he mentions the *human ability to manipulate perception of (size, distance, weight, speed, time, direction, smell, colour, shape, force, likelihood, truth and intent, among others)*. This expansion of words is undertaken in the second version of the FUSE algorithm, referred to as FUSE_2.0.

To provide fuzzy ratings of words within each of these three new categories, a series of human experiments were needed to determine which words belong to each category and determine the rating value each word should have in a category.

6.3.1 - Collection of Words for New Categories

To determine which words should be placed in the three new proposed categories, the initial step was to first gather all words relating to each category. This was done using the Oxford English Dictionary (Oxford English Dictionary, 2021), to collect all one-word synonyms belonging to each category. Only one-word synonyms such as *Hasty* for the category *Speed* were used and any synonyms that had two or more words were not collected such as *Rattle Along*. Once this was done for all three categories, human participants were needed to determine which words should remain in the three new categories.

The recruitment process for collecting human participants was the same as that described in Section 4.4.3, and only native English speakers over the age of 18 from the Northwest region of England, United Kingdom were chosen to take part in the experiment. A total of 17 participants took part in this study and they were each given the words in each category and asked to cross out the words they felt did not belong to that category. Figure 21 shows a partial example of an answer sheet from one participant for the category *Brightness* and the words they have crossed out that they feel do not belong to this category.



Please put a line thorough (~~example~~) all the words that you think **DO NOT** belong to the category **BRIGHTNESS**.

Ablaze	~~Dreary~~	Irradiated	~~Shadowy~~
Aglow	~~Dull~~	~~Lacklustre~~	~~Shady~~
Alight	~~Dun~~	~~Lambent~~	Shimmering
~~Aphotic~~	~~Dusk~~	Light	Shiny
Argent	~~Dusky~~	Lighted	Silvery
Beaming	Effulgent	~~Lightless~~	~~Sooty~~
~~Black~~	~~Faded~~	Limpid	Sparkling

Figure 21 - Partial Participant Answer Sheet for Brightness Category

This was done for all three new categories. To calculate the results and determine which words should be kept in each category, based on these human experiments, the help of English language experts was sought and a threshold of 70% was established. Any word that was crossed out by 70% of the participants was removed from that category and the remaining words kept. Table 28 shows the results of this experiment, the first column shows the category labels, the second column represents the original number of words that were collected per category using the Oxford English dictionary. The final column shows the number of words that were kept in each category, as a result of this experiment after the clean-up process and applying the 70% threshold. The words per category are now determined and the next stage involves determining the fuzzy ratings per word.

Categories	Original No. of Words	Kept No. of Words
Brightness	107	27
Strength	109	24
Speed	81	26

Table 28 - Three New Categories

6.3.2 - Modelling of Fuzzy Words for New Categories

The same process as described in Section 4.4.3 was used to rate each of the words per category. 32 native English speakers from the Northwest region of England, United Kingdom were used per category and asked to rate each of the words on a scale of [0, 10]. The words were then modelled using the Interval Type-2 FS approach and normalised on a scale of [-1, 1] to keep consistent with the rest of the words in the other six categories of the fuzzy dictionary.

Table 29 shows the total number of fuzzy words present in each of the nine categories. The full fuzzy dictionary containing all words and their defuzzified values normalised on a scale of [-1, 1] from all nine categories can be found in Appendix C and the ontological structure for the three new categories can be found in Appendix D.

Categories	No. of Words
Size/Distance	91
Temperature	36
Age	42
Frequency	48
Level of Membership	31
Worth	61
Brightness	27
Strength	24
Speed	26

Table 29 - Fuzzy Words Per Category

6.3.3 - Testing FUSE_2.0

To evaluate the performance of FUSE_2.0 with the addition of the three new fuzzy categories, it was important to test its performance, compared to the AHR with the use of a number of datasets. The results will also need to be compared with the other SSM algorithms, to determine if the addition of the new fuzzy categories achieved a better correlation with the AHR.

6.3.3.1 - Experimental Methodology

To test the correlation of the FUSE_2.0 algorithm against human ratings and to see if the presence of the fuzzy dictionary with the nine categories helped improve the correlations with the AHR, FUSE_2.0 was ran on several datasets and also compared with other SSM's. The aim of the experiments was to test the following hypothesis:

H_0: FUSE_2.0 gives a higher correlation with human ratings compared to other SSM's.

In order to test the *H_0*, FUSE_2.0 was ran against each of the five datasets (FUSE-62, SWFD (Chandran, 2013), MWFD (Chandran, 2013), STSS-65 (J. O'Shea et al., 2013) and STSS-131 (J. O'Shea et al., 2013)) and the sentence similarity results for each Sentence Pair [SP] was recorded. To test the improvement of FUSE_2.0, all five datasets were also run with STASIS (Li et al., 2006), Dandelion Semantic (SpazioDati, 2015), Dandelion Syntactic (SpazioDati, 2015) and SEMILAR (Rus et al., 2013a) algorithms and the sentence similarity results for each SP was recorded.

The above-mentioned algorithms were not designed to capture human perception-based words within short texts through relation to the context in which they were used, therefore comparing their performance with FUSE_2.0 will build a better picture, as to why a fuzzy SSM is needed to cater for the uncertainty of fuzzy words in a sentence or utterance.

Using Pearson's correlation coefficient (Kent State University, 2013), the correlation for each dataset was compared to the Average Human Ratings [AHR]. Pearson's correlation provides statistical evidence for a linear relationship between two variables *x* and *y* and can be computed as shown in Equation 19 (Kent State University, 2013):

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)} \, . \sqrt{\mathrm{var}(y)}}$$

Equation 19 (Source: Kent State University, 2013)

Where r_{xy} is the correlation coefficient, cov(*x*, *y*) is the sample covariance of *x* and *y*; var(*x*) is the sample variance of *x*; and var(*y*) is the sample variance of *y* (Kent State University, 2013).

6.3.3.2 - Datasets

Five datasets were used in total containing both fuzzy sentence pairs, and non-fuzzy sentence pairs. A full breakdown of these datasets is given in Table 30. The published gold-standard datasets STSS-65 and STSS-131 (J. O'Shea et al., 2013) did not contain any fuzzy words and the published datasets SWFD and MWFD (Chandran, 2013) contained limited fuzzy words, therefore, to fully cover the words in all nine fuzzy categories a new dataset called FUSE-62 was designed with the help of English language experts.

Dataset	Description	Fuzzy / Non-Fuzzy	Readers Age
FUSE-62	62 sentence pairs specifically designed by English language experts to contain fuzzy words from all nine categories of FUSE_2.0	Fuzzy	14-15 yrs. old
SWFD	30 sentence pairs containing one fuzzy word per sentence	Fuzzy	10-11 yrs. old
MWFD	30 sentence pairs containing two or more fuzzy word per sentence	Fuzzy	College graduate
STSS-65	65 Gold standard sentence pairs	Non-Fuzzy	10-11 yrs. old
STSS-131	131 Gold standard sentence pairs	Non-Fuzzy	8-9 yrs. old

Table 30 - Datasets Breakdown & Feasibility

The FUSE_62 dataset consisted of the 32 original sentence pairs from the MWFD dataset, and an additional 30 sentence pairs taken from the STSS-131 dataset, a constraint was to ensure there were an equal number of sentence pairs in the *low*, *medium* and *high* categories, as identified by human participants in previous published studies (J. O'Shea et al., 2013). Using the help of the English language expert, to ensure sentences still held an expressive meaning, fuzzy words from the three new proposed fuzzy categories (Brightness, Speed and Strength) were carefully added to the 30 sentence pairs taken from STSS-131 to make up 62 sentence pairs that would cover words from all nine of the fuzzy categories of FUSE_2.0.

The readers' age is obtained after examining the contents of each dataset and performing a feasibility test (Automatic Readability Checker, 2022). This feasibility is important because it influences how clearly a text can be understood by the reader. By making text as clear to understand as possible, this will allow the improvement on the participant selection (Text

Inspector, 2020). The Readers Age will help determine what the minimum age of participant recruitment can be, with the confidence that the text will be understood correctly. This ensures any ratings given, are the result of an individual's perception and not down to confusion or misunderstanding the contents of the text.

6.3.3.3 - Results and Discussion

Table 31 shows the Pearson's correlation results on a scale of [0, 1] of the five datasets with the different SSM algorithms tested. It can be seen from the results that FUSE_2.0 gave a higher correlation for each dataset compared to all the other algorithms tested, as highlighted in red.

Pearson Correlation of Results	**STASIS** (*r-value*)	**Dandelion Semantic** (*r-value*)	**Dandelion Syntactic** (*r-value*)	**SEMILAR** (*r-value*)	**FUSE_2.0** (*r-value*)
FUSE-62	0.543	0.546	0.312	0.533	0.544
SWFD	0.645	0.433	0.577	0.627	0.688
MWFD	0.745	0.629	0.736	0.758	0.768
STSS-65	0.681	0.537	0.620	0.661	0.690
STSS-131	0.502	0.406	0.152	0.491	0.518

Table 31 - FUSE_2.0 vs SSM's Correlation Results

Figure 22 shows a graphical representation of the results from Table 31 showing FUSE_2.0 achieving the highest correlations with human ratings for all datasets tested, compared to the other five SSM's. It can be seen from the results in Table 31 that the dataset containing the greatest number of fuzzy words (MWFD) gave the highest correlation (0.768) and the dataset with no fuzzy words STSS-131 gave the lowest correlation (0.518). This result showed that the more fuzzy words present in a sentence pair, the better the FUSE_2.0 algorithm performs. This has further highlighted the need to consider the presence of fuzzy words on sentence similarity.

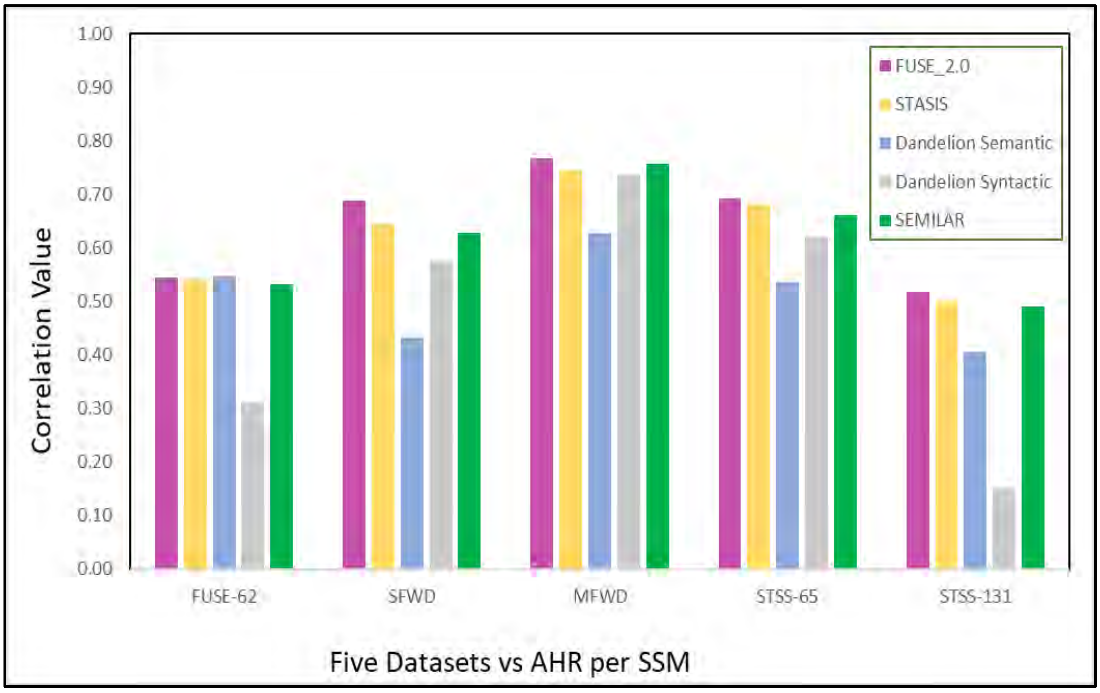


Figure 22 - Results Comparison for Five Datasets

Conduction of an Intra-Class Correlation Coefficient (ICC) (Koo and Li, 2016) also produces some positive results. ICC is important in a study as it represents the extent to which the data collected in the study is correct and a good representation of the variables measured.

Cicchetti gives the following guidelines for the interpretation of the ICC, referred to as Inter-Rater Agreement measures, also known as the *a-value* (Cicchetti, 1994):

- *a-value* < 0.40 - Poor.
- 0.40 >= *a-value* <= 0.59 - Fair.
- 0.60 >= *a-value* <= 0.74 - Good.
- 0.75 >= *a-value* <= 1.00 - Excellent.

looking at the *a-value* in Table 32 it can be seen that that four of the datasets (FUSE-62, SWFD, MWFD, STSS-65) show an *Excellent* rating based on the *a-value*. It can further be shown that the more fuzzy words present in a dataset, the higher the *a-value*. This can be seen in MWFD dataset with the *a-value* being the highest of all datasets (*a* = 0.947); this is because the MWFD has two or more fuzzy words present per sentence pair. The *p-value* is the standard method that is used in statistics to measure the significance of empirical analyses (Fenton and Neil, 2018). The *p-value* for four of the datasets (FUSE-62, SWFD, MWFD, STSS-65) is < .001 which is less than 0.05, making it statistically significant and provides support for our research hypothesis H_0 which strongly suggests that the expansion of the fuzzy dictionary and the introduction of a fuzzy ontology effects the level of similarity.

Looking at both the *a-value* (second column) and the *p-value* (fourth column) in Table 32, it can be seen that the dataset that held the highest number of non-fuzzy words (STSS-131), is the dataset that gave the lowest *a-value* result (*a* = 0.104), which is deemed as *Poor* according to Cicchetti and the *p-value* was rejected.

The result concluded that the more fuzzy words present in a dataset, the higher the *a-value*, which in turn means FUSE_2.0 performs better when more fuzzy words are present in a sentence or utterance. Most SSM's use WordNet (Miller, 1995), and since WordNet is constantly being improved, results can vary over time, therefore if this experiment was to be repeated again at a later date, results may vary slightly.

Inter-Rater Correlation Results	**a-value**	**Cicchetti Measure**	**p-value**	**Accept or Reject**
62 SP	0.872	Excellent	*p* < .001	Accept
SWFD	0.911	Excellent	*p* < .001	Accept
MWFD	0.947	Excellent	*p* < .001	Accept
STSS-65	0.883	Excellent	*p* < .001	Accept
STSS-131	0.104	Poor	0.199	Reject

Table 32 - (a-value) & (p-value) for FUSE_2.0

6.3.4 - Discussion

The experiments and results show that fuzzy words must be considered when looking at semantic similarity measures as they play a significant role in the similarity of sentences.

Looking back at the experiments conducted on the five datasets using the five algorithms, FUSE_2.0, STASIS (Li et al., 2006), Dandelion Semantic (SpazioDati, 2015), Dandelion Syntactic (SpazioDati, 2015) and SEMILAR (Rus et al., 2013a) and the original null hypothesis (*H_0: FUSE_2.0 gives a higher correlation with human ratings compared to other SSM.*), it can be concluded that H_0 can be accepted based on both the *a-value* and the *p-value* results shown in Table 32 for a confidence level of 95%.

6.4 - Logic Negation in Natural Language

Logic negation in natural language plays an important role as it can often change the polarity of a sentence from a positive one to a negative one and vice versa (Singh and Paul, 2021). For example, the two sentences S_1 and S_2:

S_1: [This food was really worth waiting for]

*S_2: [This food was really **not** worth waiting for]*

Both S_1 and S_2 have the same number of words in the same order with the only difference being the presence of the word ***not*** in S_2 before the word *worth* which completely changes the meaning of S_2 from a positive experience to a negative one.

On the other hand, referring to the following two sentences S_3 and S_4:

S_3: [The season finale was predictable]

*S_4: [The season finale was **un**predictable]*

Here, S_3 has a negative meaning in the context of a show having a *predictable* ending, but S_4 (again same number of words in the same order) with the presence of the word ***un*** takes this negative experience and turns it into something positive in the context, that the season finale of the show was actually *unpredictable*, meaning it was exciting and amusing.

The '*not*' operator has not been considered specifically within SSM measures, yet it is an important concept as illustrated by the two examples ([S_1 ,S_2] *and* [S_3 ,S_4]). Several specific complement (not) operators relating to fuzzy sets have been defined, which will be briefly reviewed. There are many different approaches to dealing with negation operators given by respected scholars and mathematicians, a selection of which will be briefly discussed below.

6.4.1 - Not Logical Operators in Fuzzy Set Theory

The original '*not*' operator introduced by Zadeh is defined by taking one minus the membership value $\sim\mu_A(x) = (1 - \mu_1)$ at each point, along the truth function, with no additional parameters needed. Since the complement of a fuzzy set is often used as a new fuzzy region in a model, the '*not*' is produced by creating and populating a new fuzzy set (Cox, 1994).

The two most popular alternatives of negation aside from Zadeh are the Yager and Sugeno (Klir and Folger, 1988) weighted complement operators. Yager defines an alternative form of the fuzzy complement having a power function defined as below:

$$\sim\mu_A(x) = (1 - \mu_A(x)^k)^{\frac{1}{k}}$$

Equation 20 (Source: Klir and Folger, 1988)

Where the class function *k* is generally in the range [>0, <5]. The class function performs the standard Zadeh complement (which is found when *k*=1). The class membership in the Yager complement, provides a convenient and flexible method of adjusting the strength of the fuzzy '*not*' operator. For the endpoint conditions of zero and one, the Yager complement, regardless of the class strength parameter, always acts like the standard Zadeh complement (Cox, 1994).

The Sugeno complement, takes a class parameter that determines the strength of the negation. The Sugeno class is defined as:

$$\sim\mu_A(x) = \frac{1 - \mu_A\,(x)}{1 + k\mu_A\,(x)}$$

Equation 21 (Source: Klir and Folger, 1988)

In this case, the class parameters are in the range [-1, ∞]. When *k*=0 the Sugeno complement has the desirable property of becoming the standard Zadeh complement (Cox, 1994).

The impact of '*not*' and negation on the similarity of sentence measures was only fully understood after the integration of the FUSE algorithm into a dialogue system (Chapter 7). Therefore, full experimental design, results, and discussion of the use of '*not*' operators in the FUSE algorithm will be explored in Chapter 7 (section 7.4.4).

6.5 - Fuzzy Influence Factor

Currently the FUSE_3.0 algorithm calculates the semantic and syntactic similarity of a sentence pair, through a weighted combination of analysis on both the syntactic and semantic elements of a short text (Algorithm 1, Section 4.5.1). This calculation was done through using the nine fuzzy categories and the consideration of the presence of hedges. One weakness in FUSE_3.0 is the lack of consideration for sentence pairs, where fuzzy words are not in the same category; for example, comparing the word "*slow*" to "*normal*". While both these words do belong to fuzzy categories (*Speed* and *Worth* respectively), they do not fall in the same fuzzy category and so WordNet is used to derive their values.

To overcome this weakness the addition of a Fuzzy Influence (FI) Factor within the FUSE algorithm was introduced referred to as FUSE_4.0. The FI factor overcomes a weakness of FUSE_3.0 by ensuring fuzzy words not in the same fuzzy categories but within the same sentence pairs have a human associated impact on determining the overall sentence similarity. The FI for a sentence pair *sn*, can be defined as:

$$FI_{sn} = \frac{1}{n-i}$$

Equation 22

where *n* is the number of all the words in the sentence pair *sn*; and *n* > 0, and *i* is the count of all the fuzzy words in *sn*. If all the words in the sentence pair are fuzzy, i.e., *n* = *i*, we set FI_{sn} ≔1, and so FI_{sn} takes values between 0 and 1. FI is applied to all sentence pair calculations within FUSE_4.0, regardless of whether fuzzy words are in the same category or not. In (Adel et al., 2018), the FUSE_1.0 algorithm was first proposed to calculate the overall similarity between two fuzzy utterances, U_1 and U_2, through the weighted addition of syntactic and semantic components.

In FUSE_4.0, the overall similarity of S(U_1, U_2) is now calculated as:

$$S(U_1, U_2\,) = S_s * w1 + S_r * w2 + FIsn * w3$$

Equation 23

Where *w1*,*w2*,*w3* ∈ [0, 1] and ∑ *w1*..*w3* = 1, S_s is the semantic similarity and S_r is the syntactic similarity, calculated using pairs of semantic and syntactic similarity vectors which were determined by a word similarity measure and a short joint word vector set comprising of word frequency information and word order.

The work in relation to Fuzzy Influence Factor (Section 6.5) was published in **Fuzzy Influence in Fuzzy Semantic Similarity Measures.** N Adel, K Crockett, JP Carvalho, V Cross. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, 2021. DOI: 10.1109/FUZZ45933.2021.9494535 and was short listed for the best paper award.

6.5.1 - Experimental Methodology

To investigate the relationships between the semantic, syntactic and FI components, an empirical experiment was conducted for FUSE_4.0 to investigate if introducing a fuzzy influence factor will affect the overall sentence similarity rating and produce a closer value to that of the average human ratings (AHR). The hypothesis for this experiment is given below:

H_0 = The inclusion of a fuzzy influence (FI) factor in the calculation of the overall semantic similarity of a sentence improves the overall correlation when compared to human ratings.

6.5.2 - Metrics

In each set of experiments, three metrics (semantic, syntactic and fuzzy influence factor) are used to measure the effectiveness of variants in the FI factor within FUSE_4.0. The Pearson's Correlation Coefficient (Kent State University, 2013) is used to show statistical evidence for a linear relationship between two variables *x* and *y* in this work between the human ratings and those generated by FUSE_4.0 and is shown in Equation 24 (Kent State University, 2013):

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)} \,.\, \sqrt{\mathrm{var}(y)}}$$

Equation 24 (Source: Kent State University, 2013)

Where r_{xy} is the correlation coefficient, *cov(x, y)* is the sample covariance of *x* and *y*; *var(x)* is the sample variance of *x*; and *var(y)* is the sample variance of *y*.

The Intra-Class Correlation Coefficient (ICC) in a study represents the extent to which the data collected in the study is correct and a good representation of the variables measured.

Cicchetti gives the following guidelines for the interpretation of the ICC, referred to as Inter-Rater Agreement measures, also known as the *a-value* (Cicchetti, 1994):

- *a-value* < 0.40 - Poor.
- 0.40 >= *a-value* <= 0.59 - Fair.
- 0.60 >= *a-value* <= 0.74 - Good.
- 0.75 >= *a-value* <= 1.00 - Excellent.

The *a-value* is important as it shows the extent to which the data, that is collected for this study, is a correct representation of the variables measured; therefore, the aim is to achieve an *Excellent* rating to maximise reliability of the human ratings of the short text pairs (McHugh, 2012).

6.5.3 - Datasets

In this work, three datasets, FI-25, FUSE-62, and the Multi Word Fuzzy Dataset (MWFD) (Chandran, 2013) were used to investigate the fuzzy influence factor.

Initial work was undertaken on a test dataset called FI-25 which comprised of a set of 25 test sentences (with inclusion criteria defined below) and they were specifically chosen based on correlations with the AHR from 3 existing datasets, STSS-131 (J. O'Shea et al., 2013), FUSE-62 (Adel et al., 2020) and MWFD (Chandran, 2013) to test the proposed methodology before undertaking further experiments on larger datasets.

SP1 – SP15 of the Sentence Pairs (SP) in FI-25 consisted of poor human rating correlations when run on FUSE_1.0. These poor human rating correlations imply that the semantic similarity measurement of FUSE_1.0 was significantly different (higher or lower) than that of the average human ratings. Ideally, similarity values derived from the measure should be as close to the human ratings as possible. The remaining 10 pairs (SP16 – SP25) gave high correlations with human ratings by FUSE_1.0. This meant that the semantic similarity measurement of FUSE_1.0 were close to the ratings given by human ratings. This dataset was created to ensure that the impact of the fuzzy influence factor was assessed against both high and low correlations. The full FI-25 sentence pairs for this dataset can be seen in Table 33, where the red words show the fuzzy words that appear in the fuzzy dictionary of the FUSE algorithm. The methodology for collecting human ratings for these sentence pairs is the same as that used in Section 4.4.3 and Section 6.2.3 for consistency.

Sentence Pairs	Sentence 1	Sentence 2
SP1	Had you married you must have been regularly acceptable	Had you married you must have been always poor
SP2	They hint that all whales on- occasion smell amazing	They hint that all whales always smell bad
SP3	An unacceptable watcher and very dietetically pathetic is Dr Bunger	A great watcher and very dietetically severe is Dr Bunger
SP4	A little quickness of voice there is which rather hurts the ear	The only living thing near was an old bony grey donkey
SP5	An automobile is a fast car.	In legends and fairy stories, a wizard is a man who has flashing magic powers.
SP6	A grin is a light smile.	An implement is a tool or other piece of lighted equipment.
SP7	The coast is an area of land that is next to the leisurely sea.	A forest is a large swift area where trees grow close together.
SP8	An automobile is a fast car.	A cushion is a fabric case filled with swift material, which you put on a seat to make it more comfortable.
SP9	A crane is a large machine that moves heavy things by lifting them in the air.	A rooster is a tough adult male chicken.
SP10	The children crossed the road very safely and fast thanks to the help of the lollipop lady.	It was feared that the child might not make a dashing recovery, because he was seriously ill.
SP11	My bedroom wall is sunlit lemon coloured but my mother says it is yellow.	Roses can be different colours, it has to be said burning red is the best though.
SP12	Cord is strong, thick string.	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.
SP13	A mound of something is a large rounded pile of it.	A stove is a piece of equipment which provides heat, either for cooking or for heating a room.

SP14	A boy is a child who will grow up to be a man.	A sage is a person who is regarded as being very wise.
SP15	The little village of Resina is also situated near the spot	He seems an excellent man and I think him uncommonly pleasing
SP16	The eyes were full of a frosty and frozen wrath a kind of utterly heartless hatred	The eyes were full of a frozen and icy wrath a kind of utterly heartless hatred
SP17	She constantly travels with her own sheets an excellent precaution	She always travels with her own sheets an excellent precaution
SP18	This is just the latest movement in a continuing trend towards open source support of business applications	This is just the latest movement in a continuing trend toward open-source support among business application vendors
SP19	Yesterday's ruling is a great first step toward better coverage for poor Maine residents he said but there is more to be done	He said the court 's ruling was a great first step toward better coverage for poor Maine residents but that there was more to be done.
SP20	A crane is a large machine that moves heavy things by lifting them in the air.	An implement is a tool or other piece of lighted equipment.
SP21	A furnace is a container or enclosed space in which a very blazing hot fire is made, for example to melt metal, burn rubbish or produce steam.	A stove is a piece of equipment which provides radiant heat, either for cooking or for heating a room.
SP22	Cord is a strong, thick string.	String is a delicate thin rope made of twisted threads, used for tying things together or tying up parcels.
SP23	A grin is a beaming broad smile.	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly and alight.
SP24	In former times, serfs were a class of people who had to work hardy on a particular person's land and could not leave without that person's permission.	A slave is someone who is the property of another person and has to work tough for that person.
SP25	When you make a journey, you travel from one place to another in a gradual manner.	A voyage is a long leisurely journey on a ship or in a spacecraft.

Table 33 - FI-25 Dataset

The FUSE-62 dataset was specifically designed with the help of English language experts to contain fuzzy words from all nine categories from the FUSE_2.0 fuzzy dictionary. The reader's age for this dataset has been calculated as 14-15 years old (Ninth to Tenth graders) using the Automatic Readability Checker (Automatic Readability Checker, 2022). The MWFD (Chandran, 2013) contains 30 sentence pairs where each sentence contains two or more fuzzy words per sentence. The reader's age for this dataset has been calculated as college graduate using the Automatic Readability Checker (Automatic Readability Checker, 2022).

6.5.4 - Experimental Results

For each of the experiments in this section, the following experimental methodology was followed, where the semantic, syntactic and FI weights were each separated, and changed using increments of 0.05 between the ranges of 0 and 1. In each case, one of the weights was fixed, whilst the other pair was changed to ensure the sum of all weights was always 1. At each iteration, Pearson's Correlation (Kent State University, 2013) was recorded each time to see which set of values gave the best results.

6.5.4.1 - Experiment 1 - FI on FI-25 Dataset

The FI factor within FUSE_4.0 was used with a range of different empirical weighting values for the semantic, syntactic, and fuzzy influence factor to see which gave the sub-optimal results. Optimal results are calculated by comparing Pearson's Correlation (Kent State University, 2013) (*r-value*) with human ratings. The higher the *r-value,* the closer the ratings to those of humans. In F1-25, the correlation was calculated for both the '*bad'* performing sentence pairs (*NPW*) (SP1-SP15) as well as the '*good'* performing sentence pairs (*PW*) (SP16-SP25), where *bad* and *good* results were generated by FUSE_1.0. Pearson's Correlation (Kent State University, 2013) of FUSE_4.0 is also compared with earlier versions of FUSE (FUSE_2.0 and FUSE_3.0) as well as four other similarity algorithms that do not cater for fuzzy words: STASIS (Li et al., 2006), SEMILAR (Rus et al., 2013a), Dandelion API Semantic (SpazioDati, 2015) and Dandelion API Syntactic (SpazioDati, 2015).

Table 34 shows the correlation findings for the five experiments (*Exp. 1.1*, *Exp. 1.2*, *Exp. 1.3*, *Exp. 1.4*, *Exp. 1.5*) ran on the (*NPW*) (SP1-SP15) sentence pairs shown on a scale of [0, 1].

Results from Table 34 show that the measures (Sem 0.5, Syn 0.2, FI 0.3) from *Exp. 1.5* gave the best overall correlation and the highest correlation, giving better results than the other algorithm measures with the exception of API Syn for *NPW* SP's as shown in Table 35. The higher the correlation, the closer the similarity ratings are to those of the human ratings (HR).

Figure 23 shows a scatter plot for the relationship between the two variables (Yi, 2019). In this instance, the two variables are the human ratings (HR) and the correlation following the fuzzy influence factor (FI) experiment. Each dot on the scatter plot shows the values for each sentence pair on the X and Y axis, with *x* being FI and *y* being HR. The scatter plot in Figure 23 shows the positive correlation of the human ratings (HR) with the fuzzy influencer factor (FI), each on a scale of [0, 1], where 0 represents no similarity and 1 represents maximum similarity. For *Exp. 1.5* for the 15 NPW SP's, the line of best fit shows the mathematically best fit for the data; also referred to as the '*trendline*'. This line shows the behaviour of a set of data, when the line goes up, this shows a positive linear relationship between the variables.

Pearson's Correlation	***r-value***	***r-value***	***r-value***	***r-value***	***r-value***
	Exp. 1.1 Sem 0.8 Syn 0.1 FI 0.1	**Exp. 1.2 Sem 0.7 Syn 0.1 FI 0.2**	**Exp. 1.3 Sem 0.75 Syn 0.15 FI 0.1**	**Exp. 1.4 Sem 0.7 Syn 0.05 FI 0.25**	**Exp. 1.5 Sem 0.5 Syn 0.2 FI 0.3**
HR vs FUSE_4.0	0.6953	0.7068	0.7174	0.6873	0.7711

Table 34 - Hyper-Parameter Optimisation for NPW SP's

SSM	**r-value**
HR vs FUSE_4.0	0.7711
HR vs FUSE_2.0	0.6817
HR vs FUSE_3.0	0.7060
HR vs STASIS	0.7126
HR vs API Semantic	0.4953
HR vs API Syntactic	0.8840
HR vs SEMILAR	0.7659

Table 35 - Comparison of SSM Best Results from Table 34

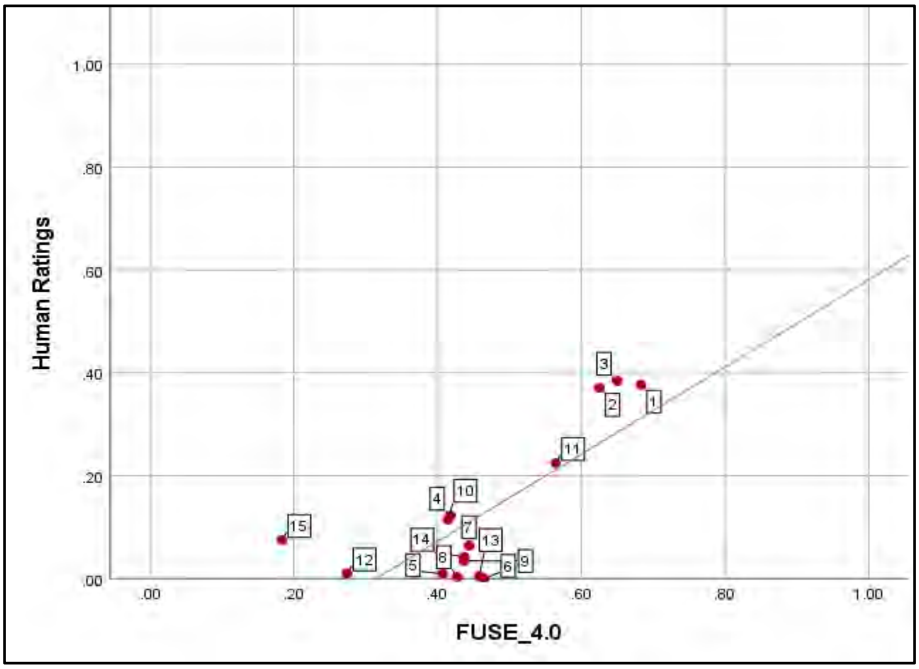


Figure 23 - NPW Dataset Scatter Plot (Sem 0.5, Syn 0.2, FI 0.3)

SP15 where SP15a = *'The little village of Resina is also situated near the spot'* and SP15b = *'He seems an excellent man and I think him uncommonly pleasing'*, is a clear outlier with the average human rating being (0.075), where FUSE_4.0 gave a measure of (0.206). SP15 contains fuzzy words [*little*, *near*, *excellent,* and *uncommonly*], *little* and *near* belong to the *Size/Distance* category, *excellent* belongs to *Worth* category and *uncommonly* belongs to *Frequency* category.

Table 36 shows the correlation findings for the five experiments (*Exp. 1.6*, *Exp. 1.7*, *Exp. 1.8*, *Exp. 1.9*, *Exp. 1.10*) ran on the PW SP's shown on a scale of [0, 1]. Results from Table 36 show that the component weightings (Sem 0.7, Syn 0.05, FI 0.25) from *Exp. 1.9* produced the best overall correlation and the highest correlation, giving better results than the other algorithm measures for PW SP's as shown in Table 37. The higher the correlation value, the closer the similarity ratings are to those of the human ratings (HR). Figure 24 shows the scatter plot for the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for *Exp. 1.9* for the PW SP's. The trendline shows a positive linear relationship between the variables. The results from the ten experiments (*Exp. 1.1*, *Exp. 1.2*, *Exp. 1.3*, *Exp. 1.4*, *Exp. 1.5*, *Exp. 1.6*, *Exp. 1.7*, *Exp. 1.8*, *Exp. 1.9*, *Exp. 1.10*) on the FI-25 dataset gave positive indicators that H_0 would be accepted.

Pearson's Correlation	***r-value***	***r-value***	***r-value***	***r-value***	***r-value***
	Exp. 1.6 **Sem 0.8** **Syn 0.1** **FI 0.1**	**Exp. 1.7** **Sem 0.7** **Syn 0.1** **FI 0.2**	**Exp. 1.8** **Sem 0.75** **Syn 0.15** **FI 0.1**	**Exp. 1.9** **Sem 0.7** **Syn 0.05** **FI 0.25**	**Exp. 1.10** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.2497	0.2334	0.1878	0.2997	0.0826

Table 36 - Hyper-Parameter Optimisation for PW SP's

SSM	**r-value**
HR vs FUSE_4.0	0.2997
HR vs FUSE_2.0	0.1914
HR vs FUSE_3.0	0.2052
HR vs STASIS	0.1677
HR vs API Semantic	0.0519
HR vs API Syntactic	0.0731
HR vs SEMILAR	0.1286

Table 37 - Comparison of SSM Best Results from Table 36

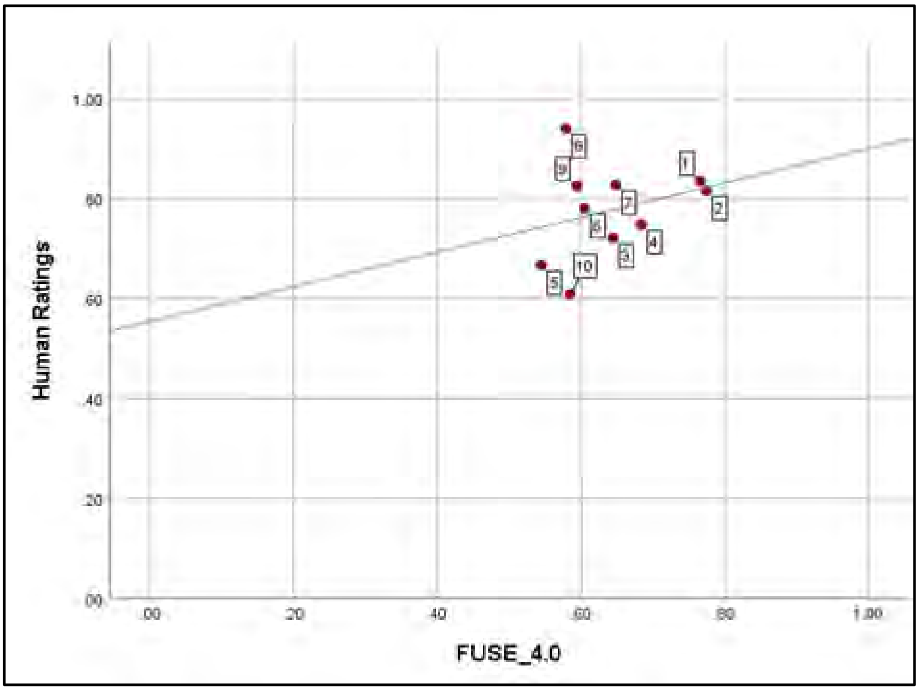


Figure 24 - PW Dataset Scatter Plot (Sem 0.7, Syn 0.05, FI 0.25)

6.5.4.2 - Experiment 2 - FI on FUSE-62 Dataset

FI-25 was a limited dataset, so a series of further empirical experiments were undertaken using a similar range of semantic, syntactic and FI factor weights using the FUSE-62 dataset. FUSE-62 consisted of 62 sentence pairs specifically designed with the help of English language experts to contain fuzzy words per sentence from all nine categories of the FUSE algorithm (Adel et al., 2019). Table 38 shows the correlation findings for the five experiments *(Exp. 2.1, Exp. 2.2, Exp. 2.3, Exp. 2.4, Exp. 2.5)* ran on the FUSE-62 dataset shown on a scale of [0, 1]. Results from Table 38 show that the measures (Sem 0.5, Syn 0.2, FI 0.3) from *Exp. 2.5* gave the best overall correlation with human ratings and also higher than competing measures as shown in Table 39. The scatter plot in Figure 25 shows the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for *Exp. 2.5* for the 62-SP dataset.

Pearson's Correlation	***r-value***	***r-value***	***r-value***	***r-value***	***r-value***
	Exp. 2.1 Sem 0.8 Syn 0.1 FI 0.1	**Exp. 2.2 Sem 0.7 Syn 0.1 FI 0.2**	**Exp. 2.3 Sem 0.75 Syn 0.15 FI 0.1**	**Exp. 2.4 Sem 0.7 Syn 0.05 FI 0.25**	**Exp. 2.5 Sem 0.5 Syn 0.2 FI 0.3**
HR vs FUSE_4.0	0.6221	0.6422	0.6462	0.6255	0.7027

Table 38 - Hyper-Parameter Optimisation for FUSE-62 Dataset

SSM	**r-value**
HR vs FUSE_4.0	0.7027
HR vs FUSE_2.0	0.5553
HR vs FUSE_3.0	0.6260
HR vs STASIS	0.5930
HR vs API Semantic	0.5263
HR vs API Syntactic	0.6712
HR vs SEMILAR	0.6646

Table 39 - Comparison of SSM Best Results from Table 38

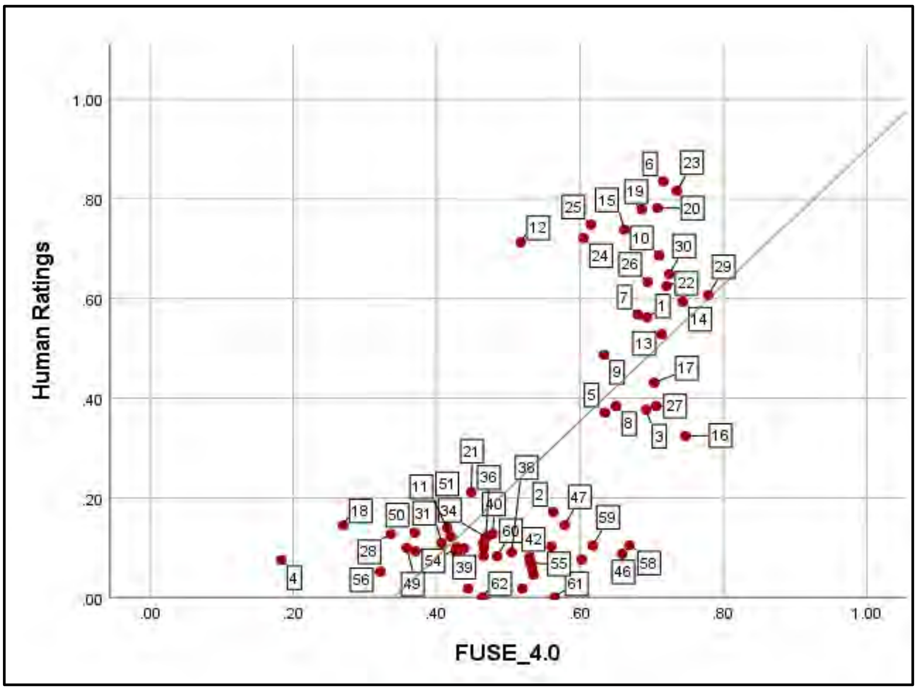


Figure 25 - FUSE-62 Dataset Scatter Plot (Sem 0.5, Syn 0.2, FI 0.3)

6.5.4.3 - Experiment 3 - FI on MWFD Dataset

The same five experiments were conducted on the published MWFD dataset (Chandran, 2013). This dataset consisted of 30 sentence pairs specifically designed by English language experts to contain at least two fuzzy words per sentence. Table 40 shows the correlation findings for the five experiments (*Exp. 3.1*, *Exp. 3.2*, *Exp. 3.3*, *Exp. 3.4*, *Exp. 3.5*) for the MWFD dataset shown on a scale of [0, 1]. Table 40 shows that the measures (Sem 0.8, Syn 0.1, FI 0.1) from *Exp. 3.1* produced the best overall correlation and the highest correlation against the other algorithm measures, with the exception of FUSE_3.0 which was slightly higher as shown in Table 41. The scatter plot in Figure 26 shows the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for *Exp. 3.1* for the MWFD dataset.

Pearson's Correlation	***r-value***	***r-value***	***r-value***	***r-value***	***r-value***
	Exp. 3.1 **Sem 0.8** **Syn 0.1** **FI 0.1**	**Exp. 3.2** **Sem 0.7** **Syn 0.1** **FI 0.2**	**Exp. 3.3** **Sem 0.75** **Syn 0.15** **FI 0.1**	**Exp. 3.4** **Sem 0.7** **Syn 0.05** **FI 0.25**	**Exp. 3.5** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.7589	0.7410	0.7559	0.7340	0.6933

Table 40 - Hyper-Parameter Optimisation for MWFD Dataset

SSM	r-value
HR vs FUSE_4.0	0.7589
HR vs FUSE_2.0	0.7538
HR vs FUSE_3.0	0.7683
HR vs STASIS	0.7452
HR vs API Semantic	0.7009
HR vs API Syntactic	0.3930
HR vs SEMILAR	0.7303

Table 41 - Comparison of SSM Best Results from Table 40

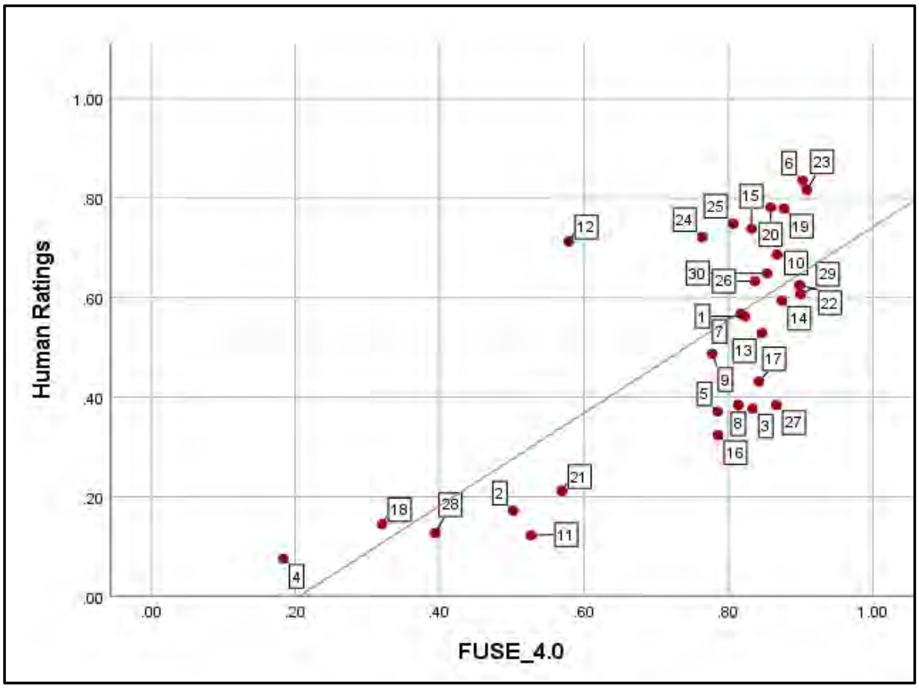


Figure 26 - MWFD Dataset Scatter Plot (Sem 0.8, Syn 0.1, FI 0.1)

6.5.5 - Discussion

Table 42 shows information with regards to the datasets that were used in the FI experiment. The *a-value* shows the Intra-Class Correlation Coefficient (ICC) for each of the datasets that was experimented on using the different algorithms. Since the *a-value* results are between 0.75 and 1.00 for each dataset, it is deemed that the Inter-Rater Agreement measure of human ratings are deemed Ex*cellent* according to Cicchetti (Cicchetti, 1994). Table 42 also shows that the *p-value* for each dataset is less than 0.05 for a confidence level of 95% and thus provides support for our research hypothesis H_0.

Datasets	FI25_NPW	FI25_PW	FI25	FUSE-62	MWFD
a-value	0.998	0.953	0.997	0.987	0.999
p-value	*p* < .001	*p* < .001	*p* < .001	*p* < .001	*p* < .001

Table 42 - (a-value) & (p-value) For Each Dataset

The snapshot of empirical experiments conducted on the four datasets indicated that the inclusion of a FI factor in a FSSM can improve the performance of the algorithm in terms of its correlation with human ratings. Although this FI is relatively simple, it has to a degree been able to model the uncertainty of human perception-based words which have already been modelled using Interval Type-2 fuzzy sets. The interaction of the FI factor with both the semantic and syntactic components of FUSE_4.0 must be kept to a minimum, to preserve the importance of the word order and ontological path length in calculating the overall similarity. FUSE_4.0 performs best when there is at least one fuzzy word present in the sentence pair being evaluated.

6.6 - Conclusion

This chapter has described the evolution of the FUSE algorithm. The first version of the algorithm, FUSE_1.0, was modified to investigate the presence of linguistic hedges in a fuzzy sentence pair and how this can be evaluated to improve sentence similarity. Results showed FUSE_1.0 achieved a higher correlation with human ratings (r = 0.8028) when compared to a traditional SSM STASIS (r = 0.7959).

The second version of the algorithm, FUSE_2.0, expanded the six existing fuzzy categories with the introduction of three new fuzzy categories, bringing the total number of fuzzy categories to nine. The FUSE_2.0 algorithm was tested on five datasets and results of correlation with AHR was compared to other SSM's. Results showed that FUSE_2.0 achieved a higher correlation with AHR for all five datasets tested compared to the SSM's it was compared with.

Version three of the algorithm, FUSE_3.0, covered the presence of negation operators in fuzzy utterances and discussed published approaches of how negation can be dealt with in fuzzy sets. A full evaluation of the effects of negation operators will be explored in Chapter 7 when the FUSE algorithm is incorporated into a dialogue system.

The final version of the algorithm, FUSE_4.0, introduced the concept of a fuzzy influence factor. Up to this point only fuzzy words in the same fuzzy category were diverted to the fuzzy dictionary when present in a fuzzy sentence pair. This meant any fuzzy word in a sentence pair that did not belong to the same fuzzy category was dismissed and diverted to WordNet (Miller, 1995) to obtain a value.

The fuzzy influence factor allows the presence of fuzzy words in a sentence pair to use the fuzzy dictionary values, regardless of the fuzzy words being in the same fuzzy category or not, allowing for a truer representation of perception in sentence similarity to be achieved. FUSE_4.0 was tested on four datasets and results of correlation with AHR compared to other SSM's as well as earlier versions of FUSE (FUSE_2.0 and FUSE_3.0). Results showed an improvement of results with correlation to, AHR compared with other SSM's and the earlier versions of the FUSE algorithm. It is important to note that the interaction of the FI factor with both the semantic and syntactic components of FUSE_4.0, must be kept to a minimum level. This is to preserve the importance of the word order and ontological path length in calculating the overall similarity.

Chapter 7 will focus on incorporating the FUSE algorithm, namely FUSE_2.0 and FUSE_4.0 into a simple dialogue system known as FUSION. The objective is to investigate if incorporating a FSSM into a dialogue system can improve language understanding using real life scenarios.

CHAPTER 7

CHAPTER 7: INTEGRATION OF FUSE INTO A DIALOGUE SYSTEM

7.1 - Introduction

Up to this point in the thesis, the emphasis has been on designing a suitable FSSM which takes into consideration many aspects of sentence similarity. The examples include capturing the presence of hedge words in a fuzzy utterance, modelling a wide coverage of fuzzy words via nine fuzzy categories, dealing with the presence of negation words such as '*not*' in sentences, as well as introducing a fuzzy influence factor which further covers the perception of fuzzy words in sentence pairs by not being limited to fuzzy words necessarily belonging to the same category in a sentence pair.

This chapter will focus on incorporating the FUSE algorithm into a simple dialogue system to help answer the second research question proposed in this thesis:

RQ2. Can a Type-2 FSSM be embedded into a Q&A dialogue system with an improved success rate of utterance - response matches compared to traditional Semantic Similarity Measures (SSM)?

The term dialogue systems mainly refer to a conversational computer system. "*It is the ability to converse with a computer using natural language*" (Deriu et al., 2021).

A question-and-answer dialogue system will be designed, where the matching of user responses to patterns in the system can be achieved through measuring the fuzzy semantic similarity between a number of prototypical sentences using the FUSE algorithm. The aim of incorporating the FUSE algorithm into a dialogue system is to evaluate the performance of a FSSM in capturing natural language dialogue and matching the human responses correctly to a set of pre-defined prototypical sentences and compare the performance to a traditional SSM. This chapter is broken up into several sections:

Section 7.2 will examine the history of dialogue systems, the different types of dialogue systems and the challenges associated with evaluating dialogue systems before assessing a suitable approach for the design of the proposed dialogue system, referred to as FUSION in this research.

Section 7.3 will discuss the design and scripting involved with the creation of the first version of the dialogue system referred to as FUSION_V1. FUSE_2.0 was integrated and tested using

FUSION_V1 with a set of questions designed around a café-based scenario asking human participants to describe their experience of the visit to the café. Subsequently moving onto the methodology and evaluation of the experimental results. Results will be evaluated against a set of prototypical sentences designed to match the similarity of the user responses. The performance of FUSE_2.0 will be compared to STASIS, a traditional SSM.

Section 7.4 will address the design and implementation of the second version of the dialogue system referred to as FUSION_V2 using the FUSE_4.0 algorithm. This second version will also address some limitations that were highlighted as a result of the FUSION_V1 experiment. FUSION_V2 was designed using two sets of questions and used a working from home scenario which was implemented online due to the Covid-19 pandemic and restrictions of social distancing. Results will be evaluated against two sets of prototypical sentences, (one for each set of questions), designed to compare the similarity of the user responses. The performance of FUSE_4.0 will be compared to STASIS (Li et al., 2006), a traditional SSM.

This phase of the project received an approval Following Manchester Metropolitan Universities ethical approval process (Ethos number: 11759).

The results of the FUSION_V1 experiment were published in **Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures**. N Adel, K Crockett, D Chandran, JP Carvalho. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), UK (virtual congress), 2020. DOI: 10.1109/FUZZ48607.2020.9177605.

7.2 - Dialogue Systems

7.2.1 - Review of Dialogue Systems

A Dialogue Systems (DS) is a computer program which interacts with a user through natural language dialogue and provides some form of service (J. O'Shea et al., 2011; Ozaeta and Graña, 2018; Harms et al., 2018; Aujogue and Aussem, 2019), Dialogue Systems (DS) are applications, which effectively replace human experts by interacting with users through natural language dialogue to provide a type of service or advice (J. O'Shea et al., 2013). For a

DS to engage with humans, they must be able to handle extended natural language dialogue relating to complex tasks and potentially engage in decision-making.

In this sense, conversational agents are helpful tools for human-machine interaction, allowing the input of data via natural language, processing sentences, and returning answers appropriately through text. DS, sometimes known as conversational agents, have been used in a wide range of applications such as customer service (J. O'Shea et al., 2013), help desk support (Ozaeta and Graña, 2018), Educational reasons (Latham et al., 2014; Aljameel et al., 2017; Aljameel et al., 2019; L. Lin et al., 2020), Cognitive Behavioural Therapy for young adults (Fitzpatrick et al., 2017), insurance (Koetter et al., 2019) and healthcare (Montenegro et al., 2019).

Dialogue understanding has become more valuable to companies, with the easier ability to gain insights from unstructured text through Google's AutoML and natural language API (Natural Language AI, 2020), to Amazon's use of supervised machine learning to allow correct interpretation of natural language vocabulary reducing, for example, the detection of false positive responses (Alexa and Alexa Device FAQs - Amazon Customer Service, 2016). For spoken DS, task-based systems which utilise deep reinforcement learning techniques in their dialogue management systems are also becoming more available to industry (Wang et al., 2020). What makes a successful DS is the ability for the machine to understand and interpret the human's natural language response in the context of the conversation.

Dialogue systems first appeared in the 1960's and 1970's and were mainly text-based, some examples of them are BASEBALL (McTear, 2020), a question and answering dialogue system that answered questions about baseball, SHRDLU (McTear, 2020) was a pragmatic component that processed non-linguistic information about the domain and GUS (McTear, 2020) a flight booking system.

Dialogue systems usually take turns with the user and based on the user's response or utterance the next dialogue is activated. The conversation can vary in method such as text based, image based or voice based (Car et al., 2020). In the work presented in this thesis, the focus is on text-based DS only.

Dialogue systems can be broken down into three main systems (Deriu et al., 2021) each one is explained further below:

1. **Task-Oriented Systems** are designed to help a user solve a specific task i.e., flight booking, virtual assistant etc. (Deriu et al., 2021).
2. **Conversational Agents** tend to have more unstructured conversation with no specific task to solve, such as chitchat systems or chatbots (Ma et al., 2020).
3. **Question & Answering (Q&A) Systems** are designed to answer questions and tend to follow a question and answer style pattern (Deriu et al., 2021).

The research in this thesis will only be concerned with the application of the FUSE algorithm within a Q&A system. Q&A systems aim to satisfy users who are looking to answer a specific question in natural language (Bouziane et al., 2015). A Q&A system typically asks the user a question in natural language and returns an answer to this question as opposed to returning a set of documents or links deemed relevant, similar to how search engines just as Google or Ask Jeeves return links.

Some applications of Q&A systems include AquaLog (Lopez et al., 2007) which takes queries expressed in natural language and an ontology as input, and returns answers drawn from one or more knowledge-bases. AquaLog uses WordNet (Miller, 1995) and a novel ontology-based relation similarity service to make sense of user queries with respect to the target knowledge-base. It also has a learning component, which ensures that the performance of the system improves over time in response to jargon used by end users (Lopez et al., 2007).

QACID (Ferrández et al., 2009) is another example of an ontology-based Q&A system that allows users to retrieve information from formal ontologies using queries in natural language in Spanish. It can offer simple adaptability to deal with inter-domain portability and changes in user information requirements (Ferrández et al., 2009).

Finally, WabiQA (Noraset et al., 2021), is a Q&A system in the Thai language that uses the Thai Wikipedia articles as the knowledge source. It firstly retrieves the Wikipedia article that is most likely to contain the answer and will then read the article and locate candid answers and rank them by confidence levels and return to the user (Noraset et al., 2021).

7.2.2 - Challenges in Interpreting Dialogue Systems

The main challenge of any dialogue system is that they are still fragile and may crash easily if deviated from the expected input. For example, considering the Amazon Customer Service DS (Alexa and Alexa Device FAQs - Amazon Customer Service, 2016); If you wish to contact a human representative from the Amazon Customer Service Team, you will first be connected to a dialogue system, which will ask you a series of questions to establish the issue you are facing. If you answer according to how the DS has been programmed to respond, it will direct you through a set of useful answers before connecting you to a human representative, if needed, to chat to. However, if you respond with sentences that do not match the pre-designed ones in its library, it will not understand you and will simply divert you to a human representative.

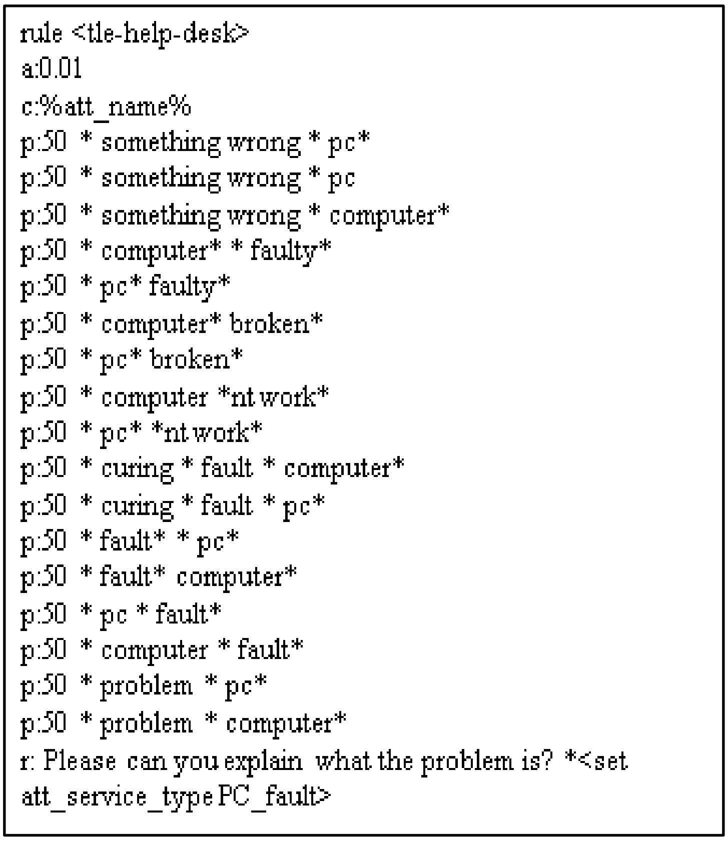
A limitation of most DS is that they only work well for the purpose they are built for, but are difficult to transfer across domains (McTear, 2020). For example, the Amazon Customer Service DS may work well as a customer service DS but transferring that directly to a different domain such as a tutoring DS for a maths subject will not work correctly without prior amendments to the system being made.

It can be difficult to evaluate just how well a DS is performing, since high-quality human dialogue (for input) may not always be available to adequately test all possible scenarios. Consequently, evaluation of a DS may also be measured in terms of task performance success. Furthermore, evaluation can also consider if the users response triggers the next correct dialogue turn response. User feedback may also not always be available or reliable and can be difficult to measure the appropriateness of the human dialogue as it can be relative to context and subjective human evaluation.

Most DS rely on grammatically correct sentences, yet most user responses are not like this (J. O'Shea et al., 2011). For example*, I don't want no refund as* opposed to *I don't want any refund*. The first sentence is grammatically incorrect as it uses a double negative (*don't* and *no*) which essentially cancel each other out, ultimately meaning you DO want a refund, when in essence the second sentence portrays the correct meaning of not wanting a refund.

Furthermore, evaluation of a DS is often not cost-effective and time consuming due to the need for human participants. Dialogue systems typically suffer from high maintenance in updating dialogue patterns for new scenarios due to the huge number of language patterns available within the scripts. Typically, DS work off scripts, which are organized into contexts, consisting of hierarchically organized rules with combining patterns and associated responses. Figure 27 is one such example of a pattern matching rule. Scripts are needed to capture a wide variety of inputs and hence many rules are required, each of which deals with an input pattern and the possible variations and an associated responses (J. O'Shea, 2010; Curry, 2018; Aljameel et al., 2019).

InfoChat is one such pattern matching system which utilises the sophisticated *PatternScript* scripting language (Michie and Sammut, 2001) and has been adapted over the years for use in intelligent conversational tutorial systems (Latham et al., 2014). Figure 27 (Source: [adapted from] Latham et al., 2014) shows an example of a pattern matching rule, <*tle-help-desk*> which has been encoded using the scripting language provided with the agent InfoChat. The *(r)ule* uses default values for *(a)ctivation* and *(p)attern* matching strength, has a *(c)ondition* (that the variable *att_name* has a value) and a response consisting both of a text and the setting of a *variable <set att_service_type PC_fault>*. Whilst pattern matching scripting engines are a mature technology and robust, to some degree to expected user input, scripting is an art form and requires good knowledge of the language and the ability to perform in-depth knowledge engineering of the domain (J. O'Shea et al., 2011; Aljameel et al., 2017; Curry, 2018).



```
rule <tle-help-desk>
a:0.01
c:%att_name%
p:50  * something wrong * pc*
p:50  * something wrong * pc
p:50  * something wrong * computer*
p:50  * computer* * faulty*
p:50  * pc* faulty*
p:50  * computer* broken*
p:50  * pc* broken*
p:50  * computer *nt work*
p:50  * pc* *nt work*
p:50  * curing * fault * computer*
p:50  * curing * fault * pc*
p:50  * fault* * pc*
p:50  * fault* computer*
p:50  * pc * fault*
p:50  * computer * fault*
p:50  * problem  * pc*
p:50  * problem  * computer*
r: Please  can you explain  what the problem is? *<set
att_service_type PC_fault>
```

Figure 27 - Pattern Matching Rule (Source: [adapted from] Latham et al., 2014)

As can be seen by the *InfoChat* example in Figure 27 , the use of traditional scripting methodologies within dialogue systems involves interpreting structural patterns of sentences by using contextualised rule based scripts relating to a particular topic (Michie and Sammut, 2001), thus illustrating that scripting patterns is inefficient, results in domain instability and high maintenance costs. Each context consists of several hierarchically organized rules possessing a list of structural sentence patterns and associated response. A user's utterance is matched against the patterns and the associated response is chosen based on a pattern scoring algorithm and retuned as output.

Scripts are usually constructed by firstly assigning each rule a base activation level, typically a number between [0, 1]. This is to ensure that if any conflict occurs between two or more rules, there are rules in place which will match the users input (Sammut, 2001). It is down to the researcher designing the scripts to choose which patterns respond to user inputs with each pattern typically being assigned a strength ranging between [10, 50] (Sammut, 2001).

As mentioned, scripting is an art form and requires good knowledge of the language and the ability to perform in-depth knowledge engineering of the domain (J. O’Shea et al., 2011; Aljameel et al., 2017; Curry, 2018), which in turn makes the scripting process time consuming. This introduction of a new rule or modifying an existing rule has a knock-on effect on the other rules, therefore a reassessment of the entire script is needed (Michie, 2001).

O’Shea (K. O'Shea et al., 2009) introduced an approach to overcome this weakness of traditional scripting by replacing the pattern matching rules with short text semantic similarity measures (SSM’s). The aim of this work was to reduce the complexity of producing scripts for use within dialogue systems and reduce the maintenance time as new topics were added. By using a sentence similarity measure, a match is determined between the user’s utterance and a set of prototypical natural language sentences. The highest ranked sentence is fired and sent as output (K. O'Shea et al., 2009). O’Shea (K. O'Shea et al., 2009) uses the three steps below to illustrate this procedure:

1. Natural language dialogue is received as input, which forms a joint word set with each rule from the script using only distinct words in the pair of sentences. The script is comprised of rules consisting of natural language sentences.
2. The joint word set forms a semantic vector using a hierarchical semantic/lexical knowledgebase. The weight of each word is based on its significance by using information content derived from a corpus.
3. Combining word order similarity with semantic similarity the overall sentence similarity is determined. The highest ranked sentence is chosen and sent as output.

The proposed method by O’Shea (K. O'Shea et al., 2009) showed effectiveness and flexibility to develop extended dialogue applications (J. O'Shea, 2010, J. O’Shea et al., 2011; Pazos et al., 2013;), especially when coupled with ruled based matching algorithms to produce controlled responses and offer flexibility to sustain dialogues with users. Utilising this new approach within a DS was more effective than traditional techniques because it replaced the scripted patterns by a few natural language sentences in each rule.

Evaluation of short text SSM based dialogue systems has been shown to improve the robustness of the system in terms of increasing the number of correctly fired rules, thus maintaining the conversational flow and increasing usability (K. O’Shea, 2012; Curry, 2018).

However, when traditional short text SSM’s are used, they do not sufficiently match the fuzziness of natural language. This is due to the presence of human perception-based words in sentences, and by not addressing the presence of fuzzy words can often lead to a fundamental meaning of the human utterance in the dialogue context being misunderstood. This in turn, can cause incorrect firing of a rule, which will lead to an incorrect flow of conversation and even wrong tasks being suggested. This suggests that a fuzzy short text SSM is a viable alternative.

7.2.3 - Evaluation of Dialogue Systems

There are many definitions around what constitutes a good dialogue system and there is no ‘one size fits all’ metric (Deriu et al., 2021). Depending on the task at hand and the data required, the desired qualities that make up a DS may differ from one system to another. Therefore, there are many approaches used to measuring the performance of a DS, such as measuring the conversation length (i.e., how long the DS can hold a conversation with a human) or asking participants to rate the system (i.e., how well did they feel the dialogue system performed or how easy it was to use) via conducting a usability questionnaire (Deriu et al., 2021).

Other relevant factors may also impact the evaluation of a DS such as the effectiveness, efficiency and user satisfaction for the task-oriented systems (i.e., was the dialogue system successful in completing the assigned task at hand?), and appropriateness and human likeness for systems oriented towards small talk (i.e., was the dialogue system able to hold meaningful conversation with the human user?) (Deriu et al., 2021). To evaluate a Q&A system some aspects to consider are the correctness of the answer provided by the DS (to the question asked by the human and/or system) and the flow of the conversation (Deriu et al., 2021).

In a semantic dialogue system, each rule is matched in accordance with a pre-determined semantic similarity threshold, which is set initially through empirical evaluation and depends upon the sensitivity of rules within a context. Figure 28 shows a simple rule comprising of a set of prototypical *(s)entences*, where the similarity with the user utterance is calculated using a short text SSM. Each rule has a series of *(r)esponses*, which are provided to the user and can be randomly selected. Each rule also has an associated default rule, which would fire if the user utterance failed to match any prototypical sentences within the rule. O'Shea (K. O'Shea, 2012) devised a semantic scripting language which incorporated a Short Text Semantic Similarity (STSS), through adapting the pattern matching language of InfoChat (Curry, 2018). This devised language encompasses the ability to extract patterns to set variables, set rule conditions and freeze, promote and demote rules.

```
rule <tle-help-desk> c:%att_name%
s: There is a problem with my computer
r: Please can you explain what the problem is?
*<set att_service_type PC_fault>
```

Figure 28 - Semantic Rule Example

In a semantic dialogue system, prototypical sentence rules are compared with user utterances using a pre-selected STSS algorithm and the rule with the highest similarity match would fire. The most obvious benefit of using semantic rules is no patterns are required and more importantly the semantic meaning of the utterance can be captured and acted upon within the dialogue context. Aljameel (Aljameel et al., 2017) used a hybrid similarity approach, combining a short text SSM with limited patterns, to construct an Arabic conversational intelligent tutoring system for the education of autistic children. The conversational agent processed Arabic utterances using a novel crisp short text SSM which utilised the cosine similarity measure to solve the word order issue associated with the Arabic language. Consequently, this reduced the number of scripts and rules required. Through empirical evaluation of two versions of the system, the use of a short text SSM reduced the number of unrecognised human utterances to 5.4% compared to 38% in the pattern scripted version, hence, the systems incorrect responses were reduced to 3.6% compared to 10.2% in the pattern scripted version (Aljameel et al., 2017). Similar improvements on the benefits of utilising a short text SSM within DS are also reported in (Kaleem et al., 2014). In this research, the traditional semantic similarity measure is replaced with a fuzzy semantic similarity measure to evaluate the effectiveness of a DS through a reduction in the incorrect responses and unrecognised human utterances compared with using a short text SSM.

7.3 - FUSION_V1 Dialogue System Using FUSE_2.0

7.3.1 - Aim and Purpose

This section describes the creation and evaluation of a simple DS referred to as FUSION_V1 which utilises the FUSE_2.0 measure to match human utterances to a set of fuzzy phrases within a rule-based system. The aim is to improve the robustness of rule matching within the DS, compared to the use of a crisp similarity measure in a market research scenario, where the capture of rich descriptive dialogue is important in gaining customer insight. A fuzzy DS can be used to automate the analysis of unstructured answers given to open ended questions, allowing for richer insight when collecting survey data. For example, an understanding of the dialogue, can lead to further probing to obtain more descriptive answers that provide greater insight into why a particular answer was given.

A simple linear Q&A semantic dialogue system will be developed, where the user will be asked a series of questions based on a given scenario. Each question will record their given answer and move onto the next question. This section will also describe an experiment that will validate the use of a fuzzy short text SSM in a DS. This experiment aims to address the following research question:

RQ2. Can a Type-2 FSSM be embedded into a Q&A dialogue system with an improved success rate of utterance - response matches compared to traditional Semantic Similarity Measures (SSM)?

The hypothesis for this experiment is:

H_0: A FSSM used in a DS will improve success rate of utterance - response matches compared with a traditional short text SSM?

7.3.2 - Design

FUSION_V1 is a simple question and answer dialogue system that utilises the FUSE_2.0 semantic similarity measure (Adel et al., 2018), to match user utterances to different categories of responses to each question. The dialogue structure is therefore a linear sequence of questions, where each questions response has three possible branches. The aim is to distinguish between human perceptions of fuzzy words in the nine fuzzy categories of FUSE_2.0 to assess if the correct rule fires in response to natural language used within the human utterance.

To establish if a FSSM could be used in a dialogue system, a simple question and answer semantic dialogue system was designed to obtain feedback from participants who visited a local café. This was done using a knowledge engineering approach (J. O'Shea et al., 2011), which involved information gathering about typical questions asked in a customer satisfaction online questionnaire, concerning customer satisfaction levels in high street cafes. Existing survey questions were a mixture of open-ended questions, dichotomous questions, multiple choice or Likert scale questions (Nemoto and Beglar, 2014).

Within the proposed café feedback DS, each question selected had to be transformed into one that would allow the user to provide descriptive textual answers to gather as much data as possible, to evaluate the impact of the fuzzy semantic measure. Therefore, open-ended questions seemed like a viable option as it would allow the participants to express their opinion without being influenced by the researcher (Reja et al., 2003). An example of one such open-ended question is taken from Reja et al., *In your opinion, what is the most critical problem the internet is facing today?* (Reja et al., 2003).

To ensure all nine fuzzy categories in FUSE_2.0 were covered, nine open-ended questions where created, each one covering responses that would contain words or synonyms of words from each fuzzy category. Table 43 shows the nine questions created and the fuzzy category each question maps to. Each question formulates a *question-rule* within the DS, where each rule can have three possible responses which represent full coverage of the categories as defuzzified word values obtained through human experts and Interval Type-2 modelling using the HMA approach (Adel et al., 2018).

Question	Category	Question Asked
Q1	**Size/Distance**	Using descriptive words, how would you describe the size of the queue?
Q2	**Temperature**	How would you describe the temperature of the cafe?
Q3	**Brightness**	How would you describe the brightness of the cafe?
Q4	**Age**	Using descriptive words, how would you describe the age of the barista that served you?
Q5	**Speed**	Once you placed your order, how quickly was your drink made and served to you?
Q6	**Strength**	Looking up from your screen to the first person you see, how would you describe their physical strength?
Q7	**Frequency**	How frequently do you visit this cafe?
Q8	**Level of Membership**	How did todays visit meet your expectation?
Q9	**Worth**	How would you describe your experience overall today?

Table 43 - FUSION_V1 - Café Scenario Questions

The rule responses were divided into three thresholds of *high*, *medium* and *low*, and words (and word synonyms) within each category fell under each threshold. The threshold for each category varies as the number of words and measurements in each category varies due to the dependency on human perceptions (Adel et al., 2018). The thresholds in each of the nine fuzzy categories were selected based on the words in that specific category. An example is shown in Figure 29 and Figure 30 for the two categories of *Frequency* and *Worth*. The thresholds for all nine fuzzy categories can be found in Appendix E.

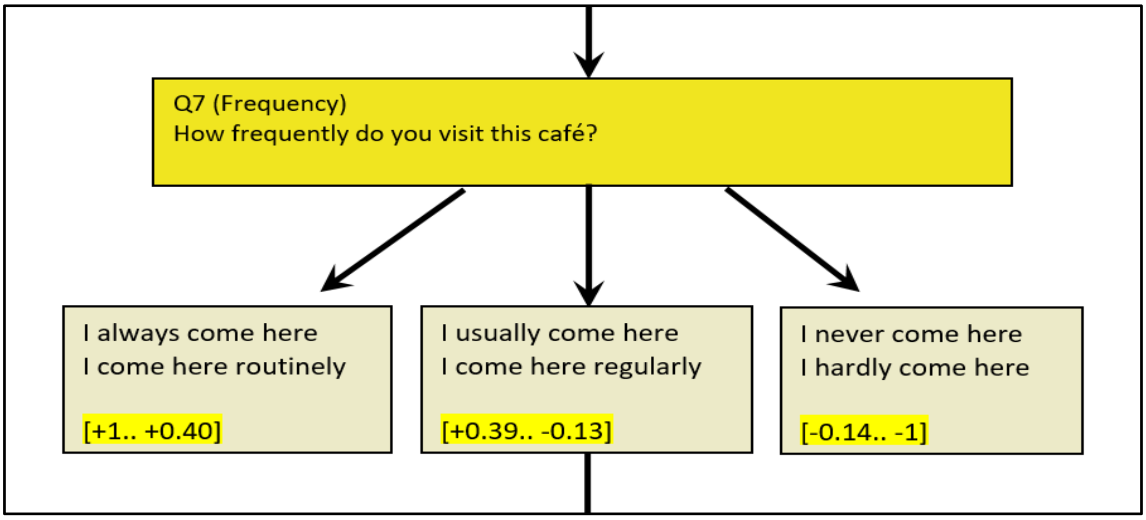


Figure 29 - Frequency Threshold - FUSION_V1

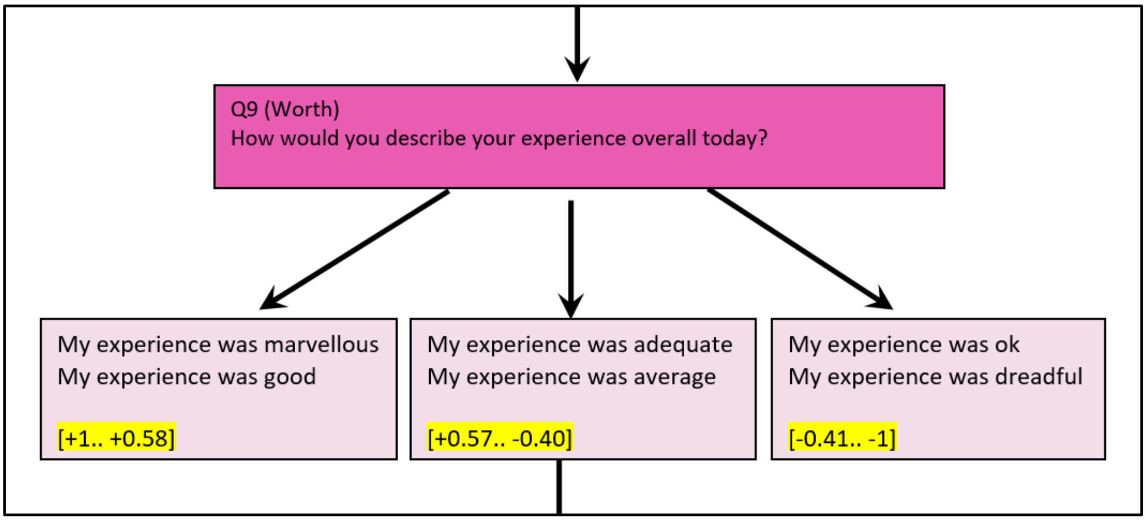

Q9 (Worth)
How would you describe your experience overall today?
My experience was marvellous
My experience was good
[+1.. +0.58]
My experience was adequate
My experience was average
[+0.57.. -0.40]
My experience was ok
My experience was dreadful
[-0.41.. -1]

Figure 30 - Worth Threshold - FUSION_V1

Referring to Figure 29, for the category *Frequency*, the *high* threshold begins at [-1] and ends at [+0.40], with the last word to fall in this threshold being *Everytime*, and the next word after this which begins the *medium* threshold is *Occasionally* at [+0.39], and this threshold continues up to [-0.13], and even though this is now a negative value, it still falls in the *medium* threshold for this category, and the *low* threshold starts at [-0.14] and ends at [-1]. Examining Figure 30 for category *Worth*, the *high* threshold starts at [+1] and ends at [+0.58], the *medium* threshold begins at [+0.57] and ends at [-0.40], and the *low* threshold begins at [- 0.41] and ends at [-1]. It is important to note that there is not a single fixed threshold for all nine fuzzy categories, as the words and their values varied in each category. To determine the specific *high*, *medium* and *low* thresholds for each fuzzy category, two English language experts independently grouped the words for each category. In the case of disagreement, a third expert was asked to cast the deciding vote.

7.3.3 - Scripting

Each question was scripted into a context which represented a category as shown in Table 43. Three pairs of English prototypical sentences were used in each rule to enable coverage of either the *high*, *medium* or the *low* thresholds. In addition, there were initialisation and conclusion contexts. Figure 31 shows three *(r)*ules from the *Size/Distance* category. Each dialogue exchange between human and machine generated a human utterance that was compared to the six prototypical sentences in each rule. In each context, the rule where the *(s)entence* gave the highest measure of similarity determined by FUSE_2.0 on a scale of [0, 1] was analysed. An attribute is set i.e., *att_size-distance-high* becomes true if *default-rule1* fires and a change in context occurs, denoted by the '*c:*' identifier. As this is a simple linear DS, the change in context is always set to the context of the next question until all questions have been asked.

```
<default-rule1><size/distance>
s: It was long
s: It was huge
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-high>
c: temperature_context

<Default-rule2><size/distance>
s: It was average
s: It was regular
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-medium>
c: temperature_context

<Default-rule3><size/distance>
s: It was tiny
s: It was small
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-low>
c: temperature_context
```

Figure 31 - Sample Rules for Size/Distance Category

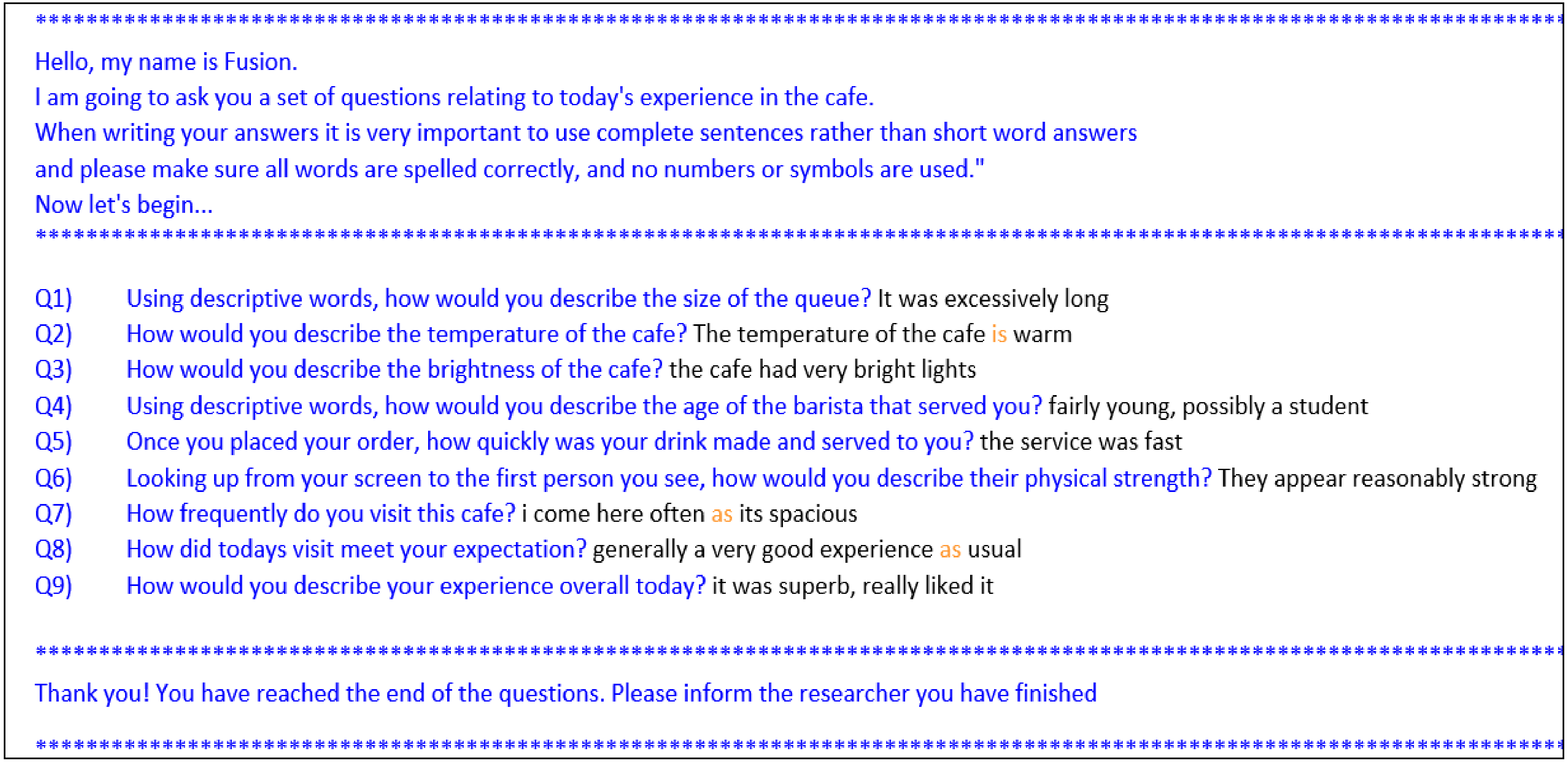
On initiation of the system, the DS begins with the simple message:

> “*Hello, My name is Fusion.*
>
> *I am going to ask you a set of questions relating to today’s experience in the cafe.*
>
> *When writing your answers it is very important to use complete sentences rather than short word answers and please make sure all words are spelled correctly, and no numbers or symbols are used.*
>
> *Now let’s begin...*”.

After all questions were asked the final message was:

> “*Thank you! You have reached the end of the questions. Please inform the researcher you have finished*.”

Figure 32 shows an example of a participant’s answers. A log file recorded all dialogue, including the semantic similarity score for each rule during the completion of the survey. In this version of the system, all human utterances were recorded. This included incorrect utterances which failed to match any rules in a given context.



```
************************************************************************************************************************************************
Hello, my name is Fusion.
I am going to ask you a set of questions relating to today's experience in the cafe.
When writing your answers it is very important to use complete sentences rather than short word answers
and please make sure all words are spelled correctly, and no numbers or symbols are used."
Now let's begin...
************************************************************************************************************************************************

Q1)     Using descriptive words, how would you describe the size of the queue? It was excessively long
Q2)     How would you describe the temperature of the cafe? The temperature of the cafe is warm
Q3)     How would you describe the brightness of the cafe? the cafe had very bright lights
Q4)     Using descriptive words, how would you describe the age of the barista that served you? fairly young, possibly a student
Q5)     Once you placed your order, how quickly was your drink made and served to you? the service was fast
Q6)     Looking up from your screen to the first person you see, how would you describe their physical strength? They appear reasonably strong
Q7)     How frequently do you visit this cafe? i come here often as its spacious
Q8)     How did todays visit meet your expectation? generally a very good experience as usual
Q9)     How would you describe your experience overall today? it was superb, really liked it

************************************************************************************************************************************************
Thank you! You have reached the end of the questions. Please inform the researcher you have finished
************************************************************************************************************************************************
```

Figure 32 - Sample Participant Answer

7.3.4 - Experimental Evaluation Methodology

Following Manchester Metropolitan Universities ethical approval process (Ethos number: 11759), 32 participants were recruited through an advertising campaign through the University. 32 participants were chosen to ensure the sample size was sufficient and allow the results to be statistically significant (J. O'Shea et al., 2013). Prior to commencing the experiment, each participant was given a Participant Information Sheet (Appendix F) explaining the experiment. Once a participant was happy to proceed and take part in the experiment, they were asked to fill a Background Information Sheet (Appendix G) and a Consent Form (Appendix H). All participant results were recorded anonymously and could not be traced back to a particular participant. Consent Forms were kept on the researchers MMU encrypted machine. Results were recorded on the researcher's machine and a Usability Questionnaire was completed upon completion of the questions by each participant (Table 44) and results saved on the researcher's machine.

After agreeing to take part, and agreeing a suitable time, participants were given a voucher to purchase a drink at one of two cafes within the University. On purchasing a beverage, the participant was asked to sit down and observe their environment for 10-15 minutes. Once the process completed, the participant notified the researcher (who was sat independently) and began to complete the café feedback survey using the FUSION_V1 DS, on the researcher's machine, relating to their experience and visit to the café. During this interaction, the typed user utterances for each answer is run through the DS and compared with the thresholds for the corresponding category. For analysis purposes, each user utterance was taken and compared with the two sentences for each of the *high*, *medium*, and *low* threshold sentences. The similarity is calculated for each sentence pair using FUSE_2.0 and the results are recorded, and the highest similarity rating is noted for each interaction. All dialogue exchanges are recorded in a log for analysis.

Upon completion of the FUSION_V1 scenario questions, participants were asked to complete a short Usability Questionnaire. Each question was measured using a Likert scale, with questions inspired by research conducted by Chen et al. (Chen et al., 2019) and Deriu et al. (Deriu et al., 2021) comparable to those used to typically assess usability of DS . *A Likert scale is a psychometric scale that has multiple categories from which respondents choose to indicate their opinions, attitudes, or feelings about a particular issue or subject* (Nemoto and

Beglar, 2014). The list of questions along with the Likert scale can be seen in Table 44. The aim of conducting the usability questionnaire is to evaluate the performance of the dialogue system in terms of usability and how easy it was to use, as discussed in Section 7.2.3.

METRIC / DESCRIPTION	**RATING**				
	Strongly Disagree	**Disagree**	**Not Sure**	**Agree**	**Strongly Agree**
1. The interaction with the CA system was ***easy, understandable***, and ***visually pleasing***.	**1**	**2**	**3**	**4**	**5**
2. I think that I would need the ***support*** of a ***technical person*** to be able to use this CA system	**1**	**2**	**3**	**4**	**5**
3. The interaction with the CA system was ***correct*** with ***no misunderstanding*** of my response	**1**	**2**	**3**	**4**	**5**
4. I did not notice ***any*** inconsistencies as I used the CA system.	**1**	**2**	**3**	**4**	**5**
5. I felt very ***confident*** using the CA system.	**1**	**2**	**3**	**4**	**5**
6. Overall, I am ***satisfied*** with how easy it is to use this CA system	**1**	**2**	**3**	**4**	**5**
7. The interaction with the CA system is ***credible***, ***realistic*** and ***believable***.	**1**	**2**	**3**	**4**	**5**
8. I felt ***comfortable*** using this CA system	**1**	**2**	**3**	**4**	**5**
9. The goal of the interaction with the CA system was achieved, - I	**1**	**2**	**3**	**4**	**5**

was able to answer ***all questions*** and complete the café feedback.					
10. I needed to ***learn*** a lot of things ***before*** I could start to use this CA system.	**1**	**2**	**3**	**4**	**5**
11. I could use this CA system ***without*** written instructions.	**1**	**2**	**3**	**4**	**5**
12. I believe that a CA system ***could*** be used to answer survey questions in the future.	**1**	**2**	**3**	**4**	**5**
13. I would ***recommend*** using this CA system to a friend.	**1**	**2**	**3**	**4**	**5**

Table 44 - Usability Questionnaire - FUSION DS

7.3.5 - Results and Discussion of Experimental Evaluation (FUSION_V1)

This section will analyse and discuss the experiments conducted in Section 7.3.4 to evaluate the hypothesis. The results of the Usability Questionnaire will be discussed in a separate section (Section 7.3.6).

To analyse the results, a dataset consisting of 288 rows was compiled from all user responses to all questions, along with the semantic similarity measurement for each rule calculated using FUSE_2.0 on a scale of [0, 1]. For the purpose of comparison, the same rules and responses were also fired through STASIS (Li et al., 2006), a traditional SSM. Table 45 shows the results from all 32 participants for the True (T) and False (F) values run for both FUSE_2.0 and STASIS and shows the percentage of correct True hits for FUSE_2.0 compared with that of STASIS. The fuzzy words assigned to each of the thresholds are examined and if the DS has picked up the correct sentence match, then this is counted as a True (T) hit and given a score of 1. If the highest similarity rating has not fallen under the correct threshold of words, then it is classed as a False (F) hit and given a score of 0.

Category	FUSE_2.0 True	FUSE_2.0 T%	FUSE_2.0 False	FUSE_2.0 F%	STASIS True	STASIS T%	STASIS False	STASIS F%
Q1) Size/Distance	26.00	81.25	6.00	18.75	20.00	62.50	12.00	37.50
Q2) Temperature	31.00	96.88	1.00	3.13	21.00	65.63	11.00	34.38
Q3) Brightness	27.00	84.38	5.00	15.63	27.00	84.38	5.00	15.63
Q4) Age	24.00	75.00	8.00	25.00	17.00	53.13	15.00	46.88
Q5) Speed	31.00	96.88	1.00	3.13	26.00	81.25	6.00	18.75
Q6) Strength	24.00	75.00	8.00	25.00	16.00	50.00	16.00	50.00
Q7) Frequency	27.00	84.38	5.00	15.63	14.00	43.75	18.00	56.25
Q8) Level of Membership	31.00	96.88	1.00	3.13	23.00	71.88	9.00	28.13
Q9) Worth	32.00	100.00	0.00	0.00	26.00	81.25	6.00	18.75
Average %True Rate	**FUSE: 87.85%**				**STASIS: 65.97%**			

Table 45 - Results of FUSION_V1 - FUSE_2.0 vs. STASIS

As seen from the results in Table 45, FUSE_2.0 has an average True rating of 87.85% and STASIS has an average True rating of only 65.97%. The average True rating represents the total number of correctly fired rules that are also correctly matched with the user utterances and are, therefore a True hit. These results show that the fuzzy dictionary of words modelled within the FUSE_2.0 categories, increase the similarity rating when compared with human utterances, as opposed to just crisp values with STASIS.

Figure 33 shows the percentage of correctly matched user utterances using FUSE_2.0 and STASIS. Each question is designed to represent a separate fuzzy category. Although STASIS does not have a fuzzy dictionary and no categories and only relies on WordNet (Miller, 1995), it can still be used in this scenario to compare the similarity with the AHR measures. It can further be seen in Figure 33 that for all the nine categories, apart from Brightness (Q3), FUSE_2.0 always resulted in a higher True rating than STASIS, this had proven that it has a higher number of True matches under the correct threshold. For Q3 (Brightness), both FUSE_2.0 and STASIS scored the same, meaning they both fired the same number of correct thresholds.

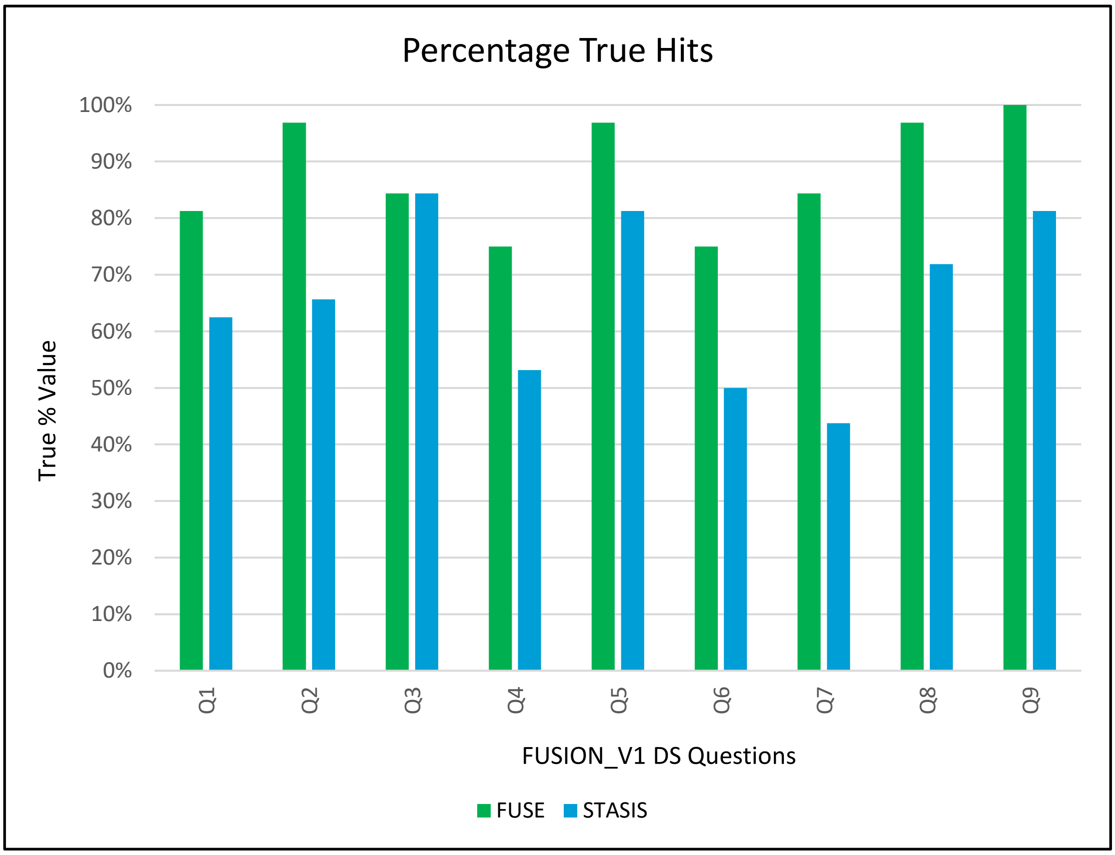


Figure 33 - Percentage of True values - FUSE_2.0 vs. STASIS

Overall, the results have shown that a DS that utilises the FUSE_2.0 measure to determine which rule fires, provides a higher average True rating using fuzzy words as opposed to STASIS that only uses crisp values. There was an improvement of 21.88% in the average True rating based on the results in Table 45, when compared to STASIS, where fuzzy words are not taken into consideration. However, there were some rules that did not fire correctly, and this section provides some in-depth analysis of those rules to feed into future work on the system.

In total, 8 (out of 288) of the user utterances contained some numerical responses as well as just words; an example is shown below of an instance where the DS asked the question relating to the fuzzy category *Age*:

Q4) Using descriptive words, how would you describe the age of the barista that served you?

User Utterance: *The physical appearance of the barista tells that she was in her 30's*

Both FUSE_2.0 and STASIS picked this up as belonging to the *low* threshold, consisting of words such as *baby*, *young*, *child*, etc; when according to the two English language experts, it should be in the *medium* threshold containing words such as *adult*, *middleaged*, *grownup* etc. Figure 34 shows the Threshold for the *Age* category.

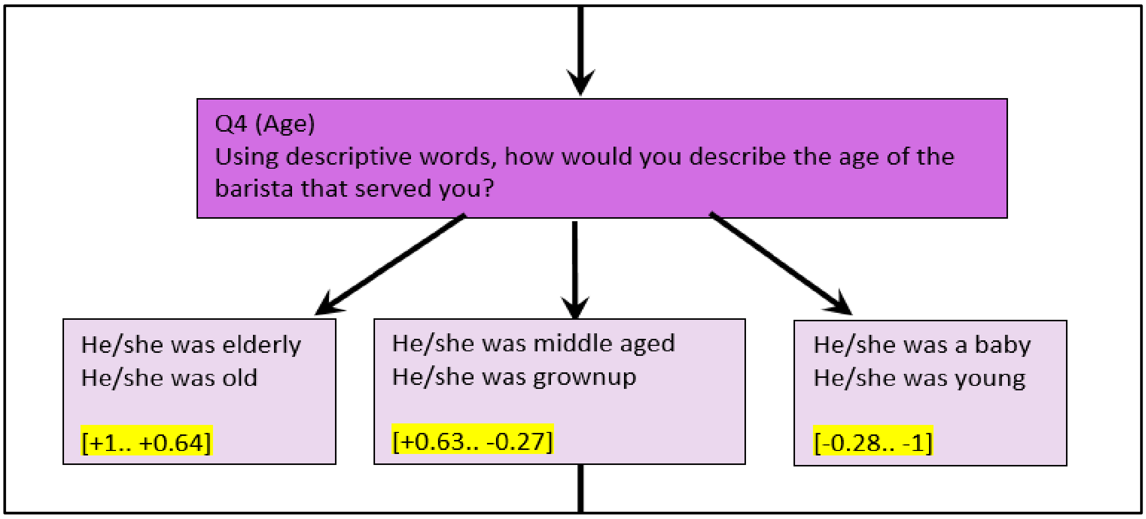


Figure 34 - Age Threshold - FUSION_V1

On the other hand, when the DS asked a question relating to the fuzzy category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of the queue?

User Utterance: *The size of the queue was 2-3 people long with a wait time of no longer than 1 minute.*

Both FUSE_2.0 and STASIS picked this up as being in the *medium* threshold, containing words such as *average*, *standard*, *middle*, and even though numbers were used in place of descriptive words as required, the two English language experts both agreed that this can be classed as a True hit, and it is in the correct threshold. Figure 35 shows the threshold for the *Size/Distance* category.

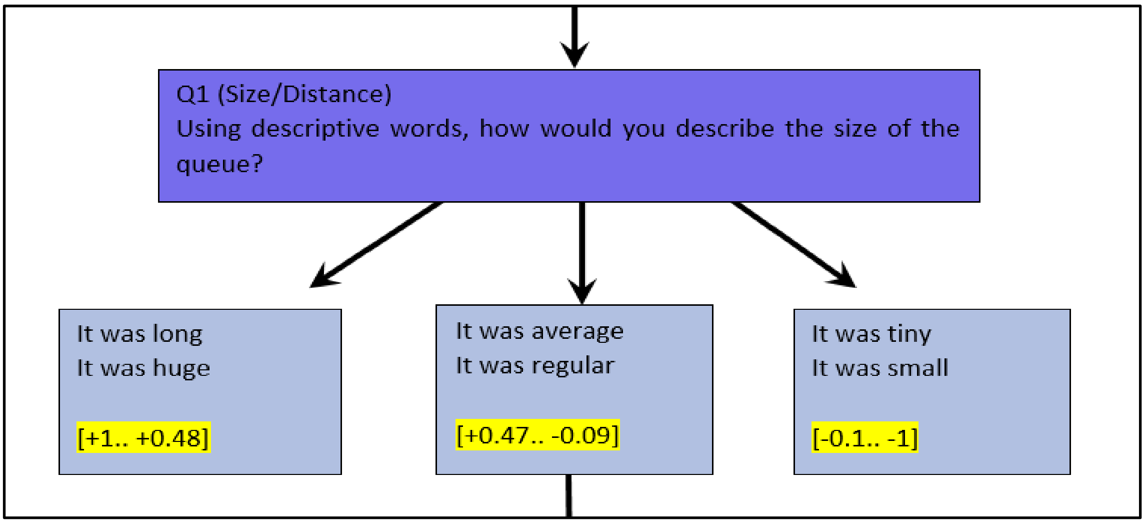


Figure 35 - Size/Distance Threshold - FUSION_V1

Neither FUSE_2.0 nor STASIS were able to deal with the effect of the inclusion of negation words within utterances. For example, when the DS asked the question relating to the fuzzy category *Brightness*:

Q3) How would you describe the brightness of the cafe?

User Utterance: *The light level of the cafe is not bright*

Both FUSE_2.0 and STASIS picked this up as belonging to the *high* threshold because of the word *bright*, when in effect due to the presence of the word *not*, it actually means it was dark. Therefore, is this case, the correct rule category did not fire (i.e., bright was identified as being in the *high* threshold by the English language experts, but the presence of the word *not* would contradict this and it should be the in the *low* threshold). Figure 36 shows the threshold for *Brightness* category.

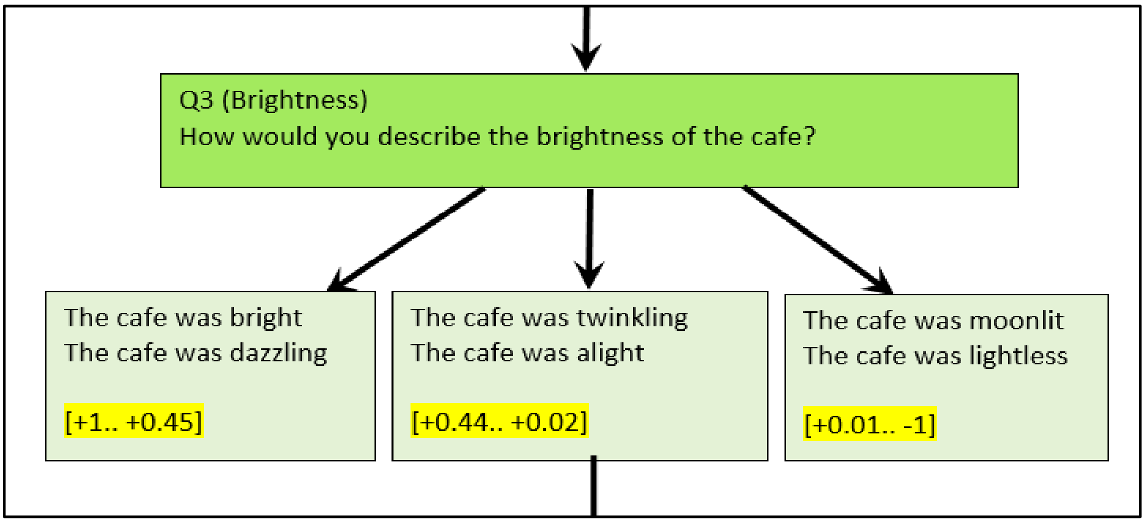


Figure 36 - Brightness Threshold - FUSION_V1

An additional example of negations leading to an incorrect rule firing was, when the DS asked the question relating to the fuzzy category *Strength*:

Q6) Looking up from your screen to the first person you see, how would you describe their physical strength?

User Utterance: *I would describe them as lean and not very strong.*

Both FUSE_2.0 and STASIS picked this up as belonging to the *high* threshold due to the word *strong* (and had an increased intensity in FUSE_2.0 due the presence of the hedge word *very*), when in fact because of the use of the word *not* it actually should belong to the *low* or *medium* threshold, and this was also confirmed by the two English language experts. Figure 37 shows the threshold for the *Strength* category.

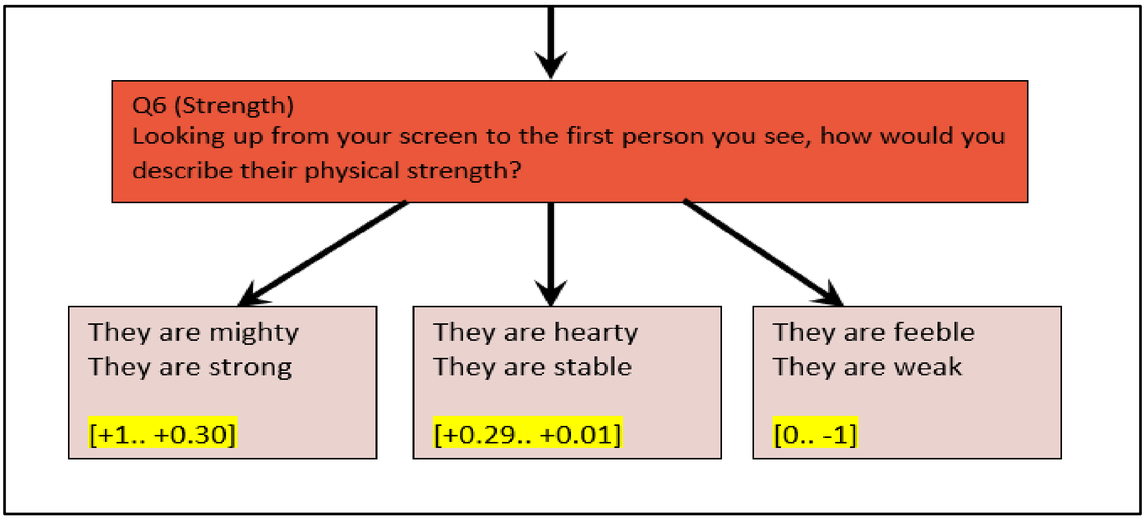


Figure 37 - Strength Threshold - FUSION_V1

There were some instances where FUSE_2.0 correctly matched a rule and STASIS did not. One example of this is when the DS asked the question relating to the fuzzy category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of the queue?

User Utterance: *The size of the queue was huge.*

FUSE_2.0 picked this up as belonging to the *high* threshold with a similarity value of ((D1) *It was long: 0.57554*), and STASIS picked this up as belonging to the *low* threshold, with a similarity value of ((D3) *It was small: 0.53459*). The *high* threshold is correct, since it holds words such as *big*, *massive* and *huge*. Although the difference in the two similarity ratings are small, it is down to the fact that the high threshold actually holds the word *huge* therefore this is the threshold it must fall under for it to be a True hit (Adel et al., 2018). Figure 38 shows the threshold for the *Size/Distance* category.

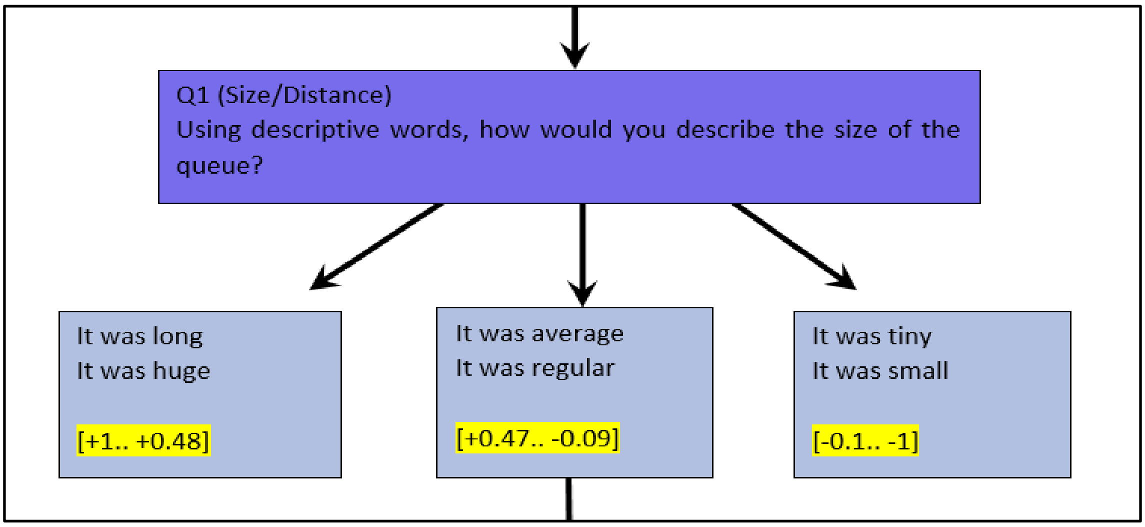


Figure 38 - Size/Distance Threshold - FUSION_V1

An instance when STASIS correctly matched a rule and FUSE_2.0 did not, when the DS asked the question relating to the category *Brightness*:

Q3) How would you describe the brightness of the cafe?

User Utterance: *It was fairly bright*

STASIS picked this up as belonging to the *high* threshold with a similarity value of ((D1) *The cafe was bright: 0.36442*), and FUSE_2.0 picked this up as belonging to the *medium* threshold with a similarity value of ((D2) *The cafe was luminous: 0.67367*). The *high* threshold is correct as it holds words such as *sunny*, *radiant* and *bright*. Figure 39 shows the threshold for the *Brightness* category.

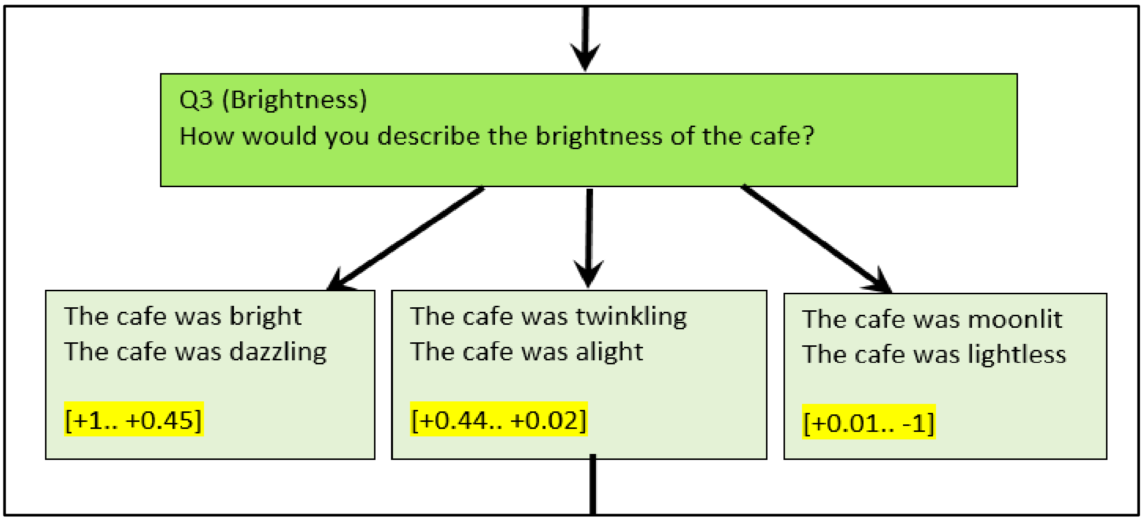


Figure 39 - Brightness Threshold - FUSION_V1

7.3.6 - Usability Questionnaire Evaluation (FUSION_V1)

All participants completed a short usability survey comprising of 13 Likert scale questions, following completion of the task. Table 46 shows the results of the usability questionnaire that each participant filled upon completion of the FUSION_V1 experiment. The aim of a usability questionnaire is to measure the performance of the system and be able to get a better insight as to where the system performed well and where it did not, and how users felt about using a dialogue system as described in Section 7.2.3.

It can be seen from the results that:

- 91% found the system easy to interact with and intuitive to use (Q1) (sum of Agree and Strongly Agree).
- 90% of participants reported no inconsistences when using the system (Q4) (sum of Agree and Strongly Agree).
- 94% of participants did not need the support of a technical person to use FUSION_V1 (Q2) (sum of Strongly Disagree and Disagree).

Overall results show that the inclusion of a FSSM into the DS did not appear to negatively affect the usability of the system, since 94% of participants felt that a DS could be used as a mechanism to answer survey questions in the future(Q12) (sum of Agree and Strongly Agree).

METRIC / DESCRIPTION	**RATING**				
	Strongly Disagree (1)	**Disagree (2)**	**Not Sure (3)**	**Agree (4)**	**Strongly Agree (5)**
1. The interaction with the CA system was ***easy, understandable***, and ***visually pleasing***.	0	2	1	13	16
2. I think that I would need the ***support*** of a ***technical person*** to be able to use this CA system	16	14	1	0	1

3. The interaction with the CA system was ***correct*** with ***no misunderstanding*** of my response	0	1	2	8	21
4. I did not notice ***any*** inconsistencies as I used the CA system.	0	2	1	8	21
5. I felt very ***confident*** using the CA system.	0	0	1	11	20
6. Overall, I am ***satisfied*** with how easy it is to use this CA system	0	1	1	9	21
7. The interaction with the CA system is ***credible***, ***realistic*** and ***believable***.	0	1	3	13	15
8. I felt ***comfortable*** using this CA system	0	0	0	11	21
9. The goal of the interaction with the CA system was achieved, - I was able to answer ***all questions*** and complete the café feedback.	0	0	0	4	28
10. I needed to ***learn*** a lot of things ***before*** I could start to use this CA system.	26	4	1	1	0
11. I could use this CA system ***without*** written instructions.	0	4	5	8	15
12. I believe that a CA system ***could*** be used to answer survey questions in the future.	0	0	2	6	24
13. I would ***recommend*** using this CA system to a friend.	0	1	1	8	22

Table 46 - Usability Results for FUSION_V1

7.3.7 - FUSION_V1 Conclusion

The FUSION_V1 experiment has described the development of a simple linear DS that incorporated the FUSE_2.0 semantic similarity algorithm. The semantic similarity of user utterances and rules were compared using both FUSE_2.0 and STASIS to determine which of the three rules in each category would fire. The results show that the average True rating of FUSE_2.0 is 87.85% which is an improvement of 21.88% when compared with STASIS rule firing rating of (65.97%).

Given the original research question in Section 7.3.1, and the hypothesis for this experiment, it can be concluded that a Fuzzy Semantic Similarity Measure (FSSM) can be incorporated into a dialogue system to improve the success rate of utterance - response matches from a user when compared with a traditional short text SSM. A weakness of utilising FUSE_2.0 was its inability to deal with negation values such as '*not*' within the dialogue, which caused misfiring of rules thus reducing the overall True rating values returned.

Despite the simplicity of the DS, several issues have been recognised. Firstly, neither measure (FUSE_2.0 nor STASIS) were able to produce correct rule firings when a negation word was used to form part of the utterance. Although hedges had been considered as an addition to the FUSE_2.0 fuzzy dictionary (Adel et al., 2018), negation words were not included in the similarity calculation within FUSE_2.0. Secondly, FUSE_2.0 is very much dependent on the fuzzy dictionary created in previous work, which was generated as a result of many empirical experiments (Adel et al., 2018), where humans rated words within categories and then within the context of general sentences. Section 7.4 will include the evaluation of a second, more substantial prototype of the DS, referred to as FUSION_V2, which will incorporate negation values, first discussed in Section 6.4, and the fuzzy influence factor using the FUSE_4.0 algorithm.

7.4 - FUSION_V2 Dialogue System Using FUSE_4.0

7.4.1 - Aim and Purpose

This section describes the creation and evaluation of a second prototype of the DS referred to as FUSION_V2, which utilises the FUSE_4.0 measure (as described in Section 6.5) to match human utterances to a set of fuzzy phrases with a rule-based system. The aim of FUSION_V2 is to improve the robustness of rule matching within the DS based on the findings of FUSION_V1, by using the FUSE_4.0 algorithm (Adel et al., 2021).

The FUSION_V2 dialogue system will look to tackle the following issued that were raised in Section 7.3 using FUSION_V1:

- Tackle negation issues and the presence of negative logical operations such as '*not*'
- Introduce the fuzzy influence factor to tackle fuzzy words present in a sentence pair which do not belong to the same fuzzy category of FUSE_4.0

This experiment conducted in this section will further contribute to the research question:

RQ2. Can a Type-2 FSSM be embedded into a Q&A dialogue system with an improved success rate of utterance - response matches compared to traditional Semantic Similarity Measures (SSM)?

The hypothesis for this experiment is:

H_0: A FSSM used in a DS will improve success rate of utterance - response matches compared with a traditional short text SSM?

7.4.2 - FUSION_V2 Scenario

Due to the Covid-19 global pandemic and the implications of shutting down many universities and offices, it meant that people were forced to work from home (WFH) on very short notice. This in turn meant certain alterations had to be made to people's homes and habits to allow these new adjustments to their working conditions. Therefore, the FUSION_V2 DS was designed around this scenario to ask participants questions relating to their WFH conditions.

To proceed with the FUSION_V2 WFH scenario, two sets of questions were created (Set 1 and Set 2) as part of this experimental methodology, both consisting of nine questions. Each question related to one of the nine fuzzy categories of FUSE_4.0 with each question being validated by an English language expert. It was decided to create two sets of questions to firstly build a bigger database of responses and secondly to have the chance to run more fuzzy responses with FUSION_V2. Table 47 and Table 48 show the questions from the two sets relating to the FUSION_V2 WFH scenario. This scenario was chosen as it was an applicable topic at the time of the COVID-19 pandemic which resulted in lockdowns. The majority of the world's population were facing this change to their working environment, making it very relatable to most, easing the recruitment of participants to express their opinion on this newly obtained experience.

Question	Category	Question Asked
Q1	**Size/Distance**	Using descriptive words, how would you describe the size of your current working environment?
Q2	**Temperature**	Using descriptive words, how would you describe the temperature of your current working environment?
Q3	**Brightness**	Using descriptive words, how would you describe the lighting of your current working environment?
Q4	**Age**	Using descriptive words, how would you describe your current age?
Q5	**Speed**	Using descriptive words, how quickly did you adapt to your current working environment?
Q6	**Strength**	Using descriptive words, how would you describe your current physical state?
Q7	**Frequency**	Using descriptive words, how frequently do you take breaks when working? (remember we are not asking about time)
Q8	**Level of Membership**	Using descriptive words, how closely does your current working environment resemble your office environment?
Q9	**Worth**	Using descriptive words, how satisfied are you with your current working environment conditions?

Table 47 - Set 1 FUSION_V2 WFH Questions

Question	Category	Question Asked
Q1	**Size/Distance**	Using descriptive words, how would you describe the distance of your computer/laptop from yourself?
Q2	**Temperature**	Using descriptive words, how would you describe the temperature of your current machine (laptop/PC) that you are using?
Q3	**Brightness**	Using descriptive words, how would you describe the brightness of your display monitor?
Q4	**Age**	Using descriptive words, how would you describe the age of your machine (laptop/PC) that you are using?
Q5	**Speed**	Using descriptive words, how quickly would you say your machine (laptop/PC) turns on?
Q6	**Strength**	Using descriptive words, think back to the last person you met, how would you describe their physical state?
Q7	**Frequency**	Using descriptive words, how frequently do you use your machine (laptop/PC) to work from home?
Q8	**Level of Membership**	Using descriptive words, how well did you adapt to working from home?
Q9	**Worth**	Using descriptive words, how satisfied are you at present with the current work furniture you use for the purpose of working from home? (chair, stool, sofa, bed, desk, table etc)

Table 48 - Set 2 FUSION_V2 WFH Questions

Calls for participation were advertised via social media platforms and once a participant expressed interest, they were sent a Participant Information Sheet (Appendix K) explaining the experiment. Once a participant was happy to proceed and take part in the experiment, they were asked to fill a Background Information Sheet (Appendix L) and a Consent Form (Appendix M). All participant results were recorded anonymously and could not be traced back to the participant. Consent forms were kept on the researcher's MMU encrypted machine. Results were recorded on the researcher's machine and a Usability Questionnaire was completed upon completion of the questions by each participant (Discussed in Section 7.42.3) and results saved on the researcher's machine.

To comply with social distancing measures, face to face experimentation were not able to go ahead in person like the FUSION_V1 experiment. Therefore, reasonable adjustments had to be made to allow the experiment to proceed safely. To overcome this issue, Microsoft Teams was used to establish a connection with each participant for each experiment once a suitable date and time was agreed between the researcher and participant. The FUSION_V2 DS was opened on the researcher's machine and control was given to the participant via MS Teams, who was then able to run the dialogue system and complete the questions without interference from the researcher.

7.4.3 - Design and Scripting

As with FUSION_V1, the words in each category were broken down into 3 thresholds of *high*, *medium,* and *low* for FUSION_V2 and using the help of an English language expert each category was broken down into these three thresholds and two prototypical sentences were created per subcategory for each set of questions.

Figure 40 shows a partial example of the thresholds for Question 1 relating to the *Size/Distance* category for Set 1 of the questions. The thresholds for all nine fuzzy categories for both Set 1 and Set 2 can be found in Appendix I and Appendix J respectively.

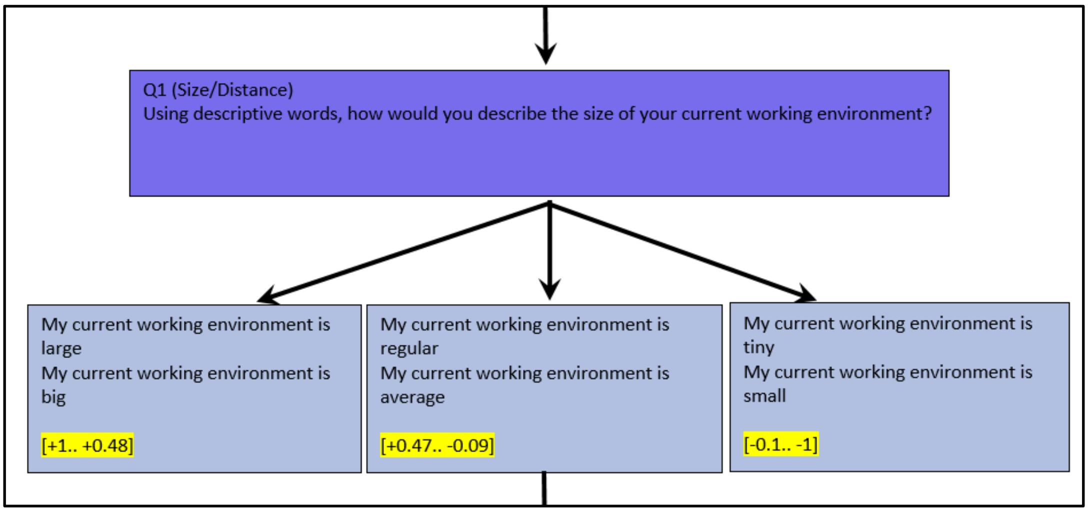

Q1 (Size/Distance)
Using descriptive words, how would you describe the size of your current working environment?
My current working environment is large
My current working environment is big
[+1.. +0.48]
My current working environment is regular
My current working environment is average
[+0.47.. -0.09]
My current working environment is tiny
My current working environment is small
[-0.1.. -1]

Figure 40 - Size/Distance Threshold - FUSION_V2 - Set1

7.4.4 - Experimental Evaluation Methodology

Following Manchester Metropolitan Universities ethical approval process (Ethos number: 11759), call for participants was placed through advertising on social media platforms. Although a sample size of 32 participants were required to allow the results to be statistically significant (J. O'Shea et al., 2013), 35 participants volunteered to take part in this experiment.

Once all 35 participants had completed both Set 1 and Set 2 of FUSION_V2, a dataset consisting of all user responses to the nine questions in each set, along with the semantic similarity measurement for each rule calculated using FUSE_4.0 on a scale of [0, 1] was compiled and the clean-up and analysis process of the results could begin. This was done in five phases which will be described below.

7.4.4.1 - Phase 1

When completing the two sets of questions using FUSION_V2, the users were asked not to use any numbers in their answers, but rather use descriptive words relating to each category. Despite this specific given instruction, some users still used numbers as part of their answers to some of the questions. One of the clean-up processes involved taking any answers that had numbers and trying to substitute them with words. Using an English language expert for help, each numerical answer was taken, and a fuzzy word closest to that numerical value was used to substitute the number. This was done to allow a larger set of natural language dialogue statements to be captured covering the nine categories since the FUSE_ 4.0 algorithm was not designed to handle numerical values. One example is the user response [*I am <u>48</u> years old*], with the help of the English language expert and the use of the English Dictionary (Oxford English Dictionary, 2021) it was agreed that the age 48 would fall under the *middleaged* range and so this sentence was changed to [*I am <u>middleaged</u>*].

7.4.4.2 - Phase 2

The second phase was to take any words that were synonyms to words already present in the fuzzy dictionary but did not exist in the fuzzy dictionary and trying to establish what word measure they would be given, to further allow a larger set of natural language dialogue statements to be captured covering the nine categories of FUSE_ 4.0. This was done with the help of the English language expert and the use of the English Dictionary (Oxford English Dictionary, 2021). An example of this is the user response [*Much warmer than at work and better air quality*], the word *warmer* was not present in the fuzzy dictionary under the category *Temperature*, but the word *warm* was present with the value of (0.480969). Therefore, the word *warmer* was also added to the *Temperature* category with the same value as (0.480969).

7.4.4.3 - Phase 3

The third phase of the clean-up process was to take words that were written by the participants that could be classed as fuzzy but were not present in the fuzzy category. In order to do this, and with the help of the English language expert, and the use of the Oxford English Dictionary (Oxford English Dictionary, 2021), each fuzzy word not present in the fuzzy dictionary was taken and the closest synonyms present in the fuzzy dictionary was used and the numerical value of that fuzzy word was taken and given to this specific word not present in the fuzzy dictionary with the word itself also being added to the fuzzy dictionary. An example of this is the user response [*When working I sometimes take a break every couple of hours*]. The word *sometimes* was not present in the *Frequency* category, but the Oxford English Dictionary (Oxford English Dictionary, 2021) stated that it was a synonym for the word *seldom*. *Seldom* has a value of (-0.365) in the *Frequency* category, and the English language expert also agreed that this word was a synonym for *sometimes*, and thus the word *sometimes* was also added to this category with the same value of (-0.365). In this way, each fuzzy category was expanded to include any words not already present in the fuzzy dictionary. Table 49 shows the number of words added to each of the nine fuzzy categories following Phase 3.

Categories	No. of Initial Words Per Category	No. of New Words Added	No. of Concluding Words Per Category
Size/Distance	91	3	94
Temperature	36	6	42
Age	42	1	43
Frequency	48	3	51
Level of Membership	31	2	33
Worth	61	5	66
Brightness	27	10	37
Strength	26	3	29
Speed	23	13	36

Table 49 - Phase 3 Word Additions to FUSE_4.0

7.4.4.4 - Phase 4

The fourth phase involved devising a methodology to correctly interpret the implications of a negation word on a fuzzy word with the context of a user utterance.

Negation was first identified as a weakness of the FUSE algorithm in Section 6.4 under FUSE_3.0, where three parameters were identified that dealt with negation class in fuzzy sets (Zadeh, Sugeno and Yager) (Section 6.4.1). To determine which method provided the best results in terms of the highest correlation to human ratings a short experiment was conducted to test the three measures of Zadeh (Cox, 1994), Yager and Sugeno (Klir and Folger, 1988).

The original '*not*' operator introduced by Zadeh (Cox, 1994) is implemented by taking one minus the membership value $\sim\mu_A(x) = (1 - \mu_1)$.

The Yager (Klir and Folger, 1988) class is defined as:

$$\sim\mu_A(x) = (1 - \mu_A(x)^k)^{\frac{1}{k}}$$

Equation 25 (Source: Klir and Folger, 1988)

Where the class function *k* is generally in the range [>0, <5]. The class function performs the standard Zadeh complement (which is found when *k*=1). The class membership in the Yager complement, provides a convenient and flexible method of adjusting the strength (class parameters) of the fuzzy '*not*' operator. Klir suggests using the following class strengths for testing [*k* = 0.5, *k* = 1, *k* = 2, *k* = 5] (Klir and Folger, 1988).

The Sugeno (Klir and Folger, 1988) class is defined as:

$$\sim\mu_A(x) = \frac{1 - \mu_A(x)}{1 + k\mu_A(x)}$$

Equation 26 (Source: Klir and Folger, 1988)

In this case the class parameters are in the range [-1, ∞]. Klir suggests using the following class strengths (class parameters) for testing [*k* = 10, *k* = 2, *k* = 0, *k* = -0.5, *k* = -0.9] (Klir and Folger, 1988).

7.4.4.4.1 - Preliminary Experiment to Evaluate Negation Operators with FUSE_4.0

Originally the sentence "*The light level of the cafe is not bright"* which was a response from a participant under the FUSION_V1 experiment, (discussed in Section 7.3) scored the highest similarity rating with the sentence *The cafe was light*. This sentence was under the *high* threshold for this category as shown in Figure 41.

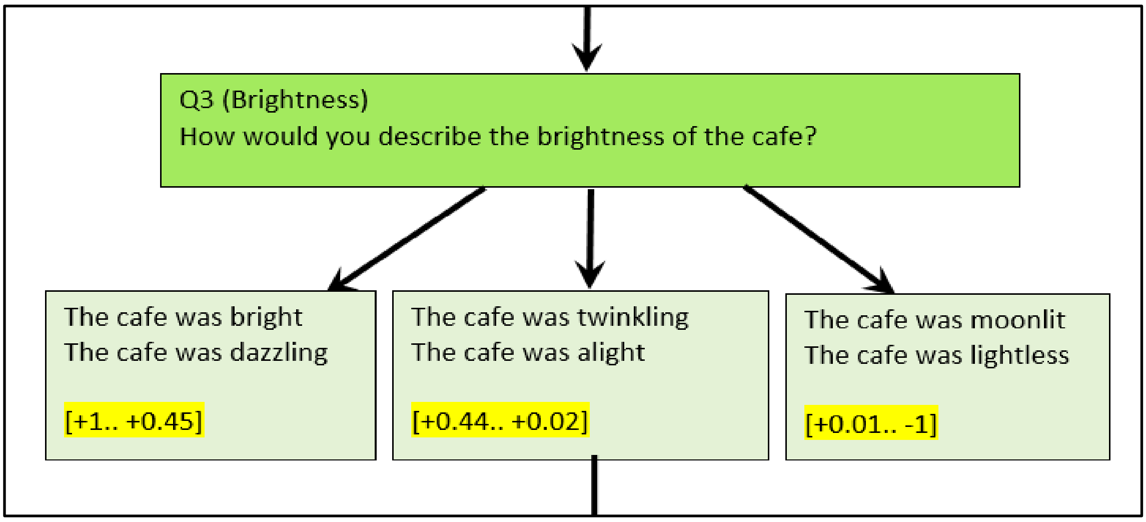


Figure 41 - Brightness Threshold - FUSION_V1

However, due to the presence of the negation word *'not'* in the participants sentence, it actually means the café is *not bright* and optimal results would be to score similarity with the *low* threshold sentences (*The cafe was moonlit* or *The cafe was lightless*). Therefore, a preliminary experiment was conducted on this sample sentence to investigate which of the Zadeh, Sugeno and Yager (and their various class strengths for Sugeno and Yager) negation classes (Section 7.4.4.4) returned the most accurate result. The correct result must fall in the threshold of *low* (to allow Sentence 1 to correctly match with Sentence 2) for FUSION_V1. This in turn would allow the most optimal class be applied to the FUSE_4.0 algorithm used within FUSION_V2.

To carry out this preliminary experiment a dataset of six sentences was created using the mentioned sentence above. Table 50 shows the participant responses in the second column (Sentence 1), the six threshold sentences for the *Brightness* category in column three (Sentence 2), and column four indicates which threshold each sentence from column three (Sentence 2) belongs to (*High*, *Medium* or *Low*).

	Sentence 1	**Sentence 2**	**Threshold**
SP 1	The light level of the cafe is not bright	The cafe was bright	High
SP 2	The light level of the cafe is not bright	The cafe was dazzling	
SP 3	The light level of the cafe is not bright	The cafe was twinkling	Medium
SP 4	The light level of the cafe is not bright	The cafe was alight	
SP 5	The light level of the cafe is not bright	The cafe was moonlit	Low
SP 6	The light level of the cafe is not bright	The cafe was lightless	

Table 50 - Not Test Sentence Pairs

Table 51 shows the results of the experiment conducted on the six sentence pairs from Table 50 presented on a scale of [0, 1]. The yellow highlighted value in each column represents the highest similarity score for that sentence pair and which class parameter it was obtained from (Zadeh, Sugeno, Yager). The original defuzzified value for the word *bright* is (0.57), and using the Yager class (k = 0.5) the word *not bright* is given a measure of (-0.8799) which provided the best results, shown in Table 51, highlighted in pink, as it matched with the correct sentence pair (SP6) with a similarity rating of (0.9277).

Therefore, the Yager class with a strength of (k = 0.5) was used with FUSE_4.0 within the FUSION_V2 DS to calculate the similarity for any sentences that presented a '*not'* immediately before the fuzzy word. For example, the word *dazzling* (with a rating of 0.6) which belonged to the *Brightness* category, and the user response of *not dazzling* (will have the rating of -0.8984).

Sentence Pair	Original	Zadeh	Sugeno					Yager			
		$1 - \mu_1$	*k* = 10	*k* = 2	*k* = 0	*k* = -0.5	*k* = -0.9	*k* = 0.5	*k* = 1	*k* = 2	*k* = 5
SP 1	0.9770	0.8024	0.7215	0.7811	0.8024	0.9531	0.9643	0.7215	0.8024	0.9643	0.9335
SP 2	0.9778	0.7787	0.7263	0.7647	0.7787	0.9335	0.9453	0.7263	0.7787	0.9453	0.9138
SP 3	0.8772	0.8759	0.6877	0.7740	0.8759	0.9257	0.8240	0.6877	0.8759	0.8240	0.7740
SP 4	0.9660	0.8507	0.7309	0.8069	0.8507	0.9912	0.9373	0.6834	0.8507	0.9373	0.8952
SP 5	0.8363	0.9410	0.8951	0.9770	0.9410	0.8459	0.7528	0.8951	0.9410	0.7528	0.7128
SP 6	0.6599	0.7552	0.8277	0.9231	0.7552	0.6502	0.5882	0.9277	0.7552	0.5882	0.5523

Table 51 - Not Empirical Experiment

7.4.4.5 - Phase 5

Any utterance that contained the negation *'not'*, which were not present immediately before a fuzzy word, could not be addressed using the method described in Phase 4 (Section 7.4.4.4), therefore it required a different approach. In the fifth phase, again using the Oxford English Dictionary (Oxford English Dictionary, 2021) any utterance that had the word *'not'* present, but it did not appear immediately before a fuzzy word, the help of an English language expert was sought, and this phrase was substituted by the equivalent negative meaning. An example of this is the user response [*I am not really satisfied with my working conditions when working from home*], here the word *really* is present between *not* and *satisfied*, and so the Yager class method approach in Phase 4 cannot be used, instead the phrase *not really satisfied* is replaced with the fuzzy word *dissatisfied*.

This concludes the clean-up process of the user responses. At this point the responses were ready to be analysed using the FUSE_4.0 algorithm. For comparison purposes, the same rules and responses were also fired through STASIS (Li et al., 2006) since STASIS is a traditional short-text SSM and is not able to capture the meaning of fuzzy words (Section 2.5.2).

7.4.5 - Results and Discussion of Experimental Evaluation (FUSION_V2)

To analyse the results, two datasets consisting of 315 rows was compiled of all user responses to all questions, per set of questions (Set 1 and Set 2) along with the semantic similarity measurement for each rule calculated using FUSE_4.0. For comparison purposes, the same rules and responses were also fired through STASIS (Li et al., 2006), a traditional SSM.

7.4.5.1 - Results for FUSION_V2 - Set 1

Table 52 shows the results for Set 1 of the FUSION_V2 experiment on both FUSE_4.0 and STASIS. If the DS picked up the correct sentence match, then this is counted as a True (T) hit and given a score of 1. If the highest similarity rating has not fallen under the correct threshold of words, then it is classed as a False (F) hit and given a score of 0. Based on the results in Table 52, it can be seen that FUSE_4.0 had an average True rating of 84.76% compared to STASIS which had a True rating of just 55.24%. Every single one of the nine fuzzy categories gave a higher True hit when run with FUSE_4.0 as opposed to STASIS.

Category	FUSE_4.0 True	FUSE_4.0 T%	FUSE_4.0 False	FUSE_4.0 F%	STASIS True	STASIS T%	STASIS False	STASIS F%
Q1) Size/Distance	30	85.71	5	14.29	14	40.00	21	60.00
Q2) Temperature	28	80.00	7	20.00	20	57.14	15	42.86
Q3) Brightness	29	82.86	6	17.14	19	54.29	16	45.71
Q4) Age	29	82.86	6	17.14	26	74.29	9	25.71
Q5) Speed	28	80.00	7	20.00	24	68.57	11	31.43
Q6) Strength	29	82.86	6	17.14	4	11.43	31	88.57
Q7) Frequency	31	88.57	4	11.43	28	80.00	7	20.00
Q8) Level of Membership	30	85.71	5	14.29	26	74.29	9	25.71
Q9) Worth	33	94.29	2	5.71	13	37.14	22	62.86
Average %True Rate	**FUSE: 84.76%**				**STASIS: 55.24%**			

Table 52 - Results of FUSION_V2 - FUSE_4.0 vs. STASIS - Set 1

One example of where FUSE_4.0 did not match correctly but STASIS did, was for the category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of your current working environment?

User response: *It is big enough to be able to do my work.*

Although *big* is in the fuzzy category for *Size/Distance*, FUSE_4.0 matched this sentence with the *medium* threshold, even though it should have been returned to the *high* threshold, as can be seen by the thresholds in Figure 42.

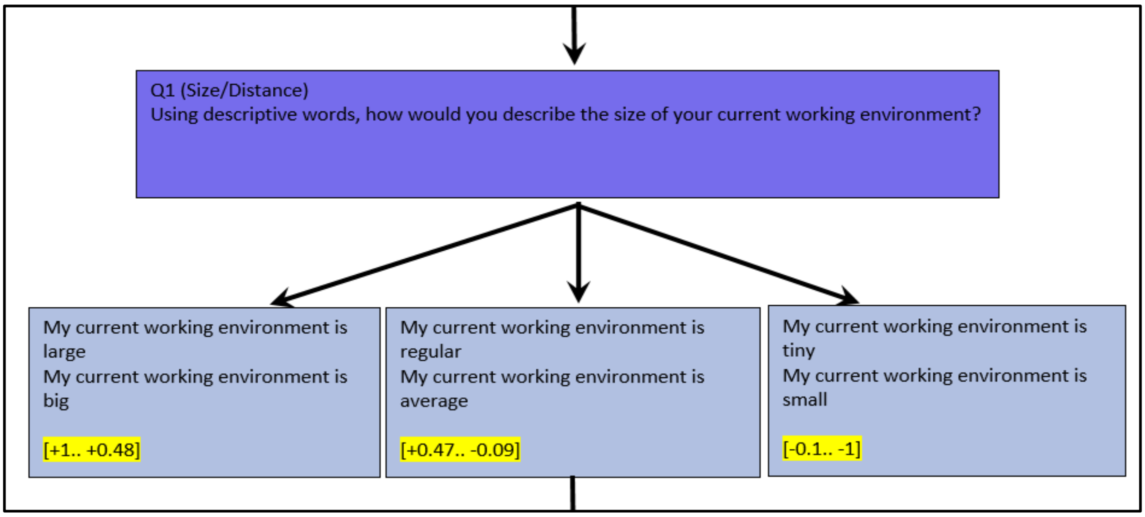

Q1 (Size/Distance)
Using descriptive words, how would you describe the size of your current working environment?
My current working environment is large
My current working environment is big
[+1.. +0.48]
My current working environment is regular
My current working environment is average
[+0.47.. -0.09]
My current working environment is tiny
My current working environment is small
[-0.1.. -1]

Figure 42 - Size/Distance Threshold - FUSION_V2 - Set 1

One example of where STASIS did not match correctly and FUSE_4.0 did was for the category *Brightness*:

Q3) Using descriptive words, how would you describe the lighting of your current working environment?

User response: *Ambient, just right. Reasonably bright, but not dazzling.*

This response uses a negation of '*not'* before the fuzzy word *dazzling*. The threshold for the *Brightness* category for FUSION_V2 Set 1 can be seen in Figure 43. STASIS returned this in the *high* category and matched it with *The lighting is bright*. FUSE_4.0 however correctly matched it to *The lighting is lightless* in the *low* threshold.

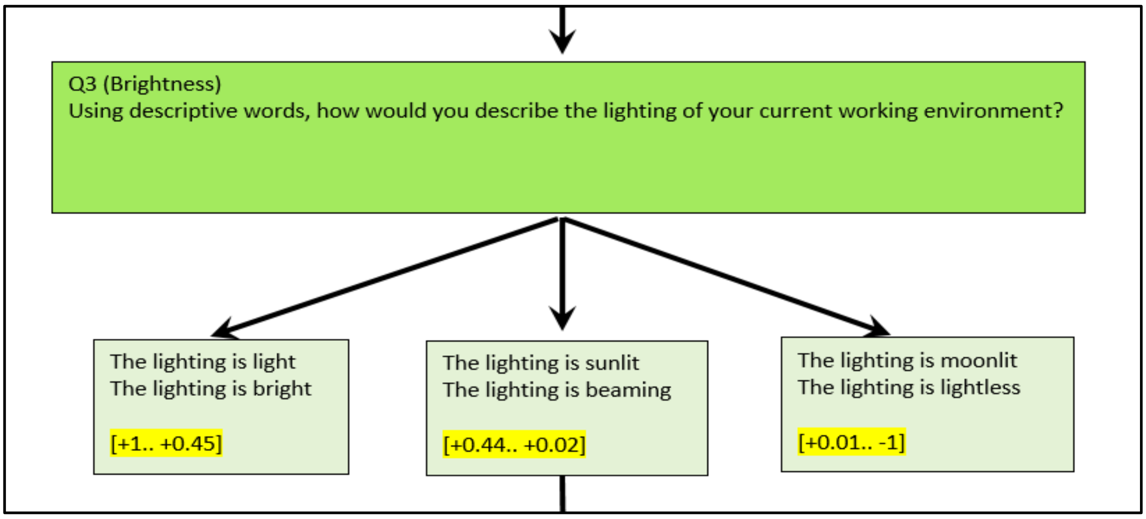


Figure 43 - Brightness Threshold - FUSION_V2 - Set 1

Figure 44 shows the percentage of correctly matched user utterances using FUSE_4.0 and STASIS. Each question is designed to represent a separate fuzzy category. Even though STASIS does not have a fuzzy dictionary and no categories and only relies on WordNet (Miller, 1995), it can still be used in this scenario to compare the similarity with the AHR measures. It can further be proven in Figure 44 that for all the nine categories, FUSE_4.0 always resulted in a higher True rating than STASIS, meaning it has a higher number of True matches that fired under the correct threshold.

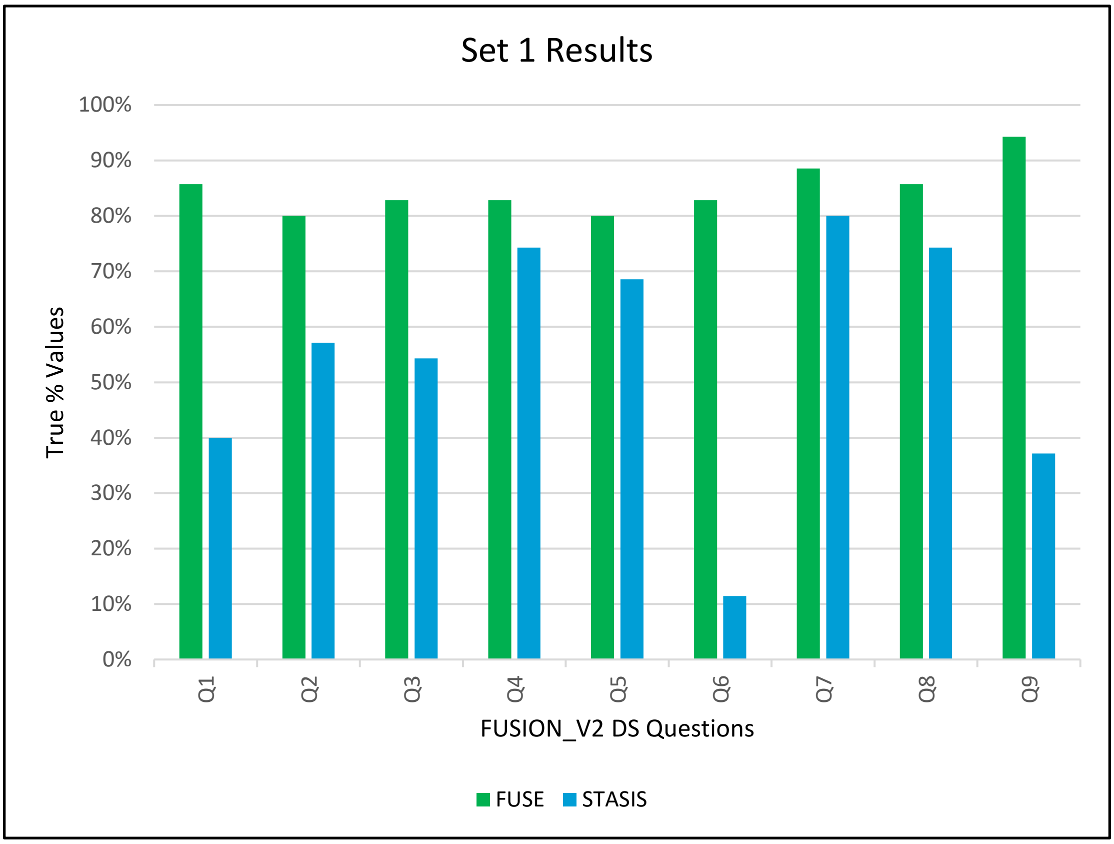


Figure 44 - Percentage of True values - FUSE_4.0 vs. STASIS - Set 1

7.4.5.2 - Results for FUSION_V2 - Set 2

Table 53 shows the results for Set 2 of the FUSION_V2 experiment on both FUSE_4.0 and STASIS. Looking at the results in Table 53, it can be seen that FUSE_4.0 had an average True rating of 93.33% compared to STASIS, which had a True rating of just 62.86%. Every single one of the nine fuzzy categories gave a higher True value when run with FUSE_4.0 as opposed to STASIS, with three of the categories (*Age*, *Frequency* and *Level of Membership*) scoring 100% True rating with FUSE_4.0. *Strength* (Q6) and *Worth* (Q9) categories performed the worst for STASIS with very low True values returned (9 and 11) respectively.

Category	FUSE_4.0 True	FUSE_4.0 T%	FUSE_4.0 False	FUSE_4.0 F%	STASIS True	STASIS T%	STASIS False	STASIS F%
Q1) Size/Distance	30	85.71	5	14.29	25	71.43	10	28.57
Q2) Temperature	32	91.43	3	8.57	23	65.71	12	34.29
Q3) Brightness	32	91.43	3	8.57	16	45.71	19	54.29
Q4) Age	35	100.00	0	0.00	31	88.57	4	11.43
Q5) Speed	33	94.29	2	5.71	26	74.29	9	25.71
Q6) Strength	33	94.29	2	5.71	9	25.71	26	74.29
Q7) Frequency	35	100.00	0	0.00	33	94.29	2	5.71
Q8) Level of Membership	35	100.00	0	0.00	24	68.57	11	31.43
Q9) Worth	29	82.86	6	17.14	11	31.43	24	68.57
Average %True Rate	**FUSE: 93.33%**				**STASIS: 62.86%**			

Table 53 - Results of FUSION_V2 - FUSE_4.0 vs. STASIS - Set 2

One example of where both FUSE_4.0 and STASIS did not match correctly was for the category *Worth:*

Q9) Using descriptive words, how satisfied are you at present with the current work furniture you use for the purpose of working from home? (chair, stool, sofa, bed, desk, table etc)

User response: *The desk is holding up ok.*

Although *ok* is in the fuzzy category for *Worth*, both FUSE_4.0 and STASIS returned this in the *medium* threshold even though it should have been returned to the *low* threshold since (ok = -0.27586) in the FUSE_4.0 fuzzy dictionary thus making it fall in the *low* threshold, as can be seen in Figure 45.

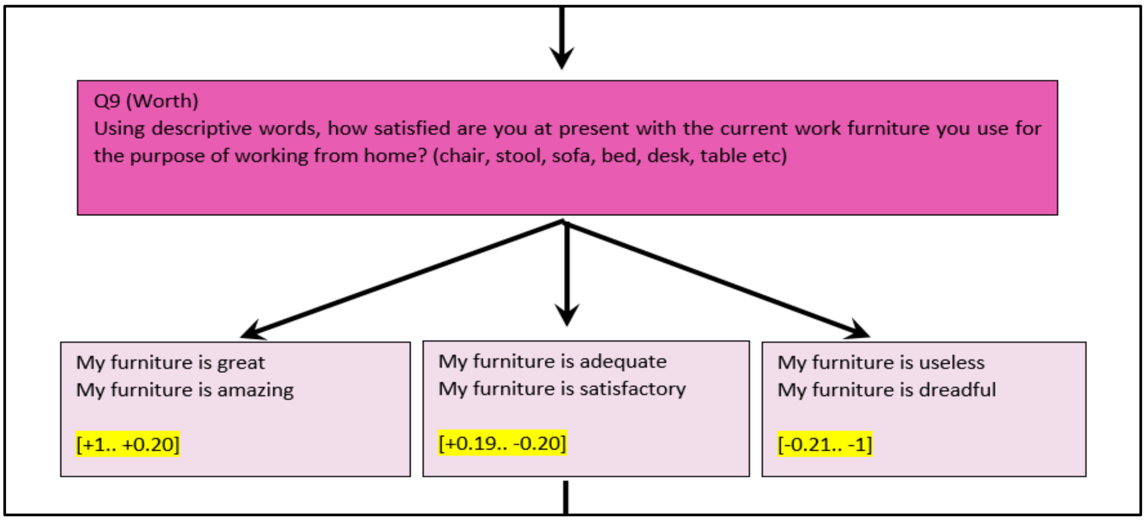

Q9 (Worth)
Using descriptive words, how satisfied are you at present with the current work furniture you use for the purpose of working from home? (chair, stool, sofa, bed, desk, table etc)
My furniture is great
My furniture is amazing
[+1.. +0.20]
My furniture is adequate
My furniture is satisfactory
[+0.19.. -0.20]
My furniture is useless
My furniture is dreadful
[-0.21.. -1]

Figure 45 - Worth Threshold - FUSION_V2 - Set 2

Figure 46 shows the percentage of correctly matched user utterances using FUSE_4.0 and STASIS. It can further be seen in Figure 46 that for all the nine categories FUSE_4.0 always resulted in a higher True rating than STASIS, meaning it has a higher number of True matches that fired under the correct threshold with Q4, Q7 and Q8 obtaining 100% True rating results.

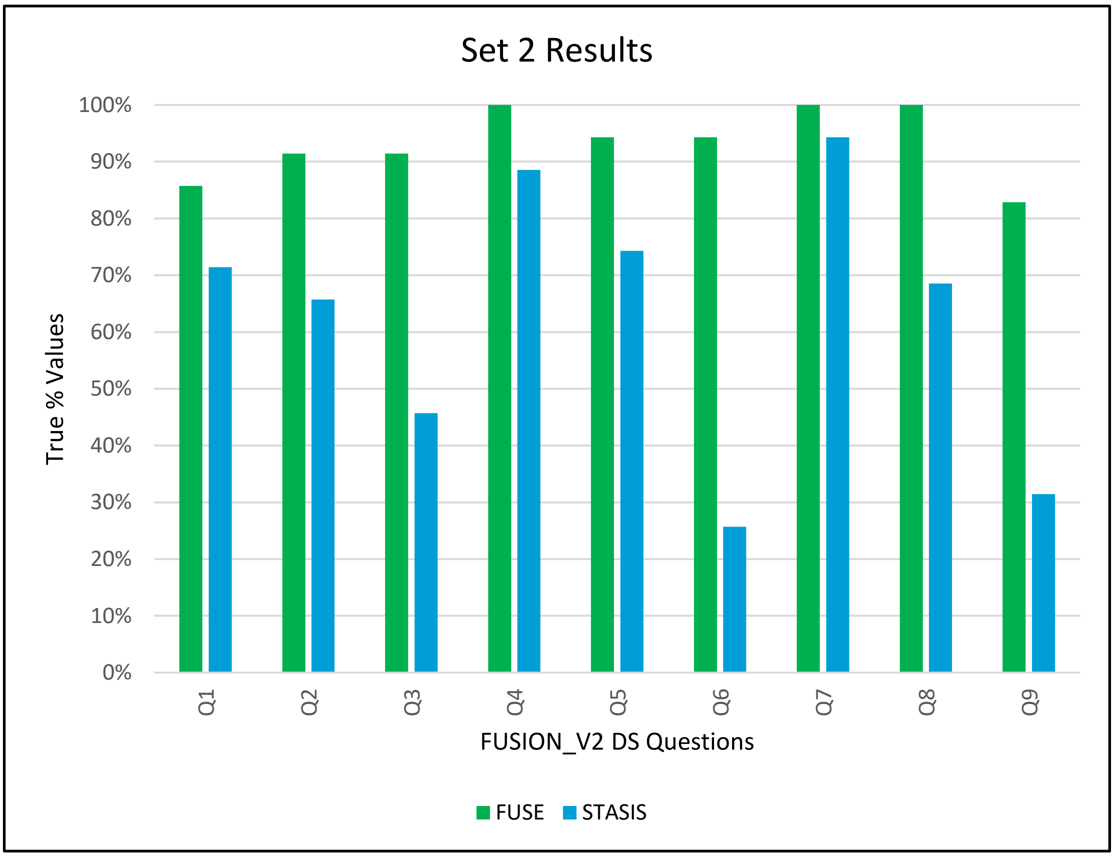


Figure 46 - Percentage of True values - FUSE_4.0 vs. STASIS - Set 2

7.4.6 - Usability Questionnaire Evaluation (FUSION_V2)

All participants completed a short Usability Questionnaire comprising of 13 Likert scale questions, following completion of the task. The questions asked are the same as those used for the FUSION_V1 experiment (section 7.3.6). Table 54 shows the results of the usability questionnaire that each participant filled upon completion of the FUSION_V2 experiment. It can be seen from the results that:

- 97% found the system easy to interact with and intuitive to use (Q1) (sum of Agree and Strongly Agree).
- 85% of participants reported no inconsistences, when using the system (Q4) (sum of Agree and Strongly Agree).
- 78% of participants did not need the support of a technical person to use FUSION_V2 (Q2) (sum of Strongly Disagree and Disagree).

Overall results show that the inclusion of a FSSM into the DS did not appear to negatively affect the usability of the system, even though the FUSION_V2 was used online as opposed to in person due to Covid-19. From Table 54 it can be seen that 87.5% of participants felt that a DS could be used as a mechanism to answer survey questions in the future(Q12) (sum of Agree and Strongly Agree).

METRIC / DESCRIPTION	RATING				
	Strongly Disagree (1)	Disagree (2)	Not Sure (3)	Agree (4)	Strongly Agree (5)
1. The interaction with the CA system was ***easy, understandable***, and ***visually pleasing***.	0	3	1	13	18
2. I think that I would need the ***support*** of a ***technical person*** to be able to use this CA system	17	8	0	5	5
3. The interaction with the CA system was ***correct*** with ***no misunderstanding*** of my response	0	2	6	11	16
4. I did not notice ***any*** inconsistencies as I used the CA system.	0	7	1	10	17
5. I felt very ***confident*** using the CA system.	0	0	5	24	6
6. Overall, I am ***satisfied*** with how easy it is to use this CA system	0	1	5	7	22
7. The interaction with the CA system is ***credible***, ***realistic*** and ***believable***.	1	4	2	12	16
8. I felt ***comfortable*** using this CA system	0	2	1	12	20

9. The goal of the interaction with the CA system was achieved, - I was able to answer ***all questions*** and complete the café feedback.	2	1	1	14	17
10. I needed to ***learn*** a lot of things ***before*** I could start to use this CA system.	10	5	2	2	16
11. I could use this CA system ***without*** written instructions.	1	4	3	11	16
12. I believe that a CA system ***could*** be used to answer survey questions in the future.	0	1	3	13	18
13. I would ***recommend*** using this CA system to a friend.	1	1	5	11	17

Table 54 - Usability Results for FUSION_V2

7.4.7 - FUSION_V2 Conclusion

This section has described the development of a second version of the FUSION dialogue system referred to as FUSION_V2, with the incorporation of negation values and the application of the fuzzy influence factor for the WFH scenario. Two sets of questions (Set 1 and Set 2) were designed, each containing nine questions to reflect the nine fuzzy categories of FUSE_4.0 for the given WFH scenario. The semantic similarity of user utterances and rules were compared using both FUSE_4.0 and STASIS to determine which of the three rules in each category would fire.

A weakness of FUSION_V1 was the lack of ability to deal with negation values such as '*not*' within the dialogue, which caused misfiring of rules. This was explored in FUSION_V2 by the preliminary investigation into the use of fuzzy negation operators within FSSM's. The Yager class (Klir and Folger, 1988) with strength of (k = 0.5) negation complement was used within FUSE_4.0, following a set of empirical experiments to overcome this weakness. The fuzzy influence factor was also introduced within FUSION_V2 as part of FUSE_4.0, which would allow fuzzy measures to be included for fuzzy words present in utterances that don't seldom come from the same fuzzy dictionary of FUSE_4.0. This further contributed to the True (T) rating of the FUSION_V2 experiment and had an improvement over FUSION_V1.

The results show that the average True rating of FUSE_4.0 run with FUSION_V2 for Set 1 is 84.76%, with an improvement of 29.52% compared to STASIS (55.24%), and a True rating for Set 2 of 93.33% with an improvement of 30.47% compared to STASIS (62.86%). The combined average rating of both Set 1 and Set 2 for FUSE_4.0 run with FUSION_V2 is 89.05% which is an improvement of 1.2% compared to the average True rating of FUSE_2.0 run with FUSUON_V1 which was 87.85%. Every single one of the nine fuzzy categories gave a higher True value when run with FUSION_V2 as opposed to STASIS, with three of the categories (*Age*, *Frequency* and *Level of Membership*) even scoring 100% True ratings with FUSE_4.0. The negation factor and fuzzy influence factor of the FUSE_4.0 play a positive role in increasing the number of True values returned for FUSION_V2 and allows the hypothesis for the experiment to be accepted.

7.5 - Suitability of FUSE Embedded Within a Dialogue System

Looking back at the implementation of a sentence similarity measure into a dialogues system, a traditional tactic would have been to use a pattern scripting approach. This is a time consuming and very intricate process as highlighted in Section 7.2.2 with limitations of modification, as any new or modified rule has a knock-on effect on the other rules present and would require a reassessment of the entire script (Michie, 2001).

Therefore, a novel approach was explored, first introduced by O'Shea (K. O'Shea et al., 2009), where traditional pattern matching rules are replaced with short text semantic similarity measures (SSM's). Thus, this approach can reduce the complexity of producing scripts for use with dialogue systems.

The FUSE algorithm was embedded into a dialogue system referred to as FUSION, which was a simple question and answer dialogue system to match user utterances to different categories of responses to each question. This algorithm used a linear sequence of questions, where each questions response has three possible branches of *high*, *medium* or *low*, with the aim of distinguishing between human perceptions of fuzzy words. This is done through user responses to nine questions reflecting the nine fuzzy categories of the FUSE algorithm to assess if the correct rule can be fired, in response to natural language used within the human utterance.

Creating the FUSION dialogue system in this way, allowed greater flexibility, in being able to adapt the DS to any scenario, using any set of questions and any desired thresholds.

The FUSE_4.0 algorithm, the last version of the FUSE algorithm is greatly adapted to be used with the FUSION dialogue system. Since it has a larger fuzzy dictionary than its previous versions, it can handle fuzzy negation and deal with fuzzy influence in a sentence, even when fuzzy words do not belong to the same category in a sentence pair. This greatly improves the accuracy of the FSSM and makes it a suitable algorithm to be used within a dialogue system, allowing semantic measures to be calculated from natural user responses.

7.6 - Conclusion

This chapter has discussed the embedment of the FUSE algorithm into a simple dialogue system referred to as FUSION. The first round of experiments explored FUSION_V1 with the FUSE_2.0 algorithm and used a café scenario to collect a set of user responses to pre-set questions. The FUSION_V1 DS was also run using the STASIS algorithm for comparison purposes between a traditional SSM and a FSSM, as STASIS does not deal with fuzzy words.

Results for FUSION_V1 were promising, and outperformed STASIS with a True (T) rating of 87.85%, an improvement of 21.88% when compared with STASIS. Some of the weaknesses of FUSION_V1 was its inability to deal with negation values such as '*not*'.

A second version of the dialogue system was designed, referred to as FUSION_V2, to test the improvements made to the FUSE algorithm, namely the approach to dealing with negation values, first explored in FUSE_3.0 and the introduction of the fuzzy influence factor in FUSE_4.0.

Due to the global pandemic of Covid-19 and the closure of many universities and workplaces, the majority of people were forced to work from home, and this was the inspiration behind the scenario for FUSION_V2 to collect user experiences relating to adapting to their new working from home conditions. Two separate sets of questions were designed (Set 1 and Set 2), each containing nine questions to reflect the nine fuzzy categories of FUSE_4.0.

The results of the FUSION_V2 experiments were very positive with some categories even scoring 100% True ratings. The results show that the average True rating of FUSE_4.0 for Set 1 is 84.76%, with an improvement of 29.52% compared to STASIS, and a True rating for Set 2 of 93.33% with an improvement of 30.47% compared to STASIS. Overall, the FUSION_V2 experiment was successful, and the coverage of negation and fuzzy influence factor played a positive and improving role in increasing the True ratings.

The combined average rating of both Set 1 and Set 2 for FUSION_V2 is 89.05% which is an improvement of 1.2% compared to the average True rating of FUSE_2.0 run with FUSION_V1 which was 87.85%. Every single one of the nine fuzzy categories gave a higher True rating when run with both FUSION_V1 and FUSION_V2 as opposed to STASIS.

FSSM's play an important part in improving language understanding with an average True rating of 88.65% for FUSION_V1 and FUSION_V2 combined as opposed to STASIS with an average True rating of 61.36%. The difference of 27.29% allows the second and final research question presented in Section 1.3 (***RQ2. Can a Type-2 FSSM be embedded into a Q&A dialogue system with an improved success rate of utterance - response matches compared to traditional Semantic Similarity Measures (SSM)?)*** to be answered positively.

The context of perception-based words does matter when using a FSSM in a dialogue system. Further work could include the introduction of numbers into the dialogue system as currently FUSE_4.0 nor FUSION_V2 cannot deal with numbers in a short text or utterance.

Chapter 8, the final chapter of this thesis, will draw conclusions to this research, by evaluating the FUSE algorithm and revisiting the proposed research questions. A summary of the key contributions this research has made will be discussed and potential future areas of work will be explored.

CHAPTER 8

CHAPTER 8: CONCLUSION AND FUTURE WORK

8.1 - Introduction

The final chapter of the thesis contains a summary of the contributions discussed throughout the previous 7 chapters. Furthermore, it performs critical analysis of the limitations of the work and the improvements it has made. The final section of this chapter illustrates the future paths of research and improvements in this field.

8.2 - Evaluation of FUSE as a Fuzzy Semantic Similarity Measure

The FUSE algorithm was developed over four versions designed in three core phases. Phase 1 investigated and reviewed the modelling of words using Type-2 fuzzy sets, specifically Interval Type-2 fuzzy sets, before developing a methodology for modelling fuzzy words using Interval Type-2 fuzzy sets. Six initial categories were taken from the existing FAST FSSM algorithm (Chandran, 2013), which used Type-1 sets to model its fuzzy words.

However, as part of the research conducted on FAST, it was established that using Type-1 sets to model fuzzy words is incorrect, since Type-1 is crisp. Therefore, initial work involved firstly expanding the number of words in the six categories of FAST (to improve natural language coverage) using human participant experimentation, and secondly to model all words in the six categories using Interval Type-2 fuzzy sets.

This resulted in a fuzzy dictionary which using human participants to offer ratings for fuzzy words allowed the modelling of fuzzy words in each category using Interval Type-2 Fuzzy Sets, using techniques developed by Hao-Mendel known as the HMA approach (Hao and Mendel, 2015). Finally, a set of fuzzy ontologies was created to represent each fuzzy category based on ideas from established SSM's, namely STASIS (Li et al., 2003), the WordNet ontology (Miller, 1995) and FAST (Chandran, 2013).

Phase 2 involved creating the first version of the FUSE algorithm referred to as FUSE_1.0. The design and development of FUSE_1.0 involved short text similarity determined by word similarity, path depth (referred to as the Lowest Common Subsumer and path length in the ontology, with fuzzy word similarity being determined using the fuzzy ontologies and non-fuzzy words using the WordNet ontology. The FUSE_1.0 algorithm was evaluated using three published datasets (MWFD (Chandran, 2013) , STSS-65 (J. O'Shea et al., 2013) and STSS-131 (J. O'Shea et al., 2013)) and result correlation with human ratings was compared with two other measures, STASIS and FAST. Results showed FUSE_1.0 gave a higher correlation with AHR compared with STASIS and FAST for all three datasets mentioned.

Phase 3 looked at how the performance of the FUSE algorithm could be improved through the development of several versions of the algorithm, each tackling an issue that was identified. FUSE_2.0, involved the introduction of linguistic hedges to the FUSE algorithm and the expansion of the six initial categories with the introduction of three new categories, bringing the total number of fuzzy categories to nine. FUSE_2.0 was evaluated on five datasets and result correlation with human ratings was compared with four other SSM's. Results showed FUSE_2.0 gave a higher correlation with AHR compared with all four SSM's for all five datasets tested. FUSE_3.0, introduced negation operators to the FUSE algorithm that involved calculating the effect of negation operators on fuzzy words in utterances. FUSE_3.0 was fully evaluated when effects of negation on natural language utterances were entirely explored with the incorporation of the FUSE algorithm in the dialogue system (Research Question 2). FUSE_4.0, saw the introduction of a Fuzzy Influence factor to the FUSE algorithm, which allowed fuzzy words not in the same fuzzy category to still have a fuzzy measure associated with them. A set of empirical experiments was conducted using four datasets and result correlation with human ratings was compared with four other SSM's as well as earlier versions of FUSE (FUSE_2.0 and FUSE_3.0).

The FUSION dialogue system was developed in two versions. FUSION_V1 was developed using FUSE_2.0 and a set of questions were designed to represent each of the nine fuzzy categories of FUSE_2.0. Participants were recruited and asked to answer the questions by evaluating their visit to a local café where FUSION_V1 asked them questions relating to their experience. A dataset of participant results and a set of prototypical answers for each question was created. The dataset was used to evaluate FUSE_2.0 in the context of FUSION_V1 and results were compared with STASIS. Results showed that FUSION_V1 gave a higher True rating for eight categories compared to STASIS and one category had the same number of True hits with STASIS.

A second version of the dialogue system, referred to as FUSION_V2 was developed incorporating FUSE_4.0 and two sets of questions (Set 1 and Set 2), were designed to represent each of the nine fuzzy categories of FUSE_4.0. FUSION_V2 also utilised the negation operator and the fuzzy influence factor. Participants were recruited and asked to answer the questions by evaluating their working from home conditions where FUSION_V2 asked them questions relating to their experience. Two datasets of participant results and two sets of prototypical answers for each question was created (reflecting each set of questions). The datasets were used to evaluate FUSE_4.0 in the context of FUSION_V2 and results were compared with STASIS. Results showed that FUSION_V2 gave a higher True rating for all nine categories compared to STASIS for Set 1 and a higher True rating for all nine categories compared to STASIS for Set 2, with three of the categories achieving a 100% True rating for Set 2 questions.

8.3 - Research Question Evaluation

The research in this thesis was designed to answer two research questions. Each one will now be evaluated:

8.3.1 - Research Question 1

RQ1. Investigate the feasibility of utilising Type-2 Fuzzy Sets and their representation of an individual's perception of fuzzy words and evaluate the suitability of the resulting fuzzy word models for incorporation into a Fuzzy Semantic Similarity Measure (FSSM).

To answer the first research question, this research proposed a new algorithm called FUSE (FUzzy Similarity mEasure). FUSE has been developed following extensive research evaluation into existing state of the art sentence similarity measures and the only published fuzzy sentence similarity measure (FAST) (Chandran, 2013). The limitations and drawbacks of existing similarity measures were identified before establishing the proposed methodology and framework for the creation of the proposed FUSE algorithm. FUSE is an ontology-based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception-based words. This new approach is more suited to modelling *intra-personal* (the uncertainty a person has about the word) and *inter-personal* (the uncertainty that a group of people have about the word) uncertainties, which are intrinsic to natural language.

The FUSE algorithm was developed over four versions designed in three core phases investigating the presence of linguistic hedges, the expansion of fuzzy categories and their use in natural language, incorporation of logical operators such as '*not*' and the introduction of the fuzzy influence factor as descried in Chapter 6. The improvements made to the FUSE algorithm contributed towards modelling human perceptions. Results of experiments conducted on the different versions of FUSE showed that the inclusion of a fuzzy influence factor in a FSSM can improve the performance of the algorithm in terms of its correlation with human ratings.

8.3.2 - Research Question 2

RQ2. Can a Type-2 FSSM be embedded into a Q&A dialogue system with an improved success rate of utterance - response matches compared to traditional Semantic Similarity Measures (SSM)?

To answer the second research question, a simple Q&A dialogue system, referred to as FUSION was designed. Two versions of FUSION were developed using FUSE_2.0 (FUSION_V1) and FUSE_4.0 (FUSION_V2) respectively on two separate scenarios, using human participants to answer scenario-based questions. FUSION_V1 used an in-person approach, where human participants visited a local café and purchased a drink of their choice and observed their surroundings, before joining the researcher and answering the questions using FUSION_V1.

FUSION_V2 adapted an online approach due to the Covid-19 pandemic and social distancing implications, and participants answered two sets of questions (Set 1 and Set 2) relating to their working from home conditions.

Each scenario also asked participants to complete a Likert scale Usability Questionnaire to evaluate the performance and ease of using a Q&A Dialogue System. FUSION was run with both a traditional SSM (STASIS) (Li et al., 2003) as well as the FSSM FUSE, to test and compare performance. The results from the experiments conducted on the FUSION dialogue system had proven that incorporating a FSSM into a dialogue system can improve language understanding, due to matching a greater number of user responses to the prototypical sentences when compared with STASIS.

8.4 - Summary of Key Contributions

The key contributions made from this research are as follows:

- A new methodology for modelling fuzzy words was created which utilised Interval Type-2 fuzzy sets to represent human perception-based words. This work led to the creation of a fuzzy dictionary for six fuzzy categories which contained defuzzified numerical measures derived from average human ratings obtained using Interval Type-2 fuzzy set approach (Chapter 4). The fuzzy dictionary is a useful resource which can be used by other researchers in the field of NLP.
- Development of a fuzzy semantic similarity measure known as FUSE (FUzzy Similarity mEasure), with its first version (FUSE_1.0) using Interval Type-2 fuzzy sets and the inclusion of the newly developed fuzzy dictionary for six fuzzy categories using Interval Type-2 fuzzy sets (Chapter 4).
- Development of four versions of the FUSE algorithm which includes the incorporation of linguistic hedges and category expansion to nine fuzzy categories (FUSE_2.0). The inclusion of negation operators (FUSE_3.0) which permits a novel ability to apply fuzzy complement operators to fuzzy words modelled by Interval Type-2 Fuzzy Sets. Up to this point, fuzzy word similarity was only computed using the fuzzy category ontologies if fuzzy words belonged to the same fuzzy category. The introduction of a fuzzy influence factor (FUSE_4.0) allowed the fuzzy measure of a word to contribute to the overall similarity measure regardless of the fuzzy words in a pair of sentences belonging to the same fuzzy category or not (Chapter 6).
- The development of three new fuzzy categories resulting in an expansion of the fuzzy dictionary for nine fuzzy categories used for the FUSE_4.0 algorithm. This presents fuzzy words and their defuzzified numerical measure derived from average human ratings obtained using Interval Type-2 fuzzy set approach. The fuzzy dictionary of FUSE_4.0 can be used by other researchers in the field of NLP with other fuzzy applications such as semantic clustering (Appendix C).

- Comparisons of different versions of the FUSE algorithm with other state of the art Semantic Similarity Measures (SSM), across a number of published and newly created datasets (Chapter 5 and 6).
- Integration of FUSE_2.0 and FUSE_4.0 into two versions of a simple Q&A Dialogue System referred to as FUSION_V1 and FUSION_V2 respectively. Textual human responses were captured using two different scenarios (visit to a local café for FUSION_V1 and working from home for FUSION_V2). The integration of the FUSE algorithm into the FUSION dialogue system demonstrated that FSSM can be used in a real-world practical implementation, by incorporation into two different scenarios of a Q&A Dialogue System. Evaluation of the FUSION Dialogue Systems was achieved through comparison with traditional semantic similarity measures, and results indicated that a FSSM incorporated into a dialogue system is able to improve language understanding (Chapter 7).

8.5 - Future Work

The research presented in this thesis has outlined a novel approach to fuzzy semantic similarity measures, through the development of the FUSE algorithm. This research has also shown its successful incorporation into the FUSION dialogue system, which demonstrated that FSSM's can be used in a real-world practical implementation, by incorporation into two different scenarios of a Q&A Dialogue System. Whilst the research at this stage, fully meets the aim and objectives of this research, there are several areas for future research and development. Some of these suggestions are discussed in subsequent sections.

8.5.1 - Expansion of Fuzzy Words and Categories

Currently FUSE_4.0 holds a total of 386 fuzzy words across nine fuzzy categories. This is not an absolute number and future work can involve the expansion of the fuzzy words in each fuzzy category to cater for more fuzzy words in the English language.

Likewise, the fuzzy categories can also be expanded by introducing further fuzzy categories such as *Price* (expensive, cheap, bargain), *Health* (sick, unwell, ok), *Personality* (friendly, moody, positive) to name just a few. This would aid in further modelling of human perception-based words and would assist in applications, where domain knowledge is important such as smart home devices like the Nest thermostat by Google or Hive Active Heating by British Gas.

Further work can also be undertaken on the presence of linguistic hedges in fuzzy utterances using a larger participant sample (higher than 16), and additional testing of the hedge category on larger datasets and comparison with other SSM's to compare correlations with AHR.

8.5.2 - Revisit Negation and Fuzzy Influence Factor

Overall, it would be beneficial to conduct further experimentation on logical negation operators and the evaluation of them using FUSE. At present FUSE addresses '*not*' when it is immediately present before a fuzzy word *(*i.e., *not bright).* Further work could involve catering for '*not*' and similar negation values anywhere in the sentence and not just directly before a fuzzy word *(*i.e., *I do <u>not</u> want a large drink).*

Additional experiments could be conducted on the fuzzy influence factor on larger datasets, with results compared with other SSM's. Datasets used by the NLP community often do not have sufficient fuzzy words to allow for rigorous testing of an FSSM. Therefore, one of the challenges is specific datasets that may need to be created or curated from existing ones.

8.5.3 - Inclusion of Conjunctions and Numbers

Conjunctions in the English language refers to words that link phrases or sentences (i.e., *for*, *and*, *nor*, *but*, *or*, *yet*). Further work could involve modelling these words in a fuzzy utterance (i.e., I am somewhere between a *large <u>and</u>* a *medium*). In this example, two fuzzy words *large* and *medium* are used, but the conjunction of *<u>and</u>* has also been used. Future work would need to assess this sentence and rather than simply apply a measure for *large* and for *medium*, consider the presence of *<u>and</u>*, and apply a measure accordingly.

At present FUSE does not cater for the presence of numbers in utterances, which can sometimes affect the similarity, as seen in the FUSION scenario experiments, where participants sometimes gave a number instead of a description (*the barrister looked like in* *her 30's*). Future work could involve the development of a methodology to address the presence of numbers in fuzzy utterances.

8.5.4 - Applications of FUSE and FUSION

In future applications of FUSE and FUSION, the ability to add speech-to-text recognition would allow a more fluent dialogue with a user, combined with the ability to recognise fuzzy words in a context, and provide appropriate user tailored interventions. An example could be two people living in one home, and both saying *Alexa, I am cold*. Person 1 may have a different interpretation of cold than Person 2. Smart devices such as Alexa, or Google Home, already have speech recognition, thus incorporation of fuzzy natural language understanding in a given context, in this case a smart home environment, would allow the heating to be adjusted based on each person's perception of temperature.

Common voice recognition systems function when there are no dysphonic (abnormal functioning of the voice) present in the voice but are poor at accurately transcribing dysphonic voices. One of the key challenges of fuzzy words is regional dialect, accents, and people with voice disorders. The FUSE dictionary could be re-evaluated to incorporate these elements by modelling the words for each of these aspects to allow the fuzzy dictionary to be changed depending on who the end user may be.

Some applications that use human utterances are Li et. al. (Li et. al., 2021) that implements interaction between the robotic system and the human operator. Tokunaga et. al. (Tokunaga et. al., 2021) uses a dialogue system to aid with dementia in older adults suffering from symptoms to maintain daily life. Clemente et. al. (Clemente et. al., 2022) uses conversational agents to assist with healthcare and wellbeing. However none of the applications mentioned incorporate fuzzy words as they only rely on key word extraction. These applications could be adapted to incorporate the fuzzy dictionary and use the FUSE algorithm to improve response rates and accuracy.

8.5.5 - Language Adaption of FUSE and FUSION

Adapting the FUSE algorithm and the FUSION Dialogue System to other languages especially low resource language such as Arabic, Urdu and Farsi, through investigating lexical resources and designing fuzzy dictionaries. Current work such as UMAIR (Urdu Machine for Artificially Intelligent Recourse) (Kaleem, 2015), a text-based goal-orientated conversational agent (CA) for the Urdu language or LANA, an Arabic Conversational Intelligent Tutoring System (CITS) (Aljameel, 2018), specifically designed to aid children with Autism Spectrum Disorder (ASD) use traditional SSM's that have been directly translated into Urdu and Arabic respectively. The development and integration of a FSSM into such applications will allow a richer modelling of human perception-based words.

References

Achananuparp, P., Hu, X. and Shen, X. (2008) ‘The evaluation of sentence similarity measures.’ *In International Conference on data warehousing and knowledge discovery*, pp. 305-316. Springer, Berlin, Heidelberg.

Adel, N., Crockett, K., Carvalho, J. P. and Cross, V. (2021) ‘Fuzzy Influence in Fuzzy Semantic Similarity Measures.’ *In 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-7. IEEE.

Adel, N., Crockett, K., Chandran, D. and Carvalho, J. P. (2020) ‘Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures.’ *In 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-8. IEEE.

Adel, N., Crockett, K., Crispin, A., Carvalho, J. P. and Chandran, D. (2019) ‘Human Hedge Perception–and its Application in Fuzzy Semantic Similarity Measures.’ *In 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-7. IEEE.

Adel, N., Crockett, K., Crispin, A., Chandran, D. and Carvalho, J. P. (2018) ‘FUSE (Fuzzy Similarity Measure)-A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets.’ *In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-8. IEEE.

Alexa and Alexa Device FAQs - Amazon Customer Service. (2016) *Alexa and Alexa Device FAQs - Amazon Customer Service*. Code Notice. [Online] [Accessed on 3 Jan 2020] https://www.amazon.com/gp/help/customer/display.html?tag=skim%201x169757-20&nodeId=201602230

Alian, M. and Awajan, A. (2020) 'Factors affecting sentence similarity and paraphrasing identification.' *International Journal of Speech* Technology, 23(4), pp. 851-859.

Aljameel, S. S. (2018) *Development of an Arabic conversational intelligent tutoring system for education of children with autism spectrum disorder.* Ph.D. Manchester Metropolitan University.

Aljameel, S. S., O'Shea, J. D., Crockett, K. A., Latham, A. and Kaleem, M. (2017) ‘Development of an Arabic conversational intelligent tutoring system for education of children with ASD.’ *In 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 24-29. IEEE.

Aljameel, S., O’Shea, J., Crockett, K., Latham, A. and Kaleem, M. (2019) ‘LANA-I: an Arabic conversational intelligent tutoring system for children with ASD.’ In Intelligent Computing-Proceedings of the Computing Conference, pp. 498-516. Springer, Cham.

Alnajran, N. A. (2019) *An integrated semantic-based framework for intelligent similarity measurement and clustering* of microblogging posts. Ph.D. Manchester Metropolitan University.

Atkinson, J., Ferreira, A. and Aravena, E. (2009) 'Discovering implicit intention-level knowledge from natural-language texts.' In International Conference on Innovative Techniques and Applications of Artificial Intelligence, pp. 249-262. Springer, London.

Aujogue, J.-B. and Aussem, A. (2019) ‘Hierarchical Recurrent Attention Networks for Context-Aware Education Chatbots.’ *In 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE.

Automatic Readability Checker. (2022) *Automatic Readability Checker*. [Online] [Accessed on 3 Jan 2018] https://readabilityformulas.com/free-readability-formula-tests.php

Balasubramanian, N., Allan, J. and Croft, W. B. (2007) ‘A comparison of sentence retrieval techniques.’ *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 813-814.

Banks, W. (2003) *Linguistic Variables: Clear Thinking with Fuzzy Logic*. [Online] [Accessed on 30 May 2016] http://www.phaedsys.co.uk/principals/bytecraft/bytecraftdata/LinguisticVariables.pdf

Batet, M. and Sánchez, D. (2015) 'A review on semantic similarity.' *In Encyclopedia of Information Science and Technology, Third Edition*. pp. 7575-7583. IGI Global.

Bengio, Y., Ducharme, R. and Vincent, P. (2000) 'A neural probabilistic language model.' *Advances in Neural Information Processing Systems*, 13

Bhuvan, N. T. and Elayidom, M. S. (2020) 'A supervised multimodal search re-ranking technique using visual semantics.' *International Journal of Intelligent Enterprise*, 7(1-3), pp. 279-290.

Bilgin, A., Hagras, H., Malibari, A., Alhaddad, M. J. and Alghazzawi, D. (2012) ‘Towards a general type-2 fuzzy logic approach for computing with words using linear adjectives.’ *In 2012 IEEE international conference on fuzzy systems* (pp. 1-8). IEEE.

Bird, S., Klein, E. and Loper, E. (2009) *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Bouziane, A., Bouchiha, D., Doumi, N. and Malki, M. (2015) 'Question answering systems: survey and trends.' *Procedia Computer Science*, 73, pp. 366-375.

Brownlee, J. (2017) *What are word embeddings for text?.* 11th October. Machine Learning Mastery. [Online] [Accessed on 11th February 2021] https://machinelearningmastery. com/what-are-wordembeddings

Car, L. T., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y.-L. and Atun, R. (2020) 'Conversational agents in health care: scoping review and conceptual analysis.' *Journal of medical Internet research*, 22(8), p. e17158.

Castillo, O. and Melin, P. (2012) *Recent Advances in Interval Type-2 Fuzzy Systems* Springer-Verlag Berlin Heidelberg.

Chandran, D. (2013) *The development of a fuzzy semantic sentence similarity measure.* Ph.D. Manchester Metropolitan University.

Chandran, D., Crockett, K., Mclean, D. and Bandar, Z. (2013) ‘FAST: A fuzzy semantic sentence similarity measure.’ *In 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-8. IEEE.

Chandrasekaran, D. and Mago, V. (2021) 'Evolution of semantic similarity—A survey.' *ACM Computing Surveys (CSUR)*, 54(2), pp. 1-37.

Cicchetti, D. V. (1994) 'Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.' *Psychological assessment*, 6(4), p. 284.

Clemente, C., Grecob, E., Sciarrettac, E. and Altieri, L. (2022) 'Alexa, How Do I Feel Today? Smart Speakers for Healthcare and Wellbeing: an Analysis About Uses and Challenges.' *Sociology and Social Work Review*, 6(1), pp. 6-24.

Colby, K. M. (1981) 'Modeling a paranoid mind.' *Behavioral and Brain Sciences*, 4(4), pp. 515-534.

Cox, E. (1994) *The fuzzy systems handbook: a practitioner's guide to building, using, and maintaining fuzzy systems.* Academic Press Professional, Inc.

Crockett, K., Adel, N., O'Shea, J., Crispin, A., Chandran, D. and Carvalho, J. P. (2017) 'Application of fuzzy semantic similarity measures to event detection within tweets.' *In 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* pp. 1-7. IEEE.

Curry, C. (2018) *A framework for developing a conversational agent to improve normal age-associated memory loss and increase subjective wellbeing.* Ph.D. Manchester Metropolitan University.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990) 'Indexing by latent semantic analysis.' *Journal of the American society for information science*, 41(6), pp.391-407.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E. and Cieliebak, M. (2021) 'Survey on evaluation methods for dialogue systems.' *Artificial Intelligence Review*, 54(1), pp. 755-810.

Dumais, S. T. (2004) 'Latent semantic analysis.' *Annual review of information science and technology*, 38(1), pp. 188-230.

Feldman, S. (1999) 'NLP meets the Jabberwocky: Natural language processing in information retrieval.' *ONLINE-WESTON THEN WILTON*, 23, pp. 62-73.

Fenton, N. and Neil, M. (2018) *Risk assessment and decision analysis with Bayesian networks.* Crc Press.

Ferrández, O., Izquierdo, R., Ferrández, S. and Vicedo, J. L. (2009) 'Addressing ontology-based question answering with collections of user queries.' *Information Processing & Management*, 45(2), pp. 175-188.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2001) 'Placing search in context: The concept revisited.' *In Proceedings of the 10th international conference on World Wide Web*, pp. 406-414.

Fitzpatrick, K. K., Darcy, A. and Vierhile, M. (2017) 'Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial.' *JMIR mental health*, 4(2), p. e7785.

Foltz, P. W. (1996) 'Latent semantic analysis for text-based research.' *Behavior Research Methods, Instruments, & Computers*, 28(2), pp. 197-202.

Francis, W. N. and Kucera, H. (1979) 'Brown corpus manual.' *Letters to the Editor*, 5(2), p. 7.

Fullér, R. (2010) *What is fuzzy logic and fuzzy ontology?* [Online] [Accessed on 16 Mar 2022] https://uni-obuda.hu/users/fuller.robert/otaniemi-2.pdf

Galindo, J. (2008) 'Introduction and trends to fuzzy logic and fuzzy databases.' *In Handbook of Research on Fuzzy Information Processing in Databases*, pp. 1-33. IGI Global.

Gomes, P., Seco, N., Pereira, F. C., Paiva, P., Carreiro, P., Ferreira, J. L. and Bento, C. (2006) 'The importance of retrieval in creative design analogies.' *Knowledge-Based Systems*, 19(7), pp. 480-488.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, T. R. G. and Person, N. (2000) 'Using latent semantic analysis to evaluate the contributions of students in AutoTutor.' *Interactive learning environments*, 8(2), pp. 129-147.

Green, S. J. (1999) 'Building hypertext links by computing semantic similarity.' *IEEE Transactions on Knowledge and Data Engineering*, 11(5), pp. 713-730.

Grefenstette, G. (1992) 'Use of syntactic context to produce term association lists for text retrieval.' *In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 89-97.

Gupta, P. K. and Muhuri, P. K. (2018) 'A novel approach based on computing with words for monitoring the heart failure patients.' *Applied Soft Computing*, 72, pp. 457-473.

Gupta, P. K. and Muhuri, P. K. (2019) 'Computing with words for student strategy evaluation in an examination.' *Granular Computing*, 4(2), pp. 167-184.

Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W. and Gao, C. (2021) 'A survey on the techniques, applications, and performance of short *text semantic similarity.' Concurrency and Computation: Practice and Experience, 33(5), p. e5971.*

Hao, M. and Mendel, J. (2015) 'Encoding words into normal interval type-2 fuzzy sets: HM approach.' IEEE Transactions on Fuzzy Systems, 24(4), pp.865-879.

Harms, J.-G., Kucherbaev, P., Bozzon, A. and Houben, G.-J. (2018) 'Approaches for dialog management in conversational agents.' *IEEE Internet Computing*, 23(2), pp. 13-22.

Jiang, J. J. and Conrath, D. W. (1997) 'Semantic similarity based on corpus statistics and lexical taxonomy.' *arXiv preprint cmp-lg/9709008*.

John, R. and Coupland, S. (2006) 'Extensions to type-1 fuzzy logic: Type-2 fuzzy logic and uncertainty.' *Computational Intelligence: Principles and Practice*, pp. 89-102.

Kaleem, M. (2015) *Methodology and algorithms for Urdu language processing in a conversational agent.* Ph.D. Manchester Metropolitan University.

Kaleem, M., O'Shea, J. D. and Crockett, K. A. (2014) 'Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents.' *In 2014 14th UK Workshop on Computational Intelligence (UKCI)*, pp. 1-8. IEEE.

Karnik, N. N. and Mendel, J. M. (2001) 'Centroid of a type-2 fuzzy set.' *information SCiences*, 132(1-4), pp. 195-220.

Kent State University. (2013) *SPSS Tutorials: Pearson Correlation.* [Online] [Accessed on 09 Dec 2017] https://libguides.library.kent.edu/SPSS/PearsonCorr

Klir, G. J. and Folger, T. A. (1988) *Fuzzy Sets, Uncertainty, and Information.* New Jersey: Prentice Hall.

Koetter, F., Blohm, M., Drawehn, J., Kochanowski, M., Goetzer, J., Graziotin, D. and Wagner, S. (2019) *Conversational agents for insurance companies: from theory to practice.* Springer.

Koo, T. K. and Li, M. Y. (2016) 'A guideline of selecting and reporting intraclass correlation coefficients for reliability research.' *Journal of chiropractic medicine*, 15(2), pp. 155-163.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998) 'An introduction to latent semantic analysis.' *Discourse processes*, 25(2-3), pp. 259-284.

Lara-Clares, A., Lastra-Díaz, J. J. and Garcia-Serrano, A. (2021) 'Protocol for a reproducible experimental survey on biomedical sentence similarity.' *Plos one*, 16(3), p. e0248663.

Latham, A., Crockett, K. and McLean, D. (2014) 'An adaptation algorithm for an intelligent natural language tutoring system.' *Computers & Education*, 71, pp. 97-110.

Lesk, M. (1986) 'Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.' *In Proceedings of the 5th annual international conference on Systems documentation*, pp. 24-26.

Li, C., Park, J., Kim, H. and Chrysostomou, D. (2021) 'How can i help you? an intelligent virtual assistant for industrial robots.' *In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 220-224.

Li, Y., Bandar, Z. A. and McLean, D. (2003) 'An approach for measuring semantic similarity between words using multiple information sources.' *IEEE Transactions on knowledge and data engineering*, 15(4), pp. 871-882.

Li, Y., Bandar, Z., McLean, D. and O'Shea, J. (2004) 'A Method for Measuring Sentence Similarity and iIts Application to Conversational Agents.' *In FLAIRS Conference*, pp. 820-825.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D. and Crockett, K. (2006) 'Sentence similarity based on semantic nets and corpus statistics.' *IEEE transactions on knowledge and data engineering*, 18(8), pp. 1138-1150.

Liddy, E. D. (2001) 'Natural language processing.' *In Encyclopedia of Library and Information Science*, 2nd Ed. NY. Marcel Decker, Inc.

Lin, C.Y. and Hovy, E. (2003) 'Automatic evaluation of summaries using n-gram co-occurrence statistics.' *In Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pp. 150-157.

Lin, D. (1998) 'An Information-Theoretic Definition of Similarity.' *In Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 296-304.

Lin, L., Ginns, P., Wang, T. and Zhang, P. (2020) 'Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain?' *Computers & Education*, 143, p. 103658.

Little, C., Mclean, D., Crockett, K. and Edmonds, B. (2020) 'A semantic and syntactic similarity measure for political tweets.' *IEEE Access*, 8, pp. 154095-154113.

Liu, F. and Mendel, J. M. (2008) 'Encoding words into interval type-2 fuzzy sets using an interval approach.' *Fuzzy Systems, IEEE Transactions on*, 16(6), pp. 1503-1521.

Liu, H. and Wang, P. (2013) 'Assessing Sentence Similarity Using WordNet based Word Similarity.' *J. Softw.*, 8(6), pp. 1451-1458.

Loper, E. and Bird, S. (2002) 'Nltk: The natural language toolkit.' *arXiv preprint cs/0205028*

Lopez, V., Uren, V., Motta, E. and Pasin, M. (2007) 'AquaLog: An ontology-driven question answering system for organizational semantic intranets.' *Journal of Web Semantics*, 5(2), pp. 72-105.

Lord, P., Stevens, R., Brass, A. and Goble, C. (2003) 'Semantic similarity measures as tools for exploring the gene ontology.' *In Proceedings of Pacific Symposium on Biocomputing*, pp. 601-612.

Luchoomun, T., Chumroo, M. and Ramnarain-Seetohul, V. (2019) 'A knowledge based system for automated assessment of short structured questions.' *In 2019 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1349-1352. IEEE.

Lytvyn, V., Vysotska, V., Rusyn, B., Pohreliuk, L., Berezin, P. and Naum, O. (2019) 'Textual Content Categorizing Technology Development Based on Ontology.' *In MoMLeT*, pp. 234-254.

Ma, Y., Nguyen, K. L., Xing, F. Z. and Cambria, E. (2020) 'A survey on empathetic dialogue systems.' *Information Fusion*, 64, pp. 50-70.

McHugh, M. L. (2012) 'Interrater reliability: the kappa statistic.' *Biochemia medica*, 22(3), pp. 276-282.

McTear, M. (2020) 'Conversational AI: dialogue systems, conversational agents, and chatbots.' *Synthesis Lectures on Human Language Technologies*, 13(3), pp. 1-251.

Mendel, J. and Wu, D. (2010) *Perceptual computing: aiding people in making subjective judgments.* Vol. 13. John Wiley & Sons.

Mendel, J.M. (2007) 'Type-2 fuzzy sets and systems: an overview.' *Computational Intelligence Magazine, IEEE*, 2(1), pp. 20-29.

Mendel, J.M. (2017) *Uncertain Rule-Based Fuzzy Systems Introduction and New Directions*. 2nd ed., Springer.

Mendel, J.M. and John, R.I.B. (2002) 'Type-2 fuzzy sets made simple.' *Fuzzy Systems, IEEE Transactions on*, 10(2), pp. 117-127.

Mendel, J.M., John, R.I. and Liu, F. (2006) 'Interval type-2 fuzzy logic systems made simple.' *Fuzzy Systems, IEEE Transactions on*, 14(6), pp. 808-821.

Mendel, J.M., Zadeh, L.A., Trillas, E., Yager, R., Lawry, J., Hagras, H. and Guadarrama, S. (2010) 'What computing with words means to me.' *IEEE Computational Intelligence Magazine*, 5(1), pp. 20-26.

Meng, L., Huang, R. and Gu, J. (2014) 'Measuring semantic similarity of word pairs using path and information content.' *Int. J. Futur. Gener. Commun. & Netw*, 7, pp. 183-194.

Mesiarová-Zemánková, A. and Ahmad, K. (2010) 'T-norms in subtractive clustering and backpropagation' *International Journal of Intelligent Systems*, 25(9), pp.909-924.

Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A. and Zobel, J. (2005) 'Similarity measures for tracking information flow.' *In Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 517-524.

Metzler, D., Dumais, S. and Meek, C. (2007) 'Similarity measures for short segments of text.' *In European conference on information retrieval*, pp. 16-27. Springer, Berlin, Heidelberg.

Michie, D. (2001) 'Return of the imitation game.' *Electronic Transactions in Artificial Intelligence*.

Michie, D. and Sammut, C. (2001) 'Infochat scripter's manual.' *Technical Repoort, Convagent Ltd, Manchester, UK*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013) 'Distributed representations of words and phrases and their compositionality.' *Advances in neural information processing systems*, 26.

Miller, G. A. (1995) 'WordNet: a lexical database for English.' *Communications of the ACM*, 38(11), pp. 39-41.

Mohammed, S. M., Jacksi, K. and Zeebaree, S. (2021) 'A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms.' *Indonesian Journal of Electrical Engineering and Computer Science*, 22(1), pp. 552-562.

Montenegro, J. L. Z., da Costa, C. A. and da Rosa Righi, R. (2019) 'Survey of conversational agents in health.' *Expert Systems with Applications*, 129, pp. 56-67.

Nadkarni, P. M., Ohno-Machado, L. and Chapman, W. W. (2011) 'Natural language processing: an introduction.' *Journal of the American Medical Informatics Association*, 18(5), pp. 544-551.

Natural Language AI. (2020) *Natural Language AI*. [Online] [Accessed on 3 Jan 2020] https://cloud.google.com/natural-language

Navigli, R. (2009) 'Word sense disambiguation: A survey.' *ACM Computing Surveys (CSUR)*, 41(2), p. 10.

Negnevitsky, M. (2005) *Artificial Intelligence: A Guide to Intelligent Systems*. 2 ed., p. 435. England: Pearson Education Limited.

Nemoto, T. and Beglar, D. (2014) 'Likert-scale questionnaires.' *In JALT 2013 conference proceedings*, pp. 1-8.

Ng, H. T. (1997) 'Getting serious about word sense disambiguation.' *In Tagging Text with Lexical Semantics: Why, What, and How?*.

Noraset, T., Lowphansirikul, L. and Tuarob, S. (2021) 'Wabiqa: A wikipedia-based thai question-answering system.' *Information processing & management*, 58(1), p. 102431.

O’Shea, J., Bandar, Z. and Crockett, K. (2011) 'Systems engineering and conversational agents.' *In Intelligence-Based Systems Engineering*, pp. 201-232. Springer.

O’Shea, J., Bandar, Z. and Crockett, K. (2013) 'A new benchmark dataset with production methodology for short text semantic similarity algorithms.' *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4), p. 19.

O’Shea, K. (2012) 'An approach to conversational agent design using semantic sentence similarity.' *Applied Intelligence*, 37(4), pp. 558-568.

O'Dell, J. W. and Dickson, J. (1984) 'Eliza as a “therapeutic” tool.' *Journal of Clinical Psychology*, 40(4), pp. 942-945.

O'Shea, J. (2010) *A framework for applying short text semantic similarity in goal-oriented conversational agents.* Ph.D. Manchester Metropolitan University.

O'Shea, K., Bandar, Z. and Crockett, K. (2009) 'Towards a new generation of conversational agents based on sentence similarity.' *In Advances in electrical engineering and computational science*. Springer, pp. 505-514.

Oxford English Dictionary. (2021) [Online]. [Accessed on 25 Apr 2017] https://www.oed.com/

Ozaeta, L. and Graña, M. (2018) *A View of the State of the Art of Dialogue Systems.* Springer.

Patwardhan, S., Banerjee, S. and Pedersen, T. (2003) ‘Using measures of semantic relatedness for word sense disambiguation.’ *In International conference on intelligent text processing and computational linguistics*, pp. 241-257. Springer, Berlin, Heidelberg.

Pazos R, R. A., González B, J. J., Aguirre L, M. A., Martínez F, J. A. and Fraire H, H. J. (2013) 'Natural language interfaces to databases: an analysis of the state of the art.' *In Recent Advances on Hybrid Intelligent Systems*, pp. 463-480. Springer.

Pedersen, T., Patwardhan, S. and Michelizzi, J. (2004) ‘WordNet:: Similarity-Measuring the Relatedness of Concepts.’ *In AAAI*, 4, pp. 25-29.

Phan, H. T., Tran, V. C., Nguyen, N. T. and Hwang, D. (2020) 'Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model.' *IEEE Access*, 8, pp. 14630-14641.

Po, D. K. (2020) 'Similarity Based Information Retrieval Using Levenshtein Distance Algorithm.' *Int. J. Adv. Sci. Res. Eng*, 6(04), pp. 06-10.

Quesada, J. (2007) 'Creating your own LSA spaces.' *In Handbook of latent semantic analysis*, pp. 83-98. Psychology Press.

Rahmanian, M., Shafieian, M. and Samie, M. E. (2021) 'Computing with words for student peer assessment in oral presentation.' *Nexo Revista Científica*, 34(01) pp. 229-241.

Reja, U., Manfreda, K. L., Hlebec, V. and Vehovar, V. (2003) 'Open-ended vs. close-ended questions in web questionnaires.' *Developments in applied statistics*, 19(1), pp. 159-177.

Resnik, P. (1995) 'Using information content to evaluate semantic similarity in a taxonomy.' *arXiv preprint cmp-lg/9511007*.

Resnik, P. (1999) 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.' *J. Artif. Intell. Res.(JAIR)*, 11, pp. 95-130.

Rocchio, J. (1971) 'Relevance feedback in information retrieval.' *The Smart retrieval system-experiments in automatic document processing*, pp. 313-323.

Ross, S. M. (2004) *Introduction to probability and statistics for engineers and scientists.* Elsevier.

Rubenstein, H. and Goodenough, J. B. (1965) 'Contextual correlates of synonymy.' *Communications of the ACM*, 8(10), pp. 627-633.

Rus, V., Banjade, R., Lintean, M., Niraula, N. and Stefanescu, D. (2013b) 'SEMILAR: A Semantic Similarity Toolkit for Assessing Students' Natural Language Inputs.' *In Educational data mining 2013.*

Rus, V., Lintean, M., Banjade, R., Niraula, N. B. and Stefanescu, D. (2013a) 'Semilar: The semantic similarity toolkit.' *In Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations*, pp. 163-168.

Rus, V., Lintean, M., Moldovan, C., Baggett, W., Niraula, N. and Morgan, B. (2012) 'The similar corpus: A resource to foster the qualitative understanding of semantic similarity of texts.' *In Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012)*, pp. 23-25.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval.' *Information processing & management*, 24(5), pp. 513-523.

Sammut, C. (2001) 'Managing context in a conversational agent.' *Linkoping Electronic Articles in Computer & Information Science*, 3(7).

Sánchez, D., Isern, D. and Millan, M. (2011) 'Content annotation for the semantic web: an automatic web-based approach.' *Knowledge and Information Systems*, 27(3), pp. 393-418.

Sebti, A. and Barfroush, A. A. (2008) 'A new word sense similarity measure in WordNet.' *In 2008 International Multiconference on Computer Science and Information Technology*, pp. 369-373. IEEE.

Singh, P. K. and Paul, S. (2021) 'Deep Learning Approach for Negation Handling in Sentiment Analysis.' *IEEE Access*, 9, pp. 102579-102592.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000) 'Content-based image retrieval at the end of the early years.' *IEEE Transactions on pattern analysis and machine intelligence*, 22(12), pp. 1349-1380.

SpazioDati. (2015) 'Dandelion API.' [Online] [Accessed on 23 Mar 2019] https://dandelion.eu/

Spiccia, C., Augello, A., Pilato, G. and Vassallo, G. (2016) 'Semantic word error rate for sentence similarity.' *In 2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pp. 266-269. IEEE.

Srihari, R. K., Zhang, Z. and Rao, A. (2000) 'Intelligent indexing and semantic retrieval of multimodal documents.' *Information Retrieval*, 2(2-3), pp. 245-275.

Srivastava, P. and Mondal, R. (2022) 'Design and Development of Intelligent Information System Using Hesitant Fuzzy Weighting Linguistic Term Sets for Computing with Words.' *In Mathematical, Computational Intelligence and Engineering Approaches for Tourism, Agriculture and Healthcare*, pp. 195-207, Springer.

Su, Z., Hu, D. and Yu, X. (2019) 'General interval approach for encoding words into interval type-2 fuzzy sets based on normal distribution and free parameter.' *Soft Computing*, 23(17), pp. 8187-8206.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B. (2014) 'Learning sentiment-specific word embedding for twitter sentiment classification.' *In ACL*, (1), (pp. 1555-1565).

Tapeh, A. G. and Rahgozar, M. (2008) 'A knowledge-based question answering system for B2C eCommerce.' *Knowledge-Based Systems*, 21(8), pp. 946-950.

Tešić, J., Tamir, D., Neumann, S., Rishe, N. and Kandel, A. (2020) 'Computing with words in maritime piracy and attack detection systems.' *In International Conference on Human-Computer Interaction,* pp. 434-444. Springer, Cham.

Text Inspector (2020) 'What is Readability and How Does It Work?' [Online] [Accessed on 16 Nov 2020] https://textinspector.com/what-is-readability-and-how-does-it-work/

Tokunaga, S., Tamura, K. and Otake-Matsuura, M. (2021) 'Implementation and Evaluation of Home-based Dialogue System for Cognitive Training of Older Adults.' *In The 23rd International Conference on Information Integration and Web Intelligence*, pp. 458-461.

Vysotska, V., Lytvyn, V., Burov, Y., Berezin, P., Emmerich, M. and Fernandes, V. B. (2019) 'Development of Information System for Textual Content Categorizing Based on Ontology.' *In COLINS*, pp. 53-70.

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (1993) *Probability and statistics for engineers and scientists.* Vol. 5. Macmillan New York.

Wang, B., He, W., Yang, Z. and Xiong, S. (2020) 'An unsupervised sentiment classification method based on multi-level fuzzy computing and multi-criteria fusion.' *IEEE Access*, 8, pp. 145422-145434.

Wang, H., Xu, Z. and Zeng, X.-J. (2018) 'Linguistic terms with weakened hedges: A model for qualitative decision making under uncertainty.' *Information Sciences*, 433, pp. 37-54.

Wang, J. and Dong, Y. (2020) 'Measurement of text similarity: a survey.' *Information*, 11(9), p. 421.

Winograd, T. (1973) 'A procedural model of language understanding.'

Woods, W. A. (1973) 'Progress in natural language understanding: an application to lunar geology.' *In Proceedings of the June 4-8, 1973, national computer conference and exposition* ,pp. 441-450.

Wu, D., Mendel, J. M. and Coupland, S. (2011) 'Enhanced interval approach for encoding words into interval type-2 fuzzy sets and its convergence analysis.' *IEEE Transactions on Fuzzy Systems*, 20(3), pp. 499-513.

Wu, Z. and Palmer, M. (1994) ‘Verbs Semantics and Lexical Selection.’ *In Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics*, Las Cruces, New Mexico.

Wubben, S. (2008) 'Using free link structure to calculate semantic relatedness.' *ILK Research Group Technical Report Series*, (08-01).

Yang, A. (2021) 'Fuzzy Language Processing in Translation of Literature: A Case Study of Dr. Zha’s Diary of Fighting the COVID-19.' *Theory and Practice in Language Studies*, 11(6), pp. 742-748.

Yi, M. (2019) 'A Complete Guide to Scatter Plots.' [Online] [Accessed on 5 Jul 2020] https://chartio.com/learn/charts/what-is-a-scatter-plot/

Zadeh, L. A. (1975a) 'The concept of a linguistic variable and its application to approximate reasoning—I.' *Information sciences*, 8(3), pp. 199-249.

Zadeh, L. A. (1975b) 'The concept of a linguistic variable and its application to approximate reasoning—II.' *Information sciences*, 8(4), pp. 301-357.

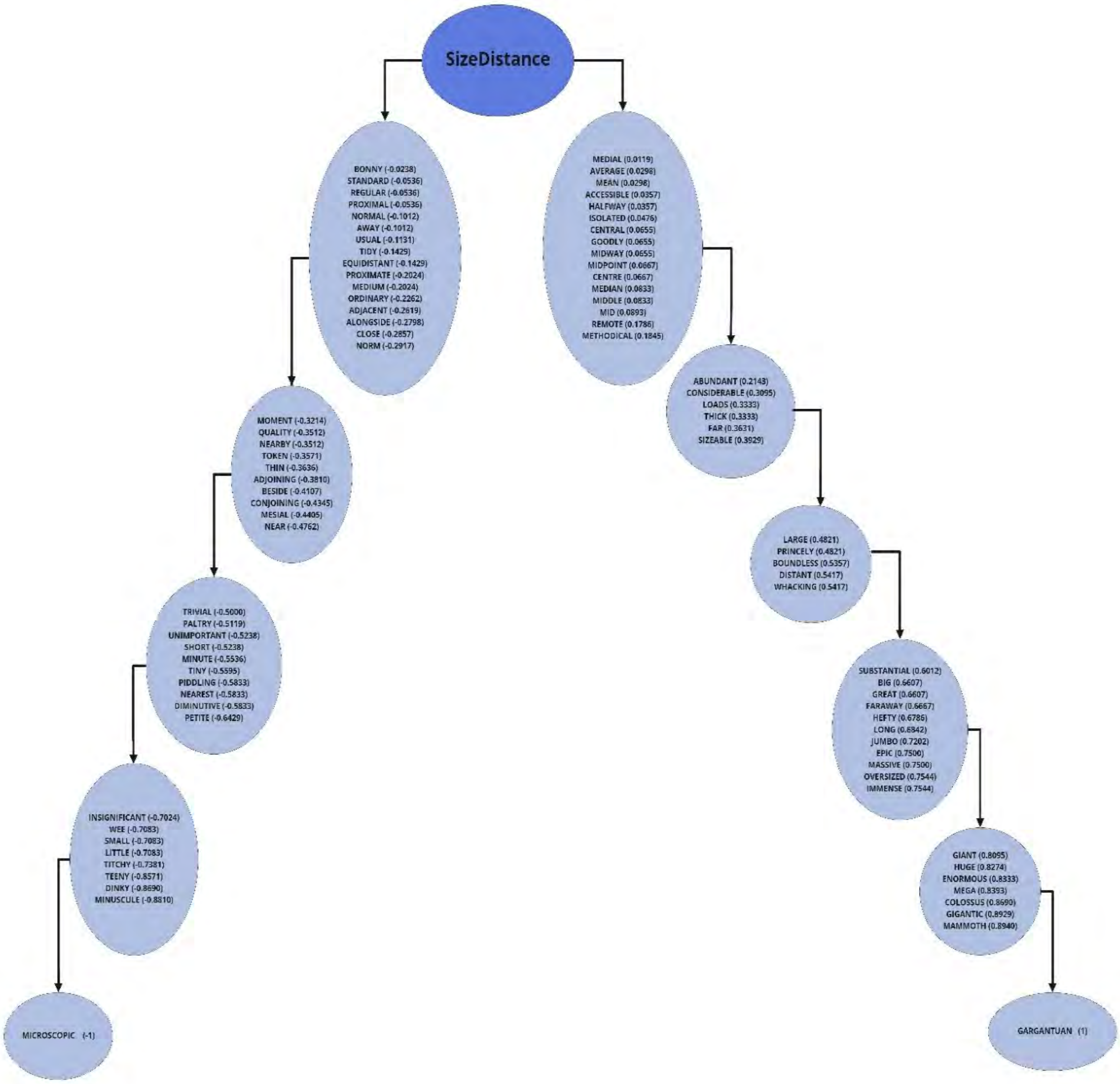
Zadeh, L. A. (1996) 'Fuzzy logic= computing with words.' *Fuzzy Systems, IEEE Transactions on*, 4(2), pp. 103-111.

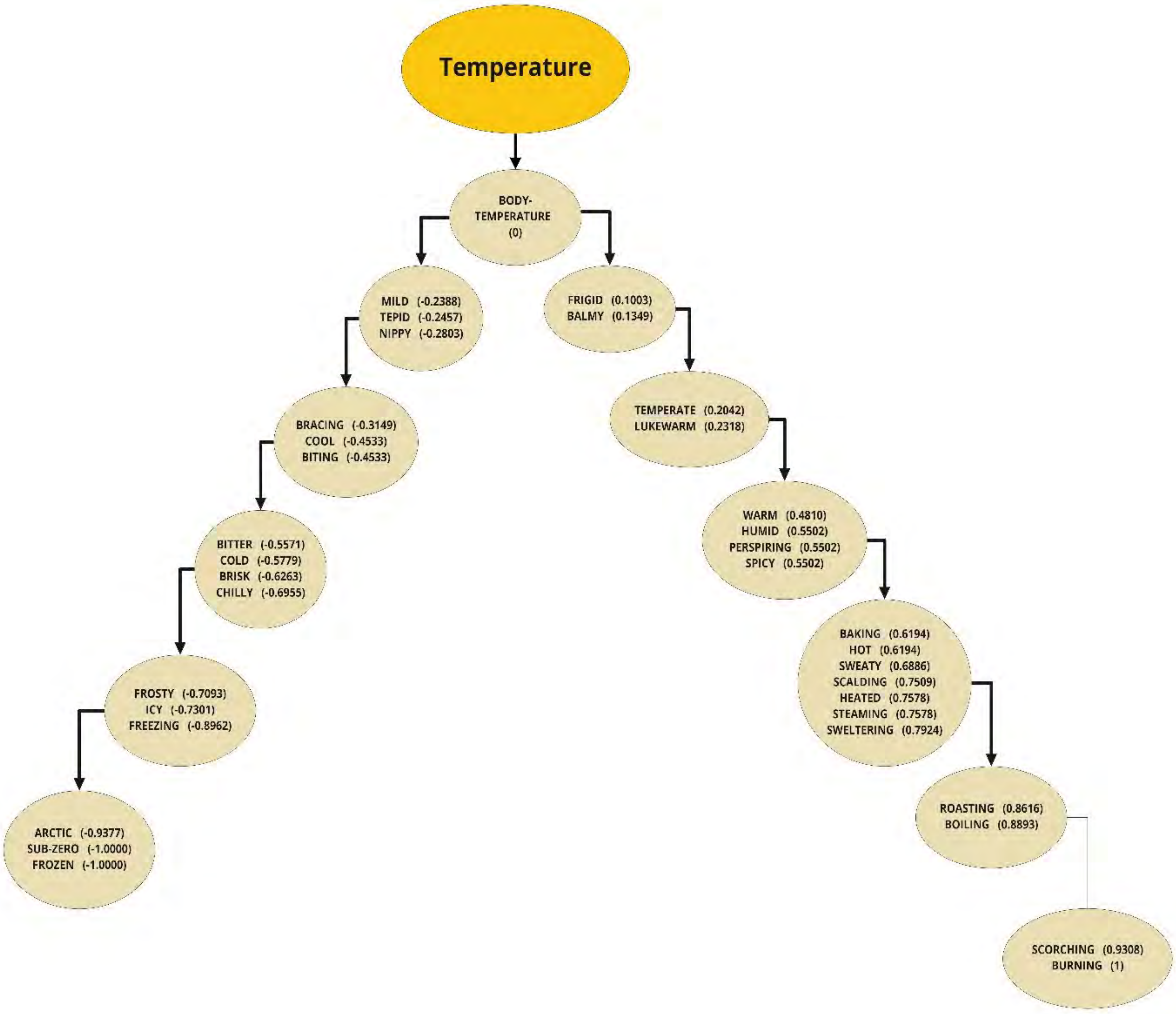
Zadeh, L. A. (1999) 'From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions.' *IEEE Transactions on circuits and systems I: fundamental theory and applications*, 46(1), pp. 105-119.

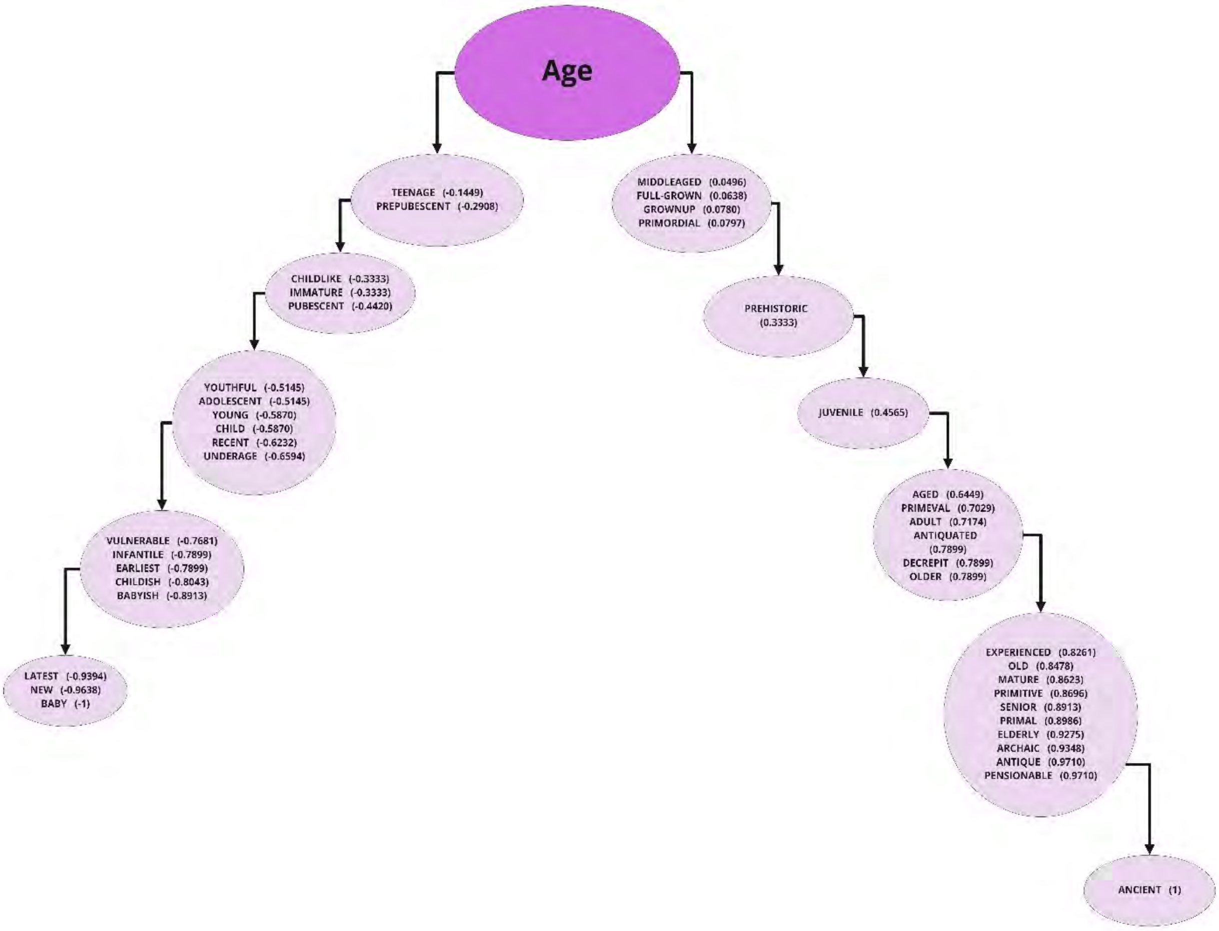
Zhang, D., Xia, X., Yang, Y., Yang, P., Xie, C., Cui, M. and Liu, Q. (2021) 'A novel word similarity measure method for IoT-enabled healthcare applications.' *Future Generation Computer Systems*, 114, pp. 209-218.

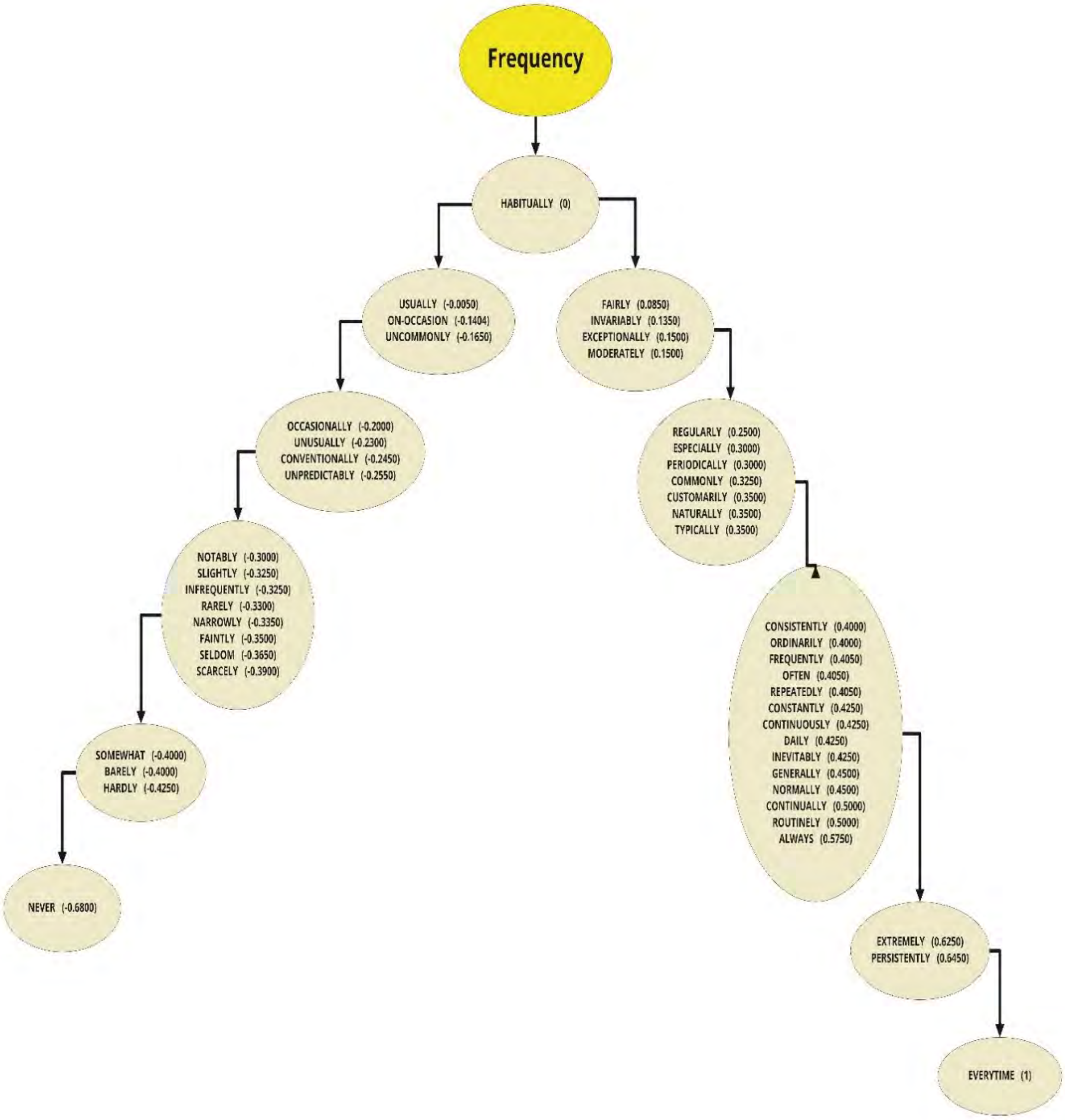
Appendices

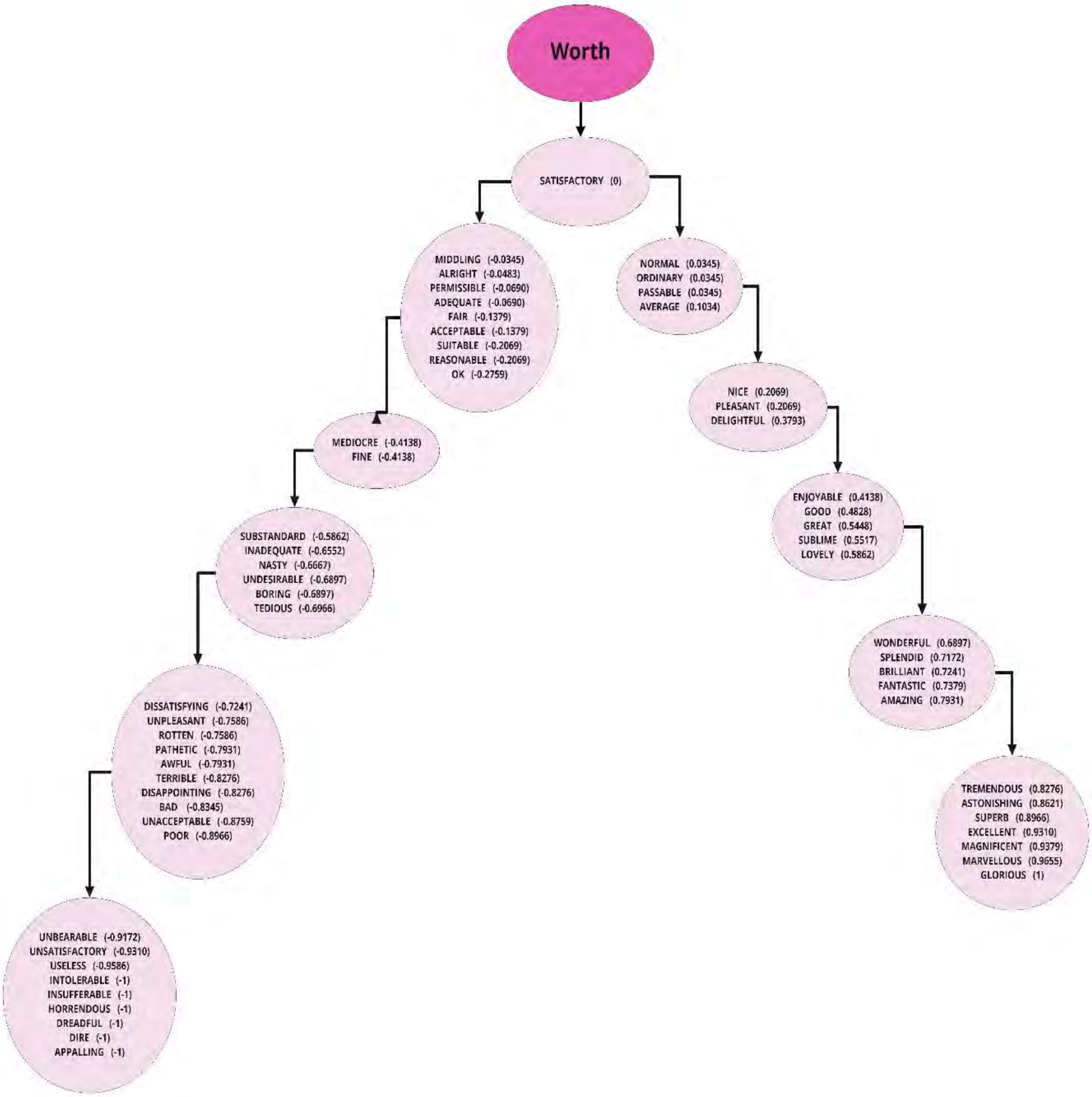
Appendix A - Ontological Structures for FUSE_1.0 (Six Fuzzy Categories)

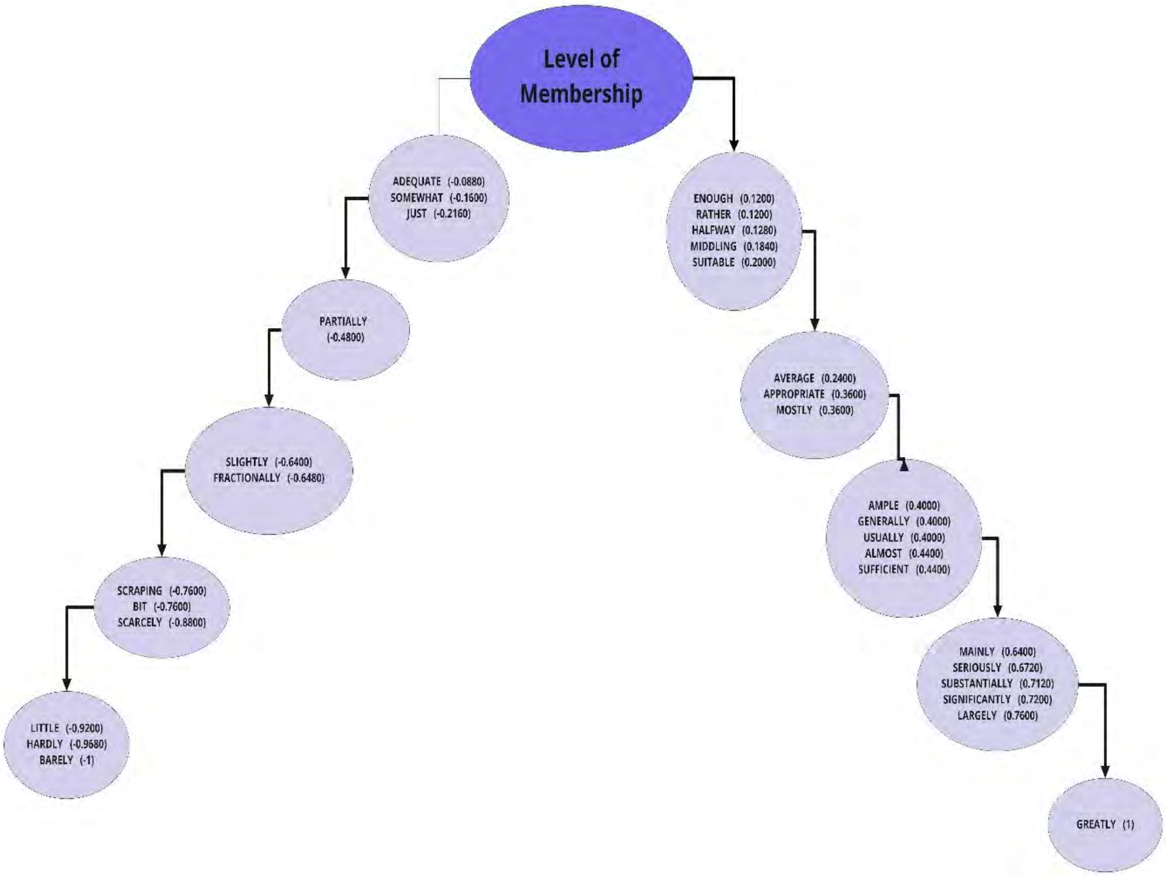



Temperature
BODY-
TEMPERATURE
(0)
MILD (-0.2388)
TEPID (-0.2457)
NIPPY (-0.2803)
FRIGID (0.1003)
BALMY (0.1349)
BRACING (-0.3149)
COOL (-0.4533)
BITING (-0.4533)
TEMPERATE (0.2042)
LUKEWARM (0.2318)
BITTER (-0.5571)
COLD (-0.5779)
BRISK (-0.6263)
CHILLY (-0.6955)
WARM (0.4810)
HUMID (0.5502)
PERSPIRING (0.5502)
SPICY (0.5502)
FROSTY (-0.7093)
ICY (-0.7301)
FREEZING (-0.8962)
BAKING (0.6194)
HOT (0.6194)
SWEATY (0.6886)
SCALDING (0.7509)
HEATED (0.7578)
STEAMING (0.7578)
SWELTERING (0.7924)
ARCTIC (-0.9377)
SUB-ZERO (-1.0000)
FROZEN (-1.0000)
ROASTING (0.8616)
BOILING (0.8893)
SCORCHING (0.9308)
BURNING (1)


Age
TEENAGE (-0.1449)
PREPUBESCENT (-0.2908)
CHILDLIKE (-0.3333)
IMMATURE (-0.3333)
PUBESCENT (-0.4420)
YOUTHFUL (-0.5145)
ADOLESCENT (-0.5145)
YOUNG (-0.5870)
CHILD (-0.5870)
RECENT (-0.6232)
UNDERAGE (-0.6594)
VULNERABLE (-0.7681)
INFANTILE (-0.7899)
EARLIEST (-0.7899)
CHILDISH (-0.8043)
BABYISH (-0.8913)
LATEST (-0.9394)
NEW (-0.9638)
BABY (-1)
MIDDLEAGED (0.0496)
FULL-GROWN (0.0638)
GROWNUP (0.0780)
PRIMORDIAL (0.0797)
PREHISTORIC (0.3333)
JUVENILE (0.4565)
AGED (0.6449)
PRIMEVAL (0.7029)
ADULT (0.7174)
ANTIQUATED (0.7899)
DECREPIT (0.7899)
OLDER (0.7899)
EXPERIENCED (0.8261)
OLD (0.8478)
MATURE (0.8623)
PRIMITIVE (0.8696)
SENIOR (0.8913)
PRIMAL (0.8986)
ELDERLY (0.9275)
ARCHAIC (0.9348)
ANTIQUE (0.9710)
PENSIONABLE (0.9710)
ANCIENT (1)


Frequency
HABITUALLY (0)
USUALLY (-0.0050)
ON-OCCASION (-0.1404)
UNCOMMONLY (-0.1650)
FAIRLY (0.0850)
INVARIABLY (0.1350)
EXCEPTIONALLY (0.1500)
MODERATELY (0.1500)
OCCASIONALLY (-0.2000)
UNUSUALLY (-0.2300)
CONVENTIONALLY (-0.2450)
UNPREDICTABLY (-0.2550)
REGULARLY (0.2500)
ESPECIALLY (0.3000)
PERIODICALLY (0.3000)
COMMONLY (0.3250)
CUSTOMARILY (0.3500)
NATURALLY (0.3500)
TYPICALLY (0.3500)
NOTABLY (-0.3000)
SLIGHTLY (-0.3250)
INFREQUENTLY (-0.3250)
RARELY (-0.3300)
NARROWLY (-0.3350)
FAINTLY (-0.3500)
SELDOM (-0.3650)
SCARCELY (-0.3900)
CONSISTENTLY (0.4000)
ORDINARILY (0.4000)
FREQUENTLY (0.4050)
OFTEN (0.4050)
REPEATEDLY (0.4050)
CONSTANTLY (0.4250)
CONTINUOUSLY (0.4250)
DAILY (0.4250)
INEVITABLY (0.4250)
GENERALLY (0.4500)
NORMALLY (0.4500)
CONTINUALLY (0.5000)
ROUTINELY (0.5000)
ALWAYS (0.5750)
SOMEWHAT (-0.4000)
BARELY (-0.4000)
HARDLY (-0.4250)
NEVER (-0.6800)
EXTREMELY (0.6250)
PERSISTENTLY (0.6450)
EVERYTIME (1)


Worth
SATISFACTORY (0)
MIDDLING (-0.0345)
ALRIGHT (-0.0483)
PERMISSIBLE (-0.0690)
ADEQUATE (-0.0690)
FAIR (-0.1379)
ACCEPTABLE (-0.1379)
SUITABLE (-0.2069)
REASONABLE (-0.2069)
OK (-0.2759)
MEDIOCRE (-0.4138)
FINE (-0.4138)
SUBSTANDARD (-0.5862)
INADEQUATE (-0.6552)
NASTY (-0.6667)
UNDESIRABLE (-0.6897)
BORING (-0.6897)
TEDIOUS (-0.6966)
DISSATISFYING (-0.7241)
UNPLEASANT (-0.7586)
ROTTEN (-0.7586)
PATHETIC (-0.7931)
AWFUL (-0.7931)
TERRIBLE (-0.8276)
DISAPPOINTING (-0.8276)
BAD (-0.8345)
UNACCEPTABLE (-0.8759)
POOR (-0.8966)
UNBEARABLE (-0.9172)
UNSATISFACTORY (-0.9310)
USELESS (-0.9586)
INTOLERABLE (-1)
INSUFFERABLE (-1)
HORRENDOUS (-1)
DREADFUL (-1)
DIRE (-1)
APPALLING (-1)
NORMAL (0.0345)
ORDINARY (0.0345)
PASSABLE (0.0345)
AVERAGE (0.1034)
NICE (0.2069)
PLEASANT (0.2069)
DELIGHTFUL (0.3793)
ENJOYABLE (0.4138)
GOOD (0.4828)
GREAT (0.5448)
SUBLIME (0.5517)
LOVELY (0.5862)
WONDERFUL (0.6897)
SPLENDID (0.7172)
BRILLIANT (0.7241)
FANTASTIC (0.7379)
AMAZING (0.7931)
TREMENDOUS (0.8276)
ASTONISHING (0.8621)
SUPERB (0.8966)
EXCELLENT (0.9310)
MAGNIFICENT (0.9379)
MARVELLOUS (0.9655)
GLORIOUS (1)


Level of Membership
ADEQUATE (-0.0880)
SOMEWHAT (-0.1600)
JUST (-0.2160)
PARTIALLY (-0.4800)
SLIGHTLY (-0.6400)
FRACTIONALLY (-0.6480)
SCRAPING (-0.7600)
BIT (-0.7600)
SCARCELY (-0.8800)
LITTLE (-0.9200)
HARDLY (-0.9680)
BARELY (-1)
ENOUGH (0.1200)
RATHER (0.1200)
HALFWAY (0.1280)
MIDDLING (0.1840)
SUITABLE (0.2000)
AVERAGE (0.2400)
APPROPRIATE (0.3600)
MOSTLY (0.3600)
AMPLE (0.4000)
GENERALLY (0.4000)
USUALLY (0.4000)
ALMOST (0.4400)
SUFFICIENT (0.4400)
MAINLY (0.6400)
SERIOUSLY (0.6720)
SUBSTANTIALLY (0.7120)
SIGNIFICANTLY (0.7200)
LARGELY (0.7600)
GREATLY (1)

Appendix B - Section 5.5 Results for FUSE_1.0

Table B-1 - MWFD Dataset Results

Sentence Pairs	Sentences	AHR	STASIS	FAST	FUSE_1.0
SP 1	How marvellous middling Piccola must have been How good poor Piccola must have been	0.5623	0.8675	0.8965	0.8952
SP 2	A frosty youthful man A hot old man	0.1715	0.4019	0.7473	0.5218
SP 3	Had you married you must have been regularly acceptable Had you married you must have been always poor	0.3769	0.7140	0.8973	0.9141
SP 4	The little village of Resina is also situated near the spot He seems an excellent man and I think him uncommonly pleasing	0.0750	0.2370	0.1978	0.2068
SP 5	They hint that all whales on-occasion smell amazing They hint that all whales always smell bad	0.3708	0.8780	0.8916	0.8543
SP 6	The eyes were full of a frosty and frozen wrath a kind of utterly heartless hatred The eyes were full of a frozen and icy wrath a kind of utterly heartless hatred	0.8350	0.9968	0.9840	0.9937
SP 7	Mr Brown broke into a mostly antiquated giggle Mr Brown broke into a rather childish giggle	0.5677	0.8979	0.9197	0.8925

SP 8	**An unacceptable watcher and very dietetically pathetic is Dr Bunger** **A great watcher and very dietetically severe is Dr Bunger**	0.3842	0.9464	0.9066	0.8933
SP 9	**Have massive mercy on the mediocre men** **Have a little mercy on the poor men**	0.4873	0.7940	0.8074	0.8428
SP 10	**Behold how fine a matter an adjacent fire kindleth** **Behold how great a matter a little fire kindleth**	0.6865	0.8989	0.9618	0.9494
SP 11	**A little quickness of voice there is which rather hurts the ear** **The only living thing near was an old bony grey donkey**	0.1223	0.5430	0.5730	0.5784
SP 12	**And he laughed almost dreadfully** **And he laughed rather unpleasantly**	0.7127	0.4997	0.6269	0.6284
SP 13	**That is somewhat the acceptable complication** **That is just the awful complication**	0.5285	0.8597	0.9095	0.9221
SP 14	**But why the fantastic youthful playthings** **But why the nice new playthings**	0.5938	0.8426	0.9403	0.9459
SP 15	**The advantages of Bath to the child are pretty sufficiently understood** **The advantages of Bath to the young are pretty generally understood**	0.7381	0.9211	0.9230	0.9111

SP 16	A thick Juvenile man A little old man	0.3238	0.6595	0.8202	0.8253
SP 17	He seems a great decrepit party, I remarked He seems a pleasant old party, I remarked	0.4312	0.8106	0.9344	0.9211
SP 18	It is as long again as almost all we have had before was scarcely less warm than hers and whose mind -- Oh	0.1446	0.3340	0.3266	0.3525
SP 19	Keeping at the midpoint of the lake we were on-occasion visited by small tame cows and calves the women and children of this routed host Keeping at the centre of the lake we were occasionally visited by small tame cows and calves the women and children of this routed host	0.7792	0.9748	0.9677	0.9677
SP 20	It is largely a sizeable story, said Turnbull smiling It is rather a long story, said Turnbull smiling	0.7815	0.9130	0.9438	0.9410
SP 21	Do not treat the little Stars so, said the good Moon Mrs Price s last baking failed for want of good barm	0.2112	0.6251	0.6251	0.6251
SP 22	We will not say how small for fear of shocking the youthful ladies We will not say how near for fear of shocking the young ladies	0.6250	0.9462	0.9891	0.9890

SP 23	**She constantly travels with her own sheets an excellent precaution** **She always travels with her own sheets an excellent precaution**	0.8162	0.9989	0.9959	0.9961
SP 24	**This is just the latest movement in a continuing trend towards open source support of business applications** **This is just the latest movement in a continuing trend toward open-source support among business application vendors**	0.7215	0.8414	0.8440	0.8433
SP 25	**Yesterday's ruling is a great first step toward better coverage for poor Maine residents he said but there is more to be done** **He said the court 's ruling was a great first step toward better coverage for poor Maine residents but that there was more to be done.**	0.7485	0.8860	0.8860	0.8861
SP 26	**Some people were habitually cross when they were temperate** **Some people were always cross when they were hot**	0.6331	0.7462	0.8820	0.9177
SP 27	**But Mr Weston is just a recent man** **But Mr Weston is almost an old man**	0.3842	0.9562	0.9709	0.9473
SP 28	**If indeed it could be restored to our poor little boy --"** **Almost sobbed the young man who was in the highest spirits**	0.1269	0.4396	0.4348	0.4372
SP 29	**So would useless diminutive Harriet** **So would poor little Harriet**	0.6069	0.7141	0.9089	0.9647

SP 30	**What’s the fine pensionable man** **What's the good old man**	0.6488	0.7478	0.9675	0.9223

Table B-2 - STSS-65 Dataset Results

Sentence Pairs	Sentences	AHR	STASIS	FAST	FUSE_1.0
SP 1	Cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.0100	0.4242	0.4242	0.4242
SP 2	A rooster is an adult male chicken. A voyage is a long journey on a ship or in a spacecraft.	0.0050	0.2923	0.2902	0.2902
SP 3	Noon is 12 o'clock in the middle of the day. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	0.0125	0.4978	0.5172	0.5172
SP 4	Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat. A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.	0.0475	0.7614	0.7552	0.7533
SP 5	An autograph is the signature of someone famous which is specially written for a fan to keep. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.0050	0.5093	0.5093	0.5093
SP 6	An automobile is a car. In legends and fairy stories, a wizard is a man who has magic powers.	0.0200	0.3334	0.3334	0.3334

SP 7	**A mound of something is a large rounded pile of it.** **A stove is a piece of equipment which provides heat, either for cooking or for heating a room.**	0.0050	0.6664	0.6664	0.6664
SP 8	**A grin is a broad smile.** **An implement is a tool or other piece of equipment.**	0.0050	0.4244	0.4244	0.4244
SP 9	**An Asylum is a psychiatric hospital.** **Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.**	0.0050	0.3743	0.3743	0.3743
SP 10	**An Asylum is a psychiatric hospital.** **A monk is a member of a male religious community that is usually separated from the outside world.**	0.0375	0.3620	0.3517	0.3576
SP 11	**A graveyard is an area of land, sometimes near a church, where dead people are buried.** **If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.**	0.0225	0.5514	0.5476	0.5651
SP 12	**Glass is a hard transparent substance that is used to make things such as windows and bottles.** **A magician is a person who entertains people by doing magic tricks.**	0.0075	0.5414	0.5414	0.5414
SP 13	**A boy is a child who will grow up to be a man.** **A rooster is an adult male chicken.**	0.1075	0.6352	0.6352	0.6352
SP 14	**A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable.**	0.0525	0.6939	0.6939	0.6939

	A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.				
SP 15	A monk is a member of a male religious community that is usually separated from the outside world. A slave is someone who is the property of another person and has to work for that person.	0.0450	0.8030	0.7860	0.7958
SP 16	An Asylum is a psychiatric hospital. A cemetery is a place where dead people's bodies or their ashes are buried.	0.0375	0.4387	0.4387	0.4387
SP 17	The coast is an area of land that is next to the sea. A forest is a large area where trees grow close together.	0.0475	0.6315	0.6315	0.6315
SP 18	A grin is a broad smile. A lad is a young man or boy.	0.0125	0.5441	0.5441	0.5441
SP 19	The shores or shore of a sea, lake or wide river is the land along the edge of it. Woodland is land with a lot of trees.	0.0825	0.7308	0.7308	0.7308
SP 20	A monk is a member of a male religious community that is usually separated from the outside world. In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.1125	0.6124	0.5917	0.5994

SP 21	**A boy is a child who will grow up to be a man.** **A sage is a person who is regarded as being very wise.**	0.0425	0.6169	0.6169	0.6169
SP 22	**An automobile is a car.** **A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable.**	0.0200	0.5235	0.5235	0.5235
SP 23	**A mound of something is a large rounded pile of it.** **The shores or shore of a sea, lake or wide river is the land along the edge of it.**	0.0350	0.6299	0.6274	0.6274
SP 24	**A lad is a young man or boy.** **In legends and fairy stories, a wizard is a man who has magic powers.**	0.0325	0.5641	0.5641	0.5641
SP 25	**A forest is a large area where trees grow close together.** **A graveyard is an area of land, sometimes near a church, where dead people are buried.**	0.0650	0.6993	0.7117	0.7001
SP 26	**Food is what people and animals eat. A rooster is an adult male chicken.**	0.0550	0.6855	0.6855	0.6855
SP 27	**A cemetery is a place where dead people's bodies or their ashes are buried.** **Woodland is land with a lot of trees.**	0.0375	0.7073	0.6981	0.6981

SP 28	**The shores or shore of a sea, lake or wide river is the land along the edge of it.** **A voyage is a long journey on a ship or in a spacecraft.**	0.0200	0.4052	0.4023	0.4023
SP 29	**A bird is a creature with feathers and wings, females lay eggs and most birds can fly.** **Woodland is land with a lot of trees.**	0.0125	0.6425	0.6346	0.6346
SP 30	**The coast is an area of land that is next to the sea.** **A hill is an area of land that is higher than the land that surrounds it.**	0.1000	0.7033	0.6919	0.6919
SP 31	**A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.** **An implement is a tool or other piece of equipment.**	0.0500	0.6229	0.6123	0.6104
SP 32	**A crane is a large machine that moves heavy things by lifting them in the air. A rooster is an adult male chicken.**	0.0200	0.5919	0.5919	0.5919
SP 33	**A hill is an area of land that is higher than the land that surrounds it. Woodland is land with a lot of trees.**	0.1450	0.7350	0.7130	0.7130
SP 34	**A car is a motor vehicle with room for a small number of passengers.** **When you make a journey, you travel from one place to another.**	0.0725	0.5316	0.5235	0.5235
SP 35	**A cemetery is a place where dead people's bodies or their ashes are buried.** **A mound of something is a large rounded pile of it.**	0.0575	0.6139	0.6139	0.6139

SP 36	**Glass is a hard transparent substance that is used to make things such as windows and bottles.** **A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.**	0.1075	0.6969	0.6969	0.6969
SP 37	**A magician is a person who entertains people by doing magic tricks.** **In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.**	0.1300	0.6122	0.6122	0.6122
SP 38	**A crane is a large machine that moves heavy things by lifting them in the air. An implement is a tool or other piece of equipment.**	0.1850	0.7155	0.7072	0.7072
SP 39	**Your brother is a boy or a man who has the same parents as you.** **A lad is a young man or boy.**	0.1275	0.7540	0.7540	0.7540
SP 40	**A sage is a person who is regarded as being very wise.** **In legends and fairy stories, a wizard is a man who has magic powers.**	0.1525	0.5495	0.5495	0.5495
SP 41	**In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.** **A sage is a person who is regarded as being very wise.**	0.2825	0.6725	0.6725	0.6725
SP 42	**A bird is a creature with feathers and wings, females lay eggs and most birds can fly.** **A crane is a large machine that moves heavy things by lifting them in the air.**	0.0350	0.7849	0.7849	0.7849

SP 43	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A cock is an adult male chicken.	0.1625	0.7973	0.7973	0.7973
SP 44	Food is what people and animals eat. Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.2425	0.6431	0.6431	0.6431
SP 45	Your brother is a boy or a man who has the same parents as you. A monk is a member of a male religious community that is usually separated from the outside world.	0.0450	0.8108	0.7928	0.8031
SP 46	An Asylum is a psychiatric hospital. If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.	0.2150	0.6793	0.6793	0.6793
SP 47	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam. A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.3475	0.7515	0.7450	0.7429
SP 48	A magician is a person who entertains people by doing magic tricks. In legends and fairy stories, a wizard is a man who has magic powers.	0.3550	0.7589	0.7589	0.7589
SP 49	A hill is an area of land that is higher than the land that surrounds it. A mound of something is a large rounded pile of it.	0.2925	0.6166	0.6166	0.6166

SP 50	Cord is strong, thick string. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	0.4700	0.8078	0.8108	0.8108
SP 51	Glass is a hard transparent substance that is used to make things such as windows and bottles. A tumbler is a drinking glass with straight sides.	0.1375	0.7873	0.7720	0.7720
SP 52	A grin is a broad smile. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.4850	0.6299	0.6299	0.6299
SP 53	In former times, serfs were a class of people who had to work on a particular person's land and could not leave without that person's permission. A slave is someone who is the property of another person and has to work for that person.	0.4825	0.7696	0.7696	0.7696
SP 54	A When you make a journey, you travel from one place to another. A voyage is a long journey on a ship or in a spacecraft.	0.3600	0.7242	0.7188	0.7188
SP 55	An autograph is the signature of someone famous which is specially written for a fan to keep. Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says.	0.4050	0.7649	0.7649	0.7579
SP 56	The coast is an area of land that is next to the sea. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.5875	0.8591	0.8591	0.8591

SP 57	A forest is a large area where trees grow close together. Woodland is land with a lot of trees.	0.6275	0.7908	0.7816	0.7816
SP 58	An implement is a tool or other piece of equipment. A tool is any instrument or simple piece of equipment that you hold in your hands and use to do a particular kind of work.	0.5900	0.8068	0.8068	0.8068
SP 59	A cock is an adult male chicken. A rooster is an adult male chicken.	0.8625	0.9999	0.9999	0.9999
SP 60	A boy is a child who will grow up to be a man. A lad is a young man or boy.	0.5800	0.7248	0.7420	0.7593
SP 61	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A pillow is a rectangular cushion which you rest your head on when you are in bed.	0.5225	0.8176	0.8176	0.8176
SP 62	A cemetery is a place where dead people's bodies or their ashes are buried. A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.7725	0.8178	0.8149	0.8149
SP 63	An automobile is a car. A car is a motor vehicle with room for a small number of passengers.	0.5575	0.7017	0.7017	0.7017
SP 64	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	0.9550	0.9983	0.9983	0.9983

SP 65	**A gem is a jewel or stone that is used in jewellery.** **A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.**	0.6525	0.8934	0.8934	0.8934

Table B-3 - STSS-131 Dataset Results

Sentence Pairs	Sentences	AHR	STASIS	FAST	FUSE_1.0
SP 1	Cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.0100	0.4242	0.4242	0.4242
SP 2	A rooster is an adult male chicken. A voyage is a long journey on a ship or in a spacecraft.	0.0050	0.2923	0.2902	0.2902
SP 3	Noon is 12 o'clock in the middle of the day. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	0.0125	0.4978	0.5172	0.5172
SP 4	Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat. A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.	0.0475	0.7614	0.7552	0.7533
SP 5	An autograph is the signature of someone famous which is specially written for a fan to keep. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.0050	0.5093	0.5093	0.5093
SP 6	An automobile is a car. In legends and fairy stories, a wizard is a man who has magic powers.	0.0200	0.3334	0.3334	0.3334

SP 7	A mound of something is a large rounded pile of it. A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.0050	0.6664	0.6664	0.6664
SP 8	A grin is a broad smile. An implement is a tool or other piece of equipment.	0.0050	0.4244	0.4244	0.4244
SP 9	An Asylum is a psychiatric hospital. Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.0050	0.3743	0.3743	0.3743
SP 10	An Asylum is a psychiatric hospital. A monk is a member of a male religious community that is usually separated from the outside world.	0.0375	0.3620	0.3517	0.3576
SP 11	A graveyard is an area of land, sometimes near a church, where dead people are buried. If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.	0.0225	0.5514	0.5476	0.5651
SP 12	Glass is a hard transparent substance that is used to make things such as windows and bottles. A magician is a person who entertains people by doing magic tricks.	0.0075	0.5414	0.5414	0.5414
SP 13	A boy is a child who will grow up to be a man. A rooster is an adult male chicken.	0.1075	0.6352	0.6352	0.6352
SP 14	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.0525	0.6939	0.6939	0.6939

SP 15	**A monk is a member of a male religious community that is usually separated from the outside world.** **A slave is someone who is the property of another person and has to work for that person.**	0.0450	0.8030	0.7860	0.7958
SP 16	**An Asylum is a psychiatric hospital.** **A cemetery is a place where dead people's bodies or their ashes are buried.**	0.0375	0.4387	0.4387	0.4387
SP 17	**The coast is an area of land that is next to the sea.** **A forest is a large area where trees grow close together.**	0.0475	0.6315	0.6315	0.6315
SP 18	**A grin is a broad smile.** **A lad is a young man or boy.**	0.0125	0.5441	0.5441	0.5441
SP 19	**The shores or shore of a sea, lake or wide river is the land along the edge of it.** **Woodland is land with a lot of trees.**	0.0825	0.7308	0.7308	0.7308
SP 20	**A monk is a member of a male religious community that is usually separated from the outside world.** **In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.**	0.1125	0.6124	0.5917	0.5994
SP 21	**A boy is a child who will grow up to be a man.** **A sage is a person who is regarded as being very wise.**	0.0425	0.6169	0.6169	0.6169

SP 22	An automobile is a car. A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable.	0.0200	0.5235	0.5235	0.5235
SP 23	A mound of something is a large rounded pile of it. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.0350	0.6299	0.6274	0.6274
SP 24	A lad is a young man or boy. In legends and fairy stories, a wizard is a man who has magic powers.	0.0325	0.5641	0.5641	0.5641
SP 25	A forest is a large area where trees grow close together. A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.0650	0.6993	0.7117	0.7001
SP 26	Food is what people and animals eat. A rooster is an adult male chicken.	0.0550	0.6855	0.6855	0.6855
SP 27	A cemetery is a place where dead people's bodies or their ashes are buried. Woodland is land with a lot of trees.	0.0375	0.7073	0.6981	0.6981
SP 28	The shores or shore of a sea, lake or wide river is the land along the edge of it. A voyage is a long journey on a ship or in a spacecraft.	0.0200	0.4052	0.4023	0.4023

SP 29	**A bird is a creature with feathers and wings, females lay eggs and most birds can fly.** **Woodland is land with a lot of trees.**	0.0125	0.6425	0.6346	0.6346
SP 30	**The coast is an area of land that is next to the sea.** **A hill is an area of land that is higher than the land that surrounds it.**	0.1000	0.7033	0.6919	0.6919
SP 31	**A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.** **An implement is a tool or other piece of equipment.**	0.0500	0.6229	0.6123	0.6104
SP 32	**A crane is a large machine that moves heavy things by lifting them in the air. A rooster is an adult male chicken.**	0.0200	0.5919	0.5919	0.5919
SP 33	**A hill is an area of land that is higher than the land that surrounds it. Woodland is land with a lot of trees.**	0.1450	0.7350	0.7130	0.7130
SP 34	**A car is a motor vehicle with room for a small number of passengers.** **When you make a journey, you travel from one place to another.**	0.0725	0.5316	0.5235	0.5235
SP 35	**A cemetery is a place where dead people's bodies or their ashes are buried.** **A mound of something is a large rounded pile of it.**	0.0575	0.6139	0.6139	0.6139
SP 36	**Glass is a hard transparent substance that is used to make things such as windows and bottles.** **A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.**	0.1075	0.6969	0.6969	0.6969

SP 37	A magician is a person who entertains people by doing magic tricks. In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.1300	0.6122	0.6122	0.6122
SP 38	A crane is a large machine that moves heavy things by lifting them in the air. An implement is a tool or other piece of equipment.	0.1850	0.7155	0.7072	0.7072
SP 39	Your brother is a boy or a man who has the same parents as you. A lad is a young man or boy.	0.1275	0.7540	0.7540	0.7540
SP 40	A sage is a person who is regarded as being very wise. In legends and fairy stories, a wizard is a man who has magic powers.	0.1525	0.5495	0.5495	0.5495
SP 41	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth. A sage is a person who is regarded as being very wise.	0.2825	0.6725	0.6725	0.6725
SP 42	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A crane is a large machine that moves heavy things by lifting them in the air.	0.0350	0.7849	0.7849	0.7849
SP 43	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A cock is an adult male chicken.	0.1625	0.7973	0.7973	0.7973
SP 44	Food is what people and animals eat. Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.2425	0.6431	0.6431	0.6431

SP 45	**Your brother is a boy or a man who has the same parents as you.** **A monk is a member of a male religious community that is usually separated from the outside world.**	0.0450	0.8108	0.7928	0.8031
SP 46	**An Asylum is a psychiatric hospital.** **If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.**	0.2150	0.6793	0.6793	0.6793
SP 47	**A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.** **A stove is a piece of equipment which provides heat, either for cooking or for heating a room.**	0.3475	0.7515	0.7450	0.7429
SP 48	**A magician is a person who entertains people by doing magic tricks.** **In legends and fairy stories, a wizard is a man who has magic powers.**	0.3550	0.7589	0.7589	0.7589
SP 49	**A hill is an area of land that is higher than the land that surrounds it.** **A mound of something is a large rounded pile of it.**	0.2925	0.6166	0.6166	0.6166
SP 50	**Cord is strong, thick string.** **String is thin rope made of twisted threads, used for tying things together or tying up parcels.**	0.4700	0.8078	0.8108	0.8108
SP 51	**Glass is a hard transparent substance that is used to make things such as windows and bottles.** **A tumbler is a drinking glass with straight sides.**	0.1375	0.7873	0.7720	0.7720
SP 52	**A grin is a broad smile.** **A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.**	0.4850	0.6299	0.6299	0.6299

SP 53	**In former times, serfs were a class of people who had to work on a particular person's land and could not leave without that person's permission.** **A slave is someone who is the property of another person and has to work for that person.**	0.4825	0.7696	0.7696	0.7696
SP 54	**A When you make a journey, you travel from one place to another.** **A voyage is a long journey on a ship or in a spacecraft.**	0.3600	0.7242	0.7188	0.7188
SP 55	**An autograph is the signature of someone famous which is specially written for a fan to keep.** **Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says.**	0.4050	0.7649	0.7649	0.7579
SP 56	**The coast is an area of land that is next to the sea.** **The shores or shore of a sea, lake or wide river is the land along the edge of it.**	0.5875	0.8591	0.8591	0.8591
SP 57	**A forest is a large area where trees grow close together.** **Woodland is land with a lot of trees.**	0.6275	0.7908	0.7816	0.7816
SP 58	**An implement is a tool or other piece of equipment.** **A tool is any instrument or simple piece of equipment that you hold in your hands and use to do a particular kind of work.**	0.5900	0.8068	0.8068	0.8068
SP 59	**A cock is an adult male chicken.** **A rooster is an adult male chicken.**	0.8625	0.9999	0.9999	0.9999

SP 60	A boy is a child who will grow up to be a man. A lad is a young man or boy.	0.5800	0.7248	0.7420	0.7593
SP 61	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A pillow is a rectangular cushion which you rest your head on when you are in bed.	0.5225	0.8176	0.8176	0.8176
SP 62	A cemetery is a place where dead people's bodies or their ashes are buried. A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.7725	0.8178	0.8149	0.8149
SP 63	An automobile is a car. A car is a motor vehicle with room for a small number of passengers.	0.5575	0.7017	0.7017	0.7017
SP 64	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	0.9550	0.9983	0.9983	0.9983
SP 65	A gem is a jewel or stone that is used in jewellery. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.6525	0.8934	0.8934	0.8934
SP 66	Would you like to go out to drink with me tonight? I really don’t know what to eat tonight so I might go out somewhere.	0.2525	0.4370	0.4370	0.4370
SP 67	I advise you to treat this matter very seriously as it is vital. You must take this most seriously, it will affect you.	0.8450	0.7271	0.7271	0.7587
SP 68	When I was going out to meet my friends there was a delay at the train station.	0.7825	0.7948	0.7948	0.7948

	The train operator announced to the passengers that the train would be delayed.				
SP 69	Does music help you to relax, or does it distract you too much? Does this sponge look wet or dry to you?	0.0250	0.5462	0.5462	0.5462
SP 70	You must realise that you will definitely be punished if you play with the alarm. He will be harshly punished for setting the fire alarm off.	0.7100	0.6840	0.6840	0.6840
SP 71	I will make you laugh so much that your sides ache. When I tell you this you will split your sides laughing.	0.9375	0.6893	0.6880	0.6880
SP 72	You shouldn't be covering what you really feel. There is no point in covering up what you said, we all know.	0.5525	0.5776	0.5793	0.5793
SP 73	Do you want to come with us to the pub behind the hill? We are going out for drinks tonight in Salford Quays if you would like to come.	0.4550	0.4169	0.4169	0.4169
SP 74	This key doesn't seem to be working, could you give me another? I dislike the word quay, it confuses me, I always think of things for locks, there's another one.	0.1800	0.4609	0.4573	0.4577
SP 75	The ghost appeared from nowhere and frightened the old man. The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.	0.3625	0.4950	0.4950	0.4950
SP 76	You're not a good friend if you're not prepared to be present when I need you. A good friend always seems to be present when you need them.	0.7850	0.4902	0.4713	0.4732
SP 77	The children crossed the road very safely thanks to the help of the lollipop lady. It was feared that the child might not recover, because he was seriously ill.	0.0325	0.4102	0.4035	0.3851
SP 78	I have invited a variety of people to my party so it should be interesting.	0.5450	0.7168	0.7168	0.7168

	A number of invitations were given out to a variety of people inviting them down the pub.				
SP 79	I offer my condolences to the parents of John Smith, who was unfortunately murdered. I express my sympathy to John Smith's parents following his murder.	0.9775	0.7837	0.7837	0.7837
SP 80	Boats come in all shapes and sizes but they all do the same thing. Chairs can be comfy and not comfy, depending on the chair.	0.1250	0.3741	0.3741	0.3741
SP 81	If you continuously use these products, I guarantee you will look very young. I assure you that, by using these products consistently over a long period of time, you will appear really young.	0.8950	0.8573	0.8021	0.8772
SP 82	We ran farther than the other children that day. You ran farther than anyone today.	0.6075	0.6954	0.6967	0.6967
SP 83	I always like to have a slice of lemon in my drink especially if it's Coke. I like to put a wedge of lemon in my drinks, especially cola.	0.9525	0.9140	0.9033	0.9148
SP 84	It seems like I've got eczema on my ear doctor, can you recommend something for me? I had to go to a chemist for a special rash cream for my ear.	0.5125	0.4542	0.4542	0.4542
SP 85	I am proud of our nation, well, most of it. I think of myself as being part of a nation.	0.4275	0.6528	0.6528	0.6528
SP 86	There was a heap of rubble left by the builders outside my house this morning. Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	0.0225	0.4989	0.4989	0.4989
SP 87	Water freezes at a certain temperature, which is zero degrees Celsius. The temperature of boiling water is 100 C and the temperature of ice is 0 C.	0.7700	0.7886	0.7956	0.7956
SP 88	We got home safely in the end, although it was a long journey.	0.7650	0.7949	0.7745	0.7745

	Though it took many hours travel, we finally reached our house safely.				
SP 89	A man called Dave gave his fiance´e a large diamond ring for their engagement. The man presented a diamond to the woman and asked her to marry him.	0.8050	0.6038	0.5886	0.5886
SP 90	I used to run quite a lot, in fact once I ran for North Tyneside. I used to climb lots at school as we had a new climbing wall put in the gym.	0.1850	0.5739	0.5705	0.5705
SP 91	I love to laugh as it makes me happy as well as those around me. I thought we bargained that it would only cost me a pound.	0.0200	0.5100	0.5100	0.5100
SP 92	Because I am the eldest one I should be more responsible. Just because of my age, people shouldn’t think I’m a responsible adult, but they do?	0.5575	0.3791	0.3742	0.3749
SP 93	I need to dash into the kitchen because I think my chip pan is on fire. In the event of a chip pan fire follow the instructions on the safety note.	0.4250	0.7556	0.7556	0.7556
SP 94	Peter was a very large youth, whose size intimidated most people, much to his delight. Now I wouldn’t say he was fat, but I’d certainly say he was one of the larger boys.	0.4900	0.3638	0.3638	0.3638
SP 95	I’m going to buy a grey jumper today, in half an hour. That’s a nice grey top, where did you get it from?	0.3125	0.4737	0.4737	0.4737
SP 96	We got soaked in the rain today, but now we are nice and dry. I was absolutely soaking wet last night, I drove my bike through the worst weather.	0.4200	0.6932	0.6932	0.6932
SP 97	Global warming is what everyone is worrying about today. The problem of global warming is a concern to every country in the world at the moment.	0.7850	0.7330	0.7275	0.7275
SP 98	He was harshly punished for setting the fire alarms off. He delayed his response, in order to create a tense atmosphere.	0.0550	0.4935	0.4935	0.4935

SP 99	**Midday is 12 o'clock in the middle of the day.** **Noon is 12 o'clock in the middle of the day.**	0.9900	0.9984	0.9984	0.9984
SP 100	**That's not a very good car, on the other hand mine is great.** **This is a terrible noise level for a new car.**	0.2625	0.4783	0.4612	0.4496
SP 101	**There was a terrible accident, a pileup, on the M16 today.** **It was a terrible accident, no one believed it was possible.**	0.5825	0.5424	0.5797	0.5808
SP 102	**After hours of getting lost we eventually arrived at the hotel.** **After walking against the strong wind for hours he finally returned home safely.**	0.2725	0.6174	0.6174	0.6174
SP 103	**The first thing I do in a morning is make myself a cup of coffee.** **The first thing I do in the morning is have a cup of coffee.**	0.9625	0.9179	0.9179	0.9179
SP 104	**Someone spilt a drink accidentally on my shirt, so I changed it.** **It appears to have shrunk, it wasn't that size before I washed it.**	0.1200	0.4546	0.4546	0.4546
SP 105	**I'm worried most seriously about the presentation, not the essay.** **It is mostly very difficult to gain full marks in today's exam.**	0.1925	0.4518	0.4421	0.5063
SP 106	**It is mostly very difficult to gain full marks in today's exam.** **The exam was really difficult, I've got no idea if I'm going to pass.**	0.6350	0.5586	0.5466	0.5466
SP 107	**Meet me on the hill behind the church in half an hour.** **Join me on the hill at the back of the church in thirty minutes time.**	0.9825	0.7828	0.7828	0.7828
SP 108	**If you don't console with a friend, there is a chance you may hurt their feelings.** **One of the qualities of a good friend is the ability to console.**	0.7525	0.6722	0.6722	0.6722
SP 109	**We tried to bargain with him but it made no difference, he still didn't change his mind.** **I tried bargaining with him, but he just wouldn't listen.**	0.8575	0.5624	0.5621	0.5628
SP 110	**It gives me great pleasure to announce the winner of this year's beauty pageant.**	0.9700	0.7351	0.7351	0.7351

	It’s a real pleasure to tell you who has won our annual beauty parade.				
SP 111	They said they were hoping to go to America on holiday. I like to cover myself up in lots of layers, I don’t like the cold.	0.0400	0.5291	0.5291	0.5291
SP 112	Will I have to drive far to get to the nearest petrol station? Is it much farther for me to drive to the next gas station?	0.9600	0.8747	0.8747	0.8344
SP 113	I think I know her from somewhere because she has a familiar face. You have a very familiar face, where do I know you from?	0.8400	0.6994	0.6754	0.6754
SP 114	I am sorry but I can’t go out as I have a heap of work to do. I’ve a heap of things to finish so I can’t go out I’m afraid.	0.9000	0.6224	0.6224	0.6224
SP 115	The responsible man felt very guilty when he crashed into the back of someone’s car. A slow driver can be annoying even though they are driving safely.	0.2200	0.6205	0.6205	0.6205
SP 116	Get that wet dog off my brand new white sofa. Make that wet hound get off my white couch – I only just bought it.	0.8975	0.9091	0.8948	0.8969
SP 117	He fought in the war in Iraq before being killed in a car crash. The prejudice I suffered whilst on holiday in Iraq was quite alarming.	0.1375	0.5117	0.5117	0.5117
SP 118	The cat was hungry so he went into the back garden to find lunch. The hen walked about in the yard eating tasty grain.	0.3000	0.6940	0.6940	0.6940
SP 119	My bedroom wall is lemon coloured but my mother says it is yellow. Roses can be different colours, it has to be said red is the best though.	0.1700	0.8521	0.8521	0.8521
SP 120	Would you like to drink this wine with your meal? Will you drink a glass of wine while you eat?	0.8900	0.7316	0.7323	0.7323
SP 121	Roses can be different colours, it has to be said red is the best though. Roses come in many varieties and colours, but yellow is my favourite.	0.7050	0.8449	0.8449	0.8449
SP 122	Flies can also carry a lot of disease and cause maggots.	0.0300	0.5695	0.5695	0.5695

	I dry my hair after I wash it or I will get ill.				
SP 123	Could you climb up the tree and save my cat from jumping please? Can you get up that tree and rescue my cat otherwise it might jump?	0.9575	0.8652	0.8657	0.8657
SP 124	The pleasure that I get from studying, is that I learn new things. I have a doubt about this exam, we never got to study for it.	0.1850	0.7366	0.7180	0.7183
SP 125	The perpetrators of war crimes are rotten to the core. There are many global issues that everybody should be aware of, such as the threat of terrorism.	0.2375	0.4961	0.4423	0.4423
SP 126	The damp was mostly in the very corner of the room. The young lady was somewhat partially burnt from the sun.	0.0275	0.3615	0.4789	0.4702
SP 127	We often ran to school because we were always late. I knew I was late for my class so I ran all the way to school.	0.7750	0.3556	0.3364	0.3236
SP 128	I hope you're taking this seriously, if not you can get out of here. The difficult course meant that only the strong would survive.	0.1250	0.3991	0.3991	0.3798
SP 129	The shores or shore of a sea, lake or wide river is the land along the edge of it. An autograph is the signature of someone famous which is specially written for a fan to keep.	0.0275	0.5093	0.5093	0.5093
SP 130	I bought a new guitar today, do you like it? The weapon choice reflects the personality of the carrier.	0.0400	0.4746	0.4672	0.4672
SP 131	I am so hungry I could eat a whole horse plus dessert. I could have eaten another meal, I'm still starving.	0.7650	0.6010	0.6010	0.6010

Appendix C - Fuzzy Dictionary for Nine Categories of FUSE_2.0

1 - SIZE/DISTANCE					
MICROSCOPIC	-1	ALONGSIDE	-0.27976	CONSIDERABLE	0.309524
MINUSCULE	-0.88095	ADJACENT	-0.26191	LOADS	0.333333
DINKY	-0.86905	ORDINARY	-0.22619	THICK	0.333333
TEENY	-0.85714	MEDIUM	-0.20238	FAR	0.363095
TITCHY	-0.7381	PROXIMATE	-0.20238	SIZEABLE	0.392857
LITTLE	-0.70833	EQUIDISTANT	-0.14286	LARGE	0.482143
SMALL	-0.70833	TIDY	-0.14286	PRINCELY	0.482143
WEE	-0.70833	USUAL	-0.1131	BOUNDLESS	0.535714
INSIGNIFICANT	-0.70238	AWAY	-0.10119	DISTANT	0.541667
PETITE	-0.64286	NORMAL	-0.10119	WHACKING	0.541667
DIMINUTIVE	-0.58333	PROXIMAL	-0.05357	SUBSTANTIAL	0.60119
NEAREST	-0.58333	REGULAR	-0.05357	BIG	0.660714
PIDDLING	-0.58333	STANDARD	-0.05357	GREAT	0.660714
TINY	-0.55952	BONNY	-0.02381	FARAWAY	0.666667
MINUTE	-0.55357	MEDIAL	0.011905	HEFTY	0.678571
SHORT	-0.52381	AVERAGE	0.029762	LONG	0.684211
UNIMPORTANT	-0.52381	MEAN	0.029762	JUMBO	0.720238
PALTRY	-0.51191	ACCESSIBLE	0.035714	EPIC	0.75
TRIVIAL	-0.5	HALFWAY	0.035714	MASSIVE	0.75
NEAR	-0.47619	ISOLATED	0.047619	OVERSIZED	0.754386
MESIAL	-0.44048	CENTRAL	0.065476	IMMENSE	0.754386
CONJOINING	-0.43452	GOODLY	0.065476	GIANT	0.809524
BESIDE	-0.41071	MIDWAY	0.065476	HUGE	0.827381
ADJOINING	-0.38095	MIDPOINT	0.066667	ENORMOUS	0.833333
THIN	-0.36364	CENTRE	0.066667	MEGA	0.839286
TOKEN	-0.35714	MEDIAN	0.083333	COLOSSUS	0.869048
NEARBY	-0.35119	MIDDLE	0.083333	GIGANTIC	0.892857
QUALITY	-0.35119	MID	0.089286	MAMMOTH	0.894
MOMENT	-0.32143	REMOTE	0.178571	GARGANTUAN	1
NORM	-0.29167	METHODICAL	0.184524		
CLOSE	-0.28571	ABUNDANT	0.214286		

2 - TEMPERATURE					
FROZEN	-1	BRACING	-0.31488	SPICY	0.550173
SUB-ZERO	-1	NIPPY	-0.28028	BAKING	0.619377
ARCTIC	-0.93772	TEPID	-0.24568	HOT	0.619377
FREEZING	-0.89619	MILD	-0.23875	SWEATY	0.688581

ICY	-0.7301	BODY-TEMPERATURE	0	SCALDING	0.750865
FROSTY	-0.70934	FRIGID	0.100346	HEATED	0.757785
CHILLY	-0.6955	BALMY	0.134948	STEAMING	0.757785
BRISK	-0.6263	TEMPERATE	0.204152	SWELTERING	0.792388
COLD	-0.57786	LUKEWARM	0.231834	ROASTING	0.861592
BITTER	-0.55709	WARM	0.480969	BOILING	0.889273
BITING	-0.45329	HUMID	0.550173	SCORCHING	0.930796
COOL	-0.45329	PERSPIRING	0.550173	BURNING	1

3 - AGE

BABY	-1	IMMATURE	-0.333333	OLDER	0.789855
NEW	-0.963768	CHILDLIKE	-0.33333	EXPERIENCED	0.8260869
LATEST	-0.93939	PREPUBESCENT	-0.29078	OLD	0.8478260
BABYISH	-0.891304	TEENAGE	-0.144927	MATURE	0.8623188
CHILDISH	-0.804347	MIDDLEAGED	0.049645	PRIMITIVE	0.8695652
EARLIEST	-0.789855	FULL-GROWN	0.06383	SENIOR	0.8913043
INFANTILE	-0.789855	GROWNUP	0.078014	PRIMAL	0.8985507
VULNERABLE	-0.768115	PRIMORDIAL	0.0797101	ELDERLY	0.9275362
UNDERAGE	-0.659420	PREHISTORIC	0.33333	ARCHAIC	0.9347826
RECENT	-0.623188	JUVENILE	0.4565217	ANTIQUE	0.9710144
CHILD	-0.586956	AGED	0.6449275	PENSIONABLE	0.9710144
YOUNG	-0.586956	PRIMEVAL	0.7028985	ANCIENT	1
ADOLESCENT	-0.514492	ADULT	0.7173913		
YOUTHFUL	-0.514492	ANTIQUATED	0.7898550		
PUBESCENT	-0.442028	DECREPIT	0.7898550		

4 - FREQUENCY

NEVER	-0.68	UNCOMMONLY	-0.165	ORDINARILY	0.4
HARDLY	-0.425	ON-OCCASION	-0.14035	FREQUENTLY	0.405
BARELY	-0.4	USUALLY	-0.005	OFTEN	0.405
SOMEWHAT	-0.4	HABITUALLY	0	REPEATEDLY	0.405
SCARCELY	-0.39	FAIRLY	0.085	CONSTANTLY	0.425
SELDOM	-0.365	INVARIABLY	0.135	CONTINUOUSLY	0.425
FAINTLY	-0.35	EXCEPTIONALLY	0.15	DAILY	0.425
NARROWLY	-0.335	MODERATELY	0.15	INEVITABLY	0.425
RARELY	-0.33	REGULARLY	0.25	GENERALLY	0.45
INFREQUENTLY	-0.325	ESPECIALLY	0.3	NORMALLY	0.45
SLIGHTLY	-0.325	PERIODICALLY	0.3	CONTINUALLY	0.5

NOTABLY	-0.3	COMMONLY	0.325	ROUTINELY	0.5
UNPREDICTABLY	-0.255	CUSTOMARILY	0.35	ALWAYS	0.575
CONVENTIONALLY	-0.245	NATURALLY	0.35	EXTREMELY	0.625
UNUSUALLY	-0.23	TYPICALLY	0.35	PERSISTENTLY	0.645
OCCASIONALLY	-0.2	CONSISTENTLY	0.4		

5 - LEVEL OF MEMBERSHIP

BARELY	-1	ADEQUATE	-0.088	USUALLY	0.4
HARDLY	-0.968	ENOUGH	0.12	ALMOST	0.44
LITTLE	-0.92	RATHER	0.12	SUFFICIENT	0.44
SCARCELY	-0.88	HALFWAY	0.128	MAINLY	0.64
BIT	-0.76	MIDDLING	0.184	SERIOUSLY	0.672
SCRAPING	-0.76	SUITABLE	0.2	SUBSTANTIALLY	0.712
FRACTIONALLY	-0.648	AVERAGE	0.24	SIGNIFICANTLY	0.72
SLIGHTLY	-0.64	APPROPRIATE	0.36	LARGELY	0.76
PARTIALLY	-0.48	MOSTLY	0.36	GREATLY	1
JUST	-0.216	AMPLE	0.4	SUITABLE	0.2
SOMEWHAT	-0.16	GENERALLY	0.4		

6 - WORTH

APPALLING	-1	UNDESIRABLE	-0.68965	PLEASANT	0.2068965
DIRE	-1	NASTY	-0.66667	DELIGHTFUL	0.3793103
DREADFUL	-1	INADEQUATE	-0.65517	ENJOYABLE	0.4137931
HORRENDOUS	-1	SUBSTANDARD	-0.58620	GOOD	0.4827586
INSUFFERABLE	-1	FINE	-0.41379	GREAT	0.5448275
INTOLERABLE	-1	MEDIOCRE	-0.41379	SUBLIME	0.5517241
USELESS	-0.95862	OK	-0.27586	LOVELY	0.5862068
UNSATISFACTORY	-0.93103	REASONABLE	-0.20689	WONDERFUL	0.6896551
UNBEARABLE	-0.91724	SUITABLE	-0.20689	SPLENDID	0.7172413
POOR	-0.89655	ACCEPTABLE	-0.13793	BRILLIANT	0.7241379
UNACCEPTABLE	-0.87586	FAIR	-0.137931	FANTASTIC	0.7379310
BAD	-0.83448	ADEQUATE	-0.068965	AMAZING	0.7931034
DISAPPOINTING	-0.82758	PERMISSIBLE	-0.068965	TREMENDOUS	0.8275862
TERRIBLE	-0.82758	ALRIGHT	-0.048275	ASTONISHING	0.8620689
AWFUL	-0.79310	MIDDLING	-0.034482	SUPERB	0.8965517
PATHETIC	-0.79310	SATISFACTORY	0	EXCELLENT	0.9310344
ROTTEN	-0.75862	NORMAL	0.0344827	MAGNIFICENT	0.9379310
UNPLEASANT	-0.75862	ORDINARY	0.0344827	MARVELLOUS	0.9655172

DISSATISFYING	-0.72413	PASSABLE	0.0344827	GLORIOUS	1
TEDIOUS	-0.69655	AVERAGE	0.1034482		
BORING	-0.68965	NICE	0.2068965		

7 - BRIGHTNESS

LIGHTLESS	-0.64	LIGHTED	0.35	GOLDEN	0.55
MOONLIT	-0.38	GLITTERING	0.35	SHINY	0.55
BURNISHED	-0.35	LUMINOUS	0.38	SPARKLING	0.55
AGLOW	-0.2	SHIMMERING	0.4	SUNNY	0.55
TWINKLING	0.02	SUNLIT	0.4	BLAZING	0.55
BURNING	0.1	ILLUMINATED	0.4	RADIANT	0.55
BEAMING	0.25	FLASHING	0.45	GLISTENING	0.55
ALIGHT	0.35	GLARING	0.45	BRIGHT	0.57
ILLUMINED	0.35	LIGHT	0.5	DAZZLING	0.6

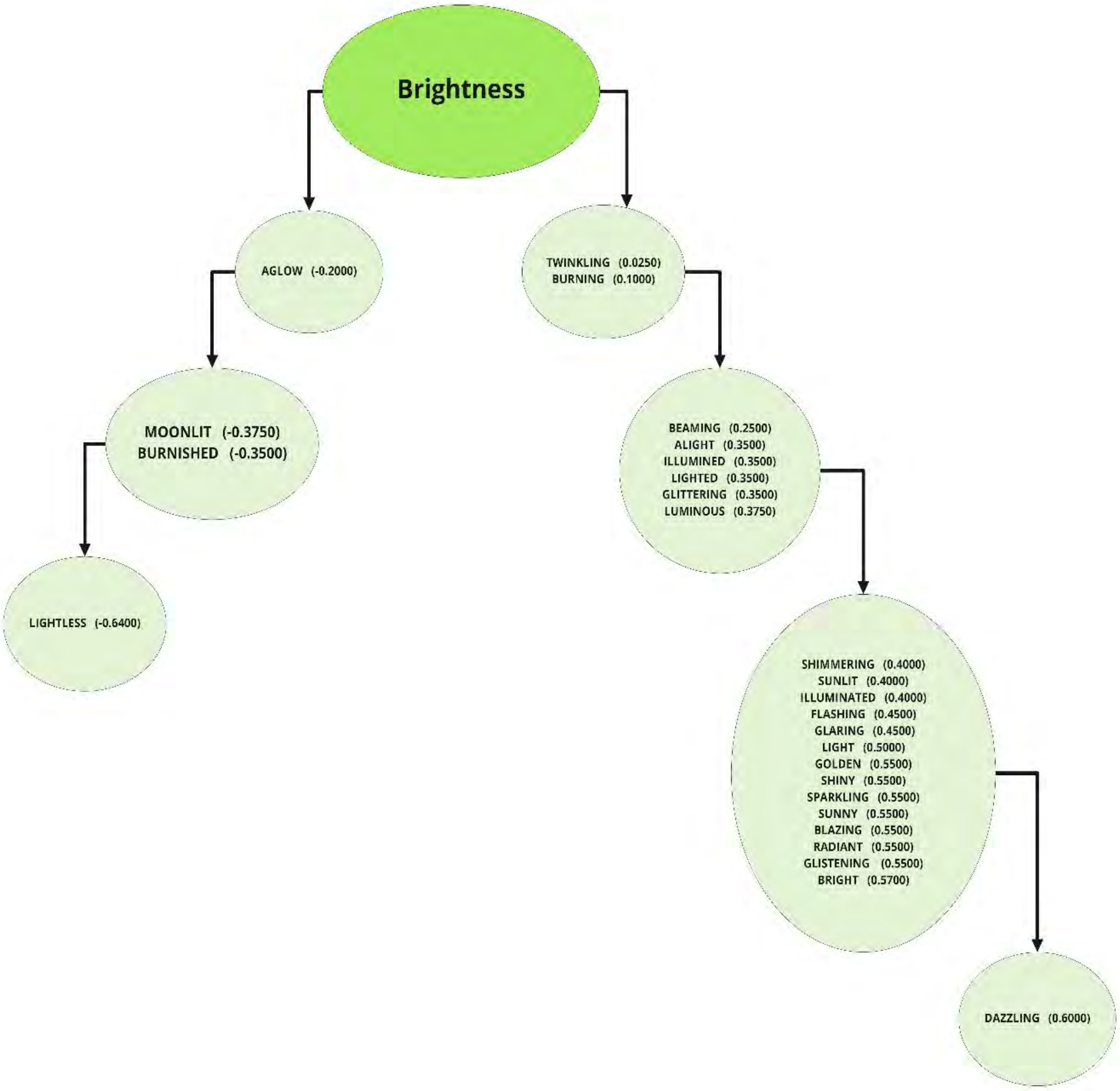
8 - SPEED

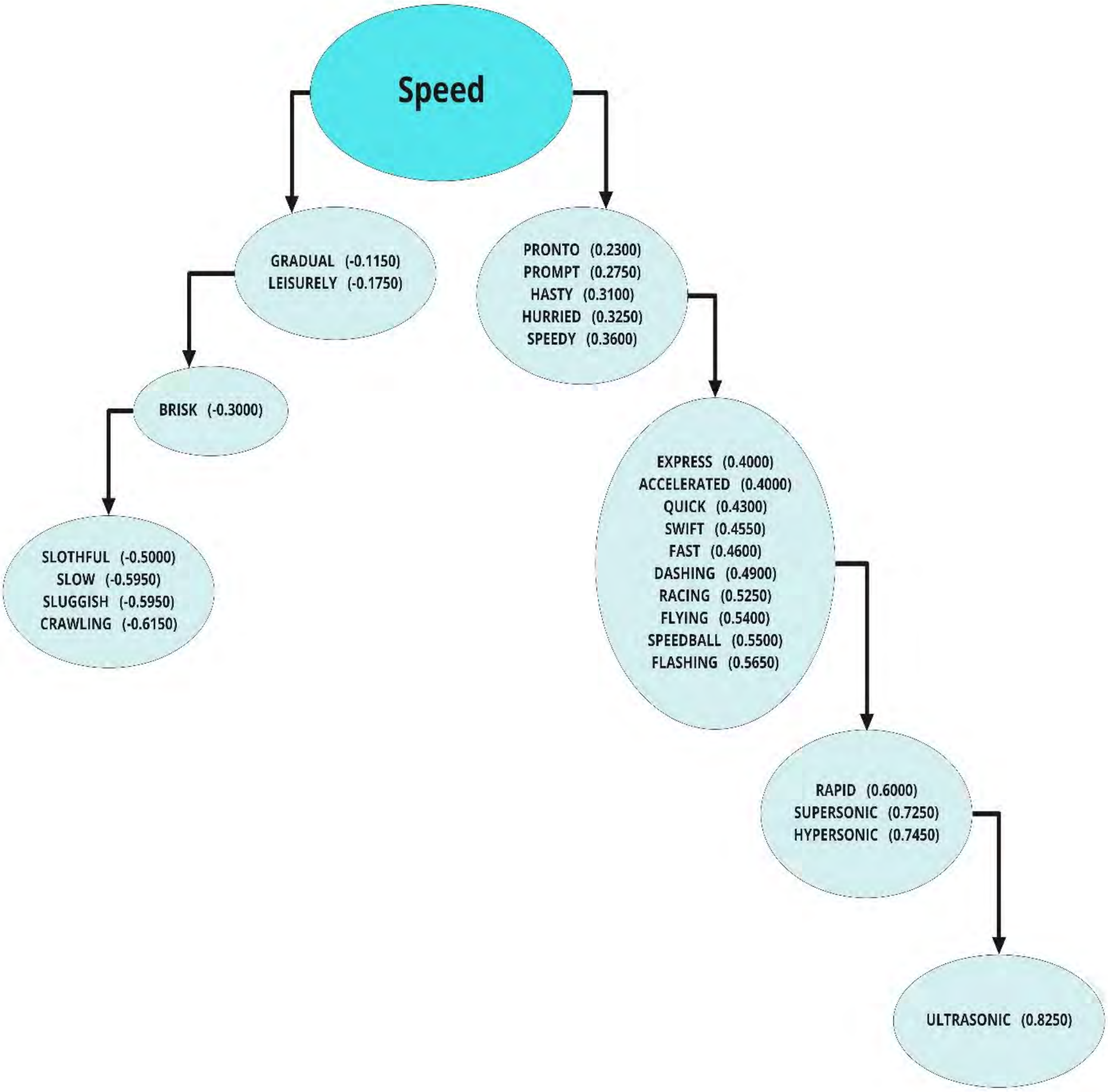
CRAWLING	-0.615	HASTY	0.31	RACING	0.525
SLUGGISH	-0.595	HURRIED	0.325	FLYING	0.54
SLOW	-0.595	SPEEDY	0.36	SPEEDBALL	0.55
SLOTHFUL	-0.5	EXPRESS	0.4	FLASHING	0.565
BRISK	-0.3	ACCELERATED	0.4	RAPID	0.6
LEISURELY	-0.175	QUICK	0.43	SUPERSONIC	0.725
GRADUAL	-0.115	SWIFT	0.455	HYPERSONIC	0.745
PRONTO	0.23	FAST	0.46	ULTRASONIC	0.825
PROMPT	0.275	DASHING	0.49		

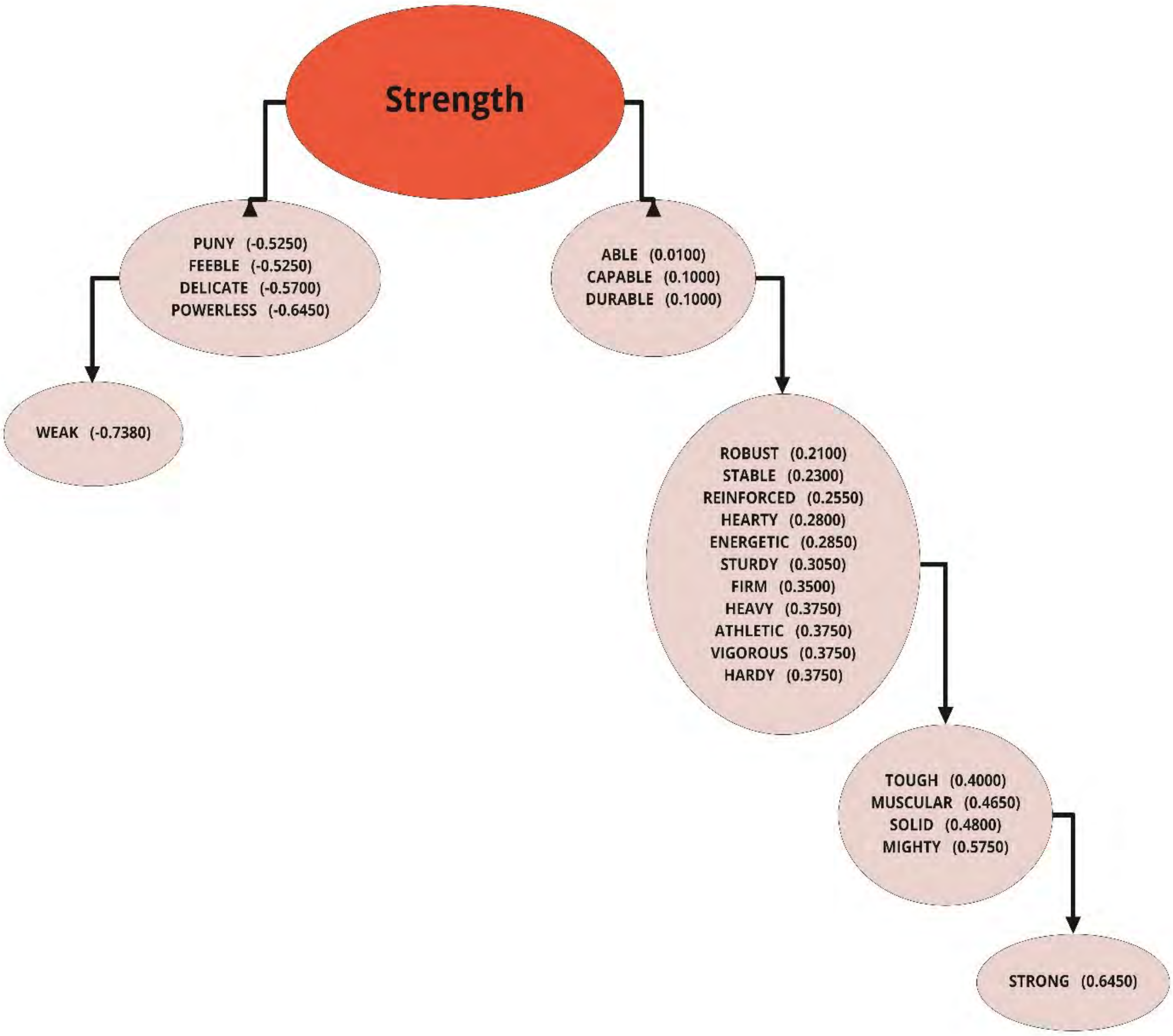
9 - STRENGTH

WEAK	-0.738	ROBUST	0.21	ATHLETIC	0.375
POWERLESS	-0.645	STABLE	0.23	VIGOROUS	0.375
DELICATE	-0.57	REINFORCED	0.255	HARDY	0.375
FEEBLE	-0.525	HEARTY	0.28	TOUGH	0.4
PUNY	-0.525	ENERGETIC	0.285	MUSCULAR	0.465
ABLE	0.01	STURDY	0.305	SOLID	0.48
CAPABLE	0.1	FIRM	0.35	MIGHTY	0.575
DURABLE	0.1	HEAVY	0.375	STRONG	0.645

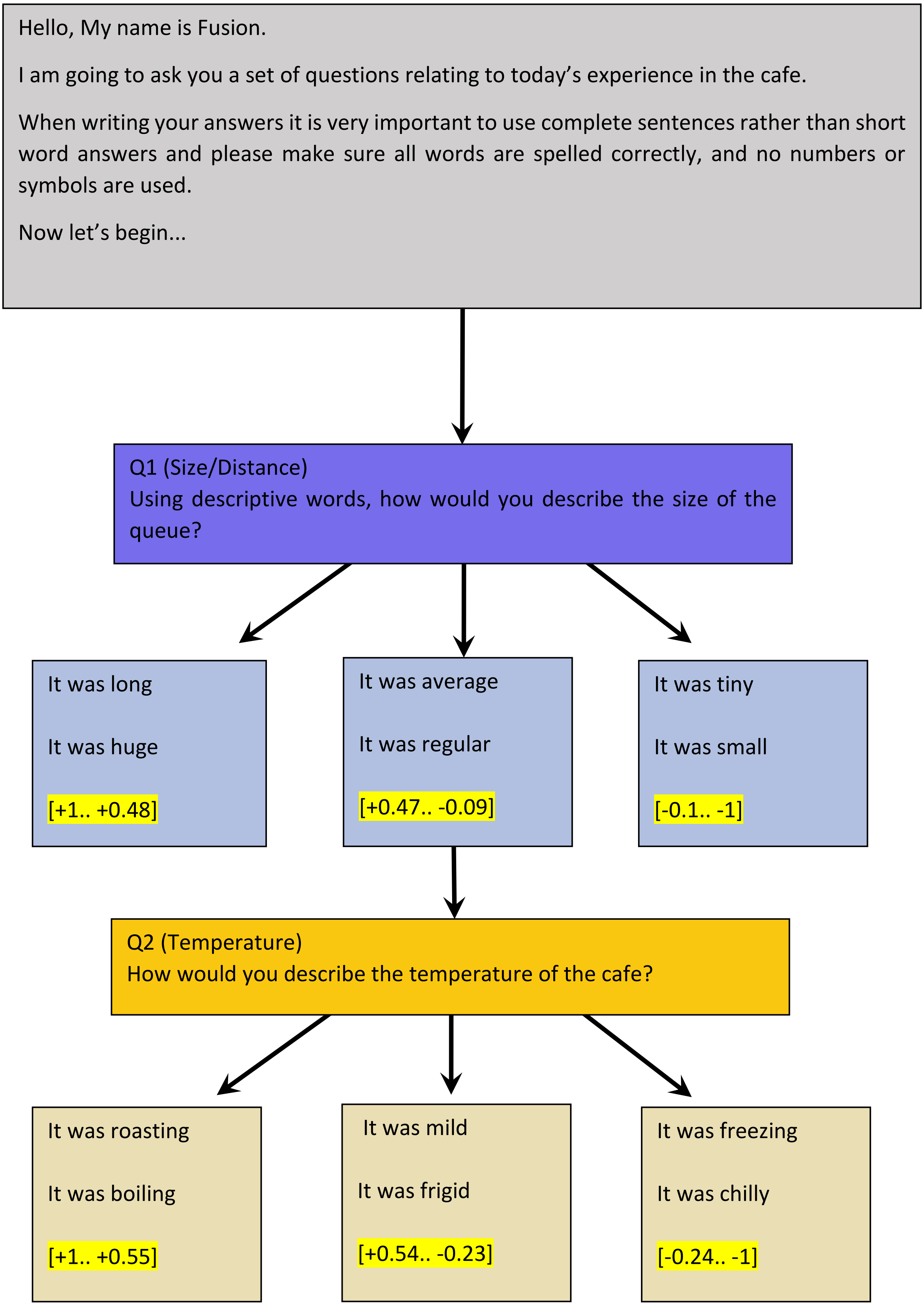
Appendix D - Ontological Structures for FUSE_2.0 (Three New Fuzzy Categories)




Speed
GRADUAL (-0.1150)
LEISURELY (-0.1750)
BRISK (-0.3000)
SLOTHFUL (-0.5000)
SLOW (-0.5950)
SLUGGISH (-0.5950)
CRAWLING (-0.6150)
PRONTO (0.2300)
PROMPT (0.2750)
HASTY (0.3100)
HURRIED (0.3250)
SPEEDY (0.3600)
EXPRESS (0.4000)
ACCELERATED (0.4000)
QUICK (0.4300)
SWIFT (0.4550)
FAST (0.4600)
DASHING (0.4900)
RACING (0.5250)
FLYING (0.5400)
SPEEDBALL (0.5500)
FLASHING (0.5650)
RAPID (0.6000)
SUPERSONIC (0.7250)
HYPERSONIC (0.7450)
ULTRASONIC (0.8250)


Strength
PUNY (-0.5250)
FEEBLE (-0.5250)
DELICATE (-0.5700)
POWERLESS (-0.6450)
WEAK (-0.7380)
ABLE (0.0100)
CAPABLE (0.1000)
DURABLE (0.1000)
ROBUST (0.2100)
STABLE (0.2300)
REINFORCED (0.2550)
HEARTY (0.2800)
ENERGETIC (0.2850)
STURDY (0.3050)
FIRM (0.3500)
HEAVY (0.3750)
ATHLETIC (0.3750)
VIGOROUS (0.3750)
HARDY (0.3750)
TOUGH (0.4000)
MUSCULAR (0.4650)
SOLID (0.4800)
MIGHTY (0.5750)
STRONG (0.6450)

Appendix E - Thresholds for FUSE_2.0 Fuzzy Categories Used in FUSION_V1



Hello, My name is Fusion.

I am going to ask you a set of questions relating to today’s experience in the cafe.

When writing your answers it is very important to use complete sentences rather than short word answers and please make sure all words are spelled correctly, and no numbers or symbols are used.

Now let’s begin...

Q1 (Size/Distance)
Using descriptive words, how would you describe the size of the queue?

It was long

It was huge

[+1.. +0.48]

It was average

It was regular

[+0.47.. -0.09]

It was tiny

It was small

[-0.1.. -1]

Q2 (Temperature)
How would you describe the temperature of the cafe?

It was roasting

It was boiling

[+1.. +0.55]

It was mild

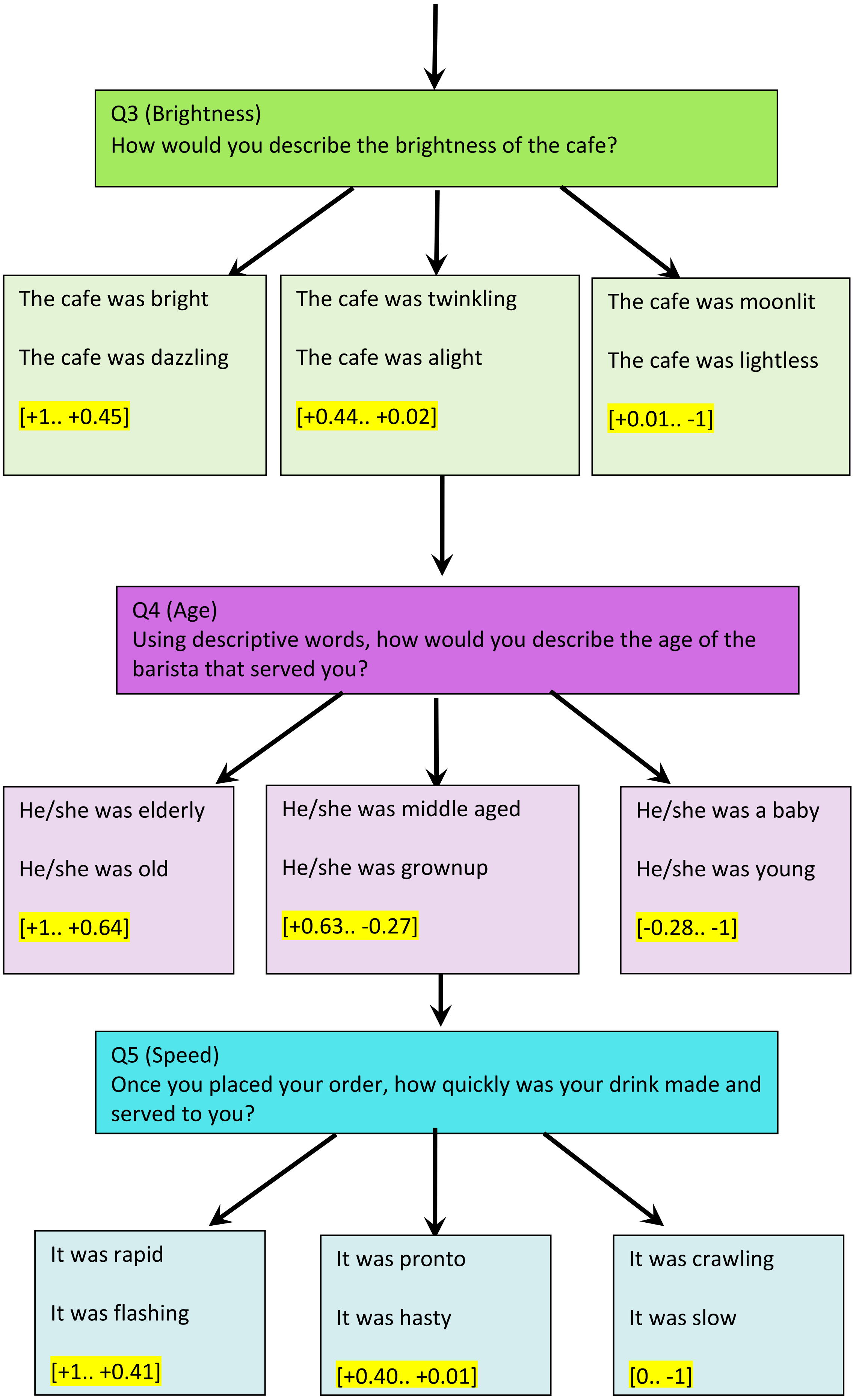
It was frigid

[+0.54.. -0.23]

It was freezing

It was chilly

[-0.24.. -1]



Q3 (Brightness)
How would you describe the brightness of the cafe?

The cafe was bright

The cafe was dazzling

[+1.. +0.45]

The cafe was twinkling

The cafe was alight

[+0.44.. +0.02]

The cafe was moonlit

The cafe was lightless

[+0.01.. -1]

Q4 (Age)
Using descriptive words, how would you describe the age of the barista that served you?

He/she was elderly

He/she was old

[+1.. +0.64]

He/she was middle aged

He/she was grownup

[+0.63.. -0.27]

He/she was a baby

He/she was young

[-0.28.. -1]

Q5 (Speed)
Once you placed your order, how quickly was your drink made and served to you?

It was rapid

It was flashing

[+1.. +0.41]

It was pronto

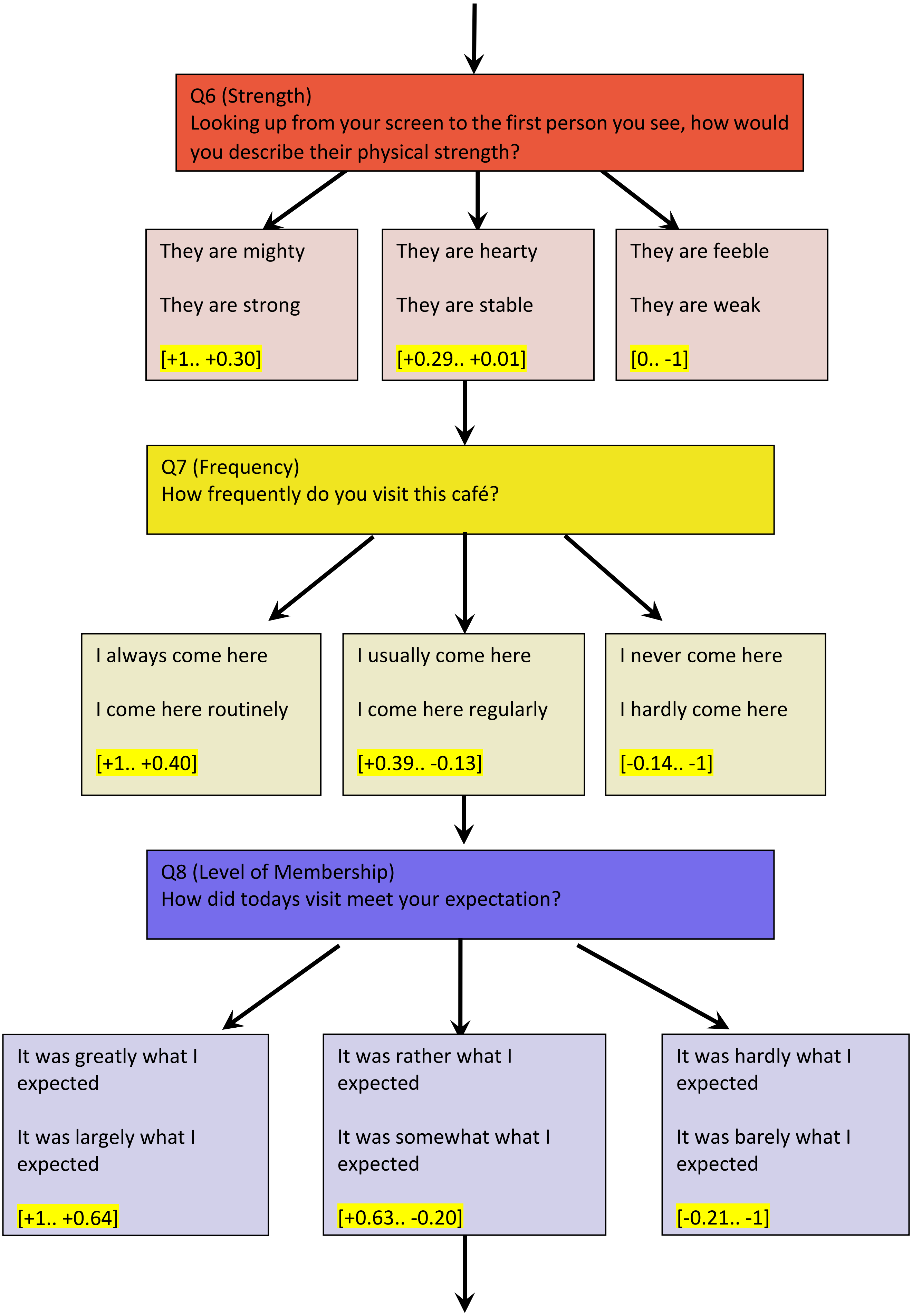
It was hasty

[+0.40.. +0.01]

It was crawling

It was slow

[0.. -1]



Q6 (Strength)
Looking up from your screen to the first person you see, how would you describe their physical strength?

They are mighty

They are strong

[+1.. +0.30]

They are hearty

They are stable

[+0.29.. +0.01]

They are feeble

They are weak

[0.. -1]

Q7 (Frequency)
How frequently do you visit this café?

I always come here

I come here routinely

[+1.. +0.40]

I usually come here

I come here regularly

[+0.39.. -0.13]

I never come here

I hardly come here

[-0.14.. -1]

Q8 (Level of Membership)
How did todays visit meet your expectation?

It was greatly what I expected

It was largely what I expected

[+1.. +0.64]

It was rather what I expected

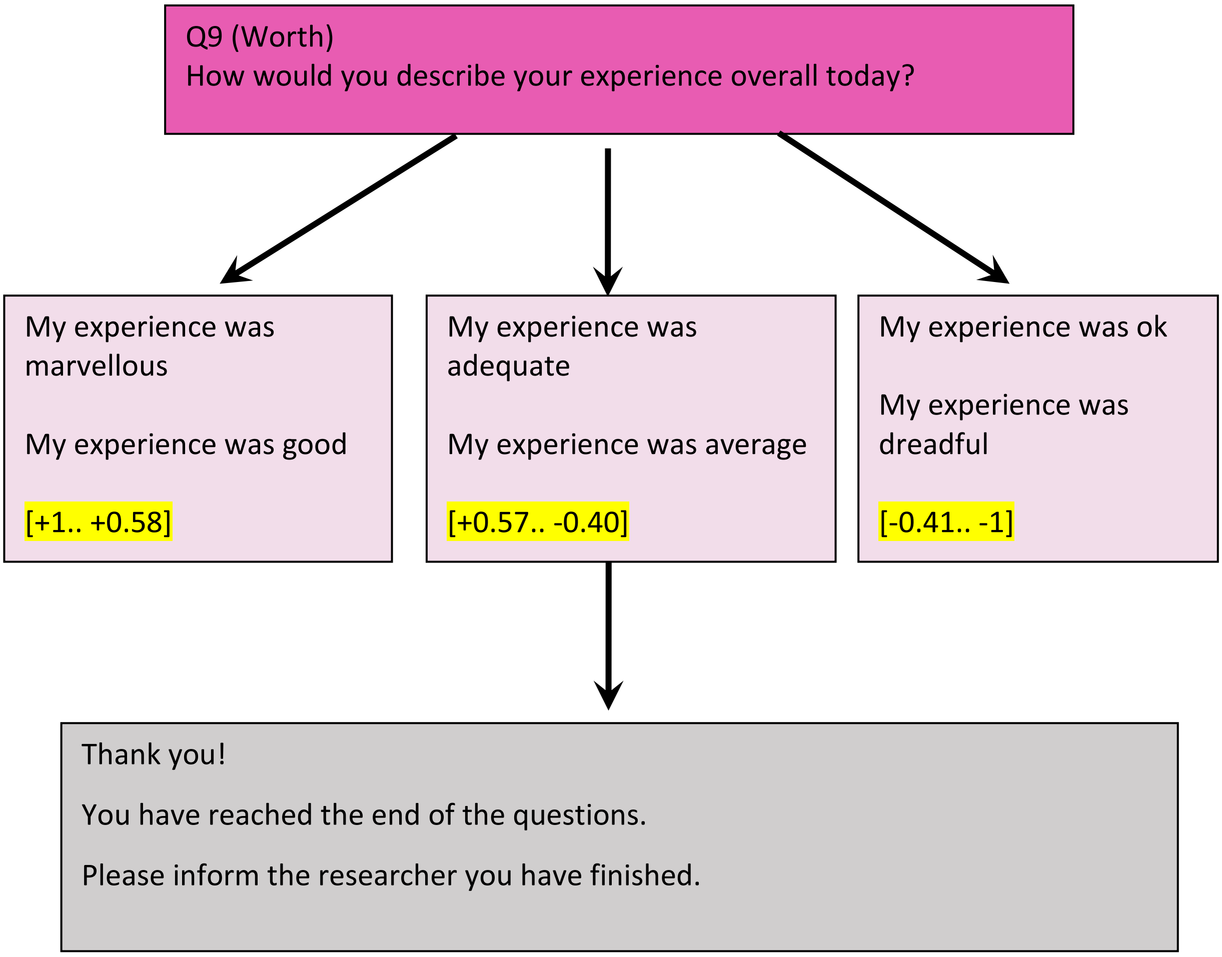
It was somewhat what I expected

[+0.63.. -0.20]

It was hardly what I expected

It was barely what I expected

[-0.21.. -1]



Q9 (Worth)
How would you describe your experience overall today?

My experience was marvellous

My experience was good

[+1.. +0.58]

My experience was adequate

My experience was average

[+0.57.. -0.40]

My experience was ok

My experience was dreadful

[-0.41.. -1]

Thank you!

You have reached the end of the questions.

Please inform the researcher you have finished.

Appendix F - Participant Information Sheet for FUSION_V1

Version: [Participant Information Sheet V2.0 July 2019]
Date: 01/07/2019
Name: Naomi Adel
Course: PhD
Title of Project: Fuzzy natural language similarity measures through CWW
Department: Science and Engineering
Building: School of Computing, Maths and Digital Technology,
Faculty of Science and Engineering, Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester, M1 5GD
Tel: 0161 247 6790

Thank you for volunteering to take part in this scientific study, in the field of semantic sentence similarity. You may still withdraw before starting the task or at any point while doing it. Before you proceed, you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully. Ask questions if anything you read is not clear or you would like more information. Take time to decide whether or not to take part. Should you have any questions please do not hesitate to contact me using the information below:

Name: Naomi Adel
Email: N.Adel@mmu.ac.uk
Telephone: 0161 247 6790

What is the purpose of the study?

The purpose of this study is to collect your response after an assigned simple task. The task first involves purchasing a hot or cold drink at a chosen café. You will then enjoy your drink sat inside the café and once finished, you will be asked to complete a short Q&A with a conversational agent who will ask about your visit to the café. You will then complete a usability questionnaire. The study should last no more than 30 minutes.

Why have I been invited?

Because you are a Native English speaker.

Do I have to take part?

The study will be explained to you after which you are asked to sign a consent form to show you agreed to take part. You are free to withdraw at any time, without giving a reason.

What will happen to me if I take part?

If you agree to take part, you will be asked by the researcher what hot drink you would like. You will be given money to pay for the drink. You will be asked to purchase a hot drink at a chosen café. You will purchase your chosen drink, sit down in the café and observe your surroundings. Once you have finished your drink you join the researcher, who will ask you to complete the relevant consent forms and then you will be asked to answer a set of questions using a computerised conversational agent (CA). The questions will relate to your visit of the *cafe and experience. Finally, you will be asked to complete a usability questionnaire rating your overall experience with the CA.*

No personal data is stored electronically. The only personal information stored will be your name and signature on the paper based consent form. Paper based consent forms will be stored independently by the researcher in a locked cabinet in a locked office in MMU

What are the possible disadvantages and risks of taking part?
None

What are the possible benefits of taking part?
Help contribute towards computer systems that will help understand the English language and enjoy a free drink.

What if there is a problem?
If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions.

Will my taking part in the study be kept confidential?
Yes. When you sign the consent form you will provide your name which is kept on a hard copy paper based consent form. You will also fill a background information sheet which will hold your age range and qualification level. Both these forms will be stored in a locked cabinet at Manchester Metropolitan University. You will then be allocated a participant number, but this will not be linked to your consent form. No personal information is recorded from participants electronically, and responses cannot be traced back to you. All information collected is private and confidential and solely for the purpose of this study.

What will happen if I don't carry on with the study?
If you withdraw from the study all the information and data collected from you, to date, will be destroyed.

What will happen to the results of the research study?
Your responses to the CA Q&A and the information you provide in the background information sheet will be analysed by the researchers and used to develop new algorithms that can determine the similarity of English language phrases. Analysis of results may be used in academic publications.

What if I have concerns about the study?
If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions:

Researchers:
[**Naomi Adel** - N.Adel@mmu.ac.uk]
[**Dr Keeley Crockett** - K.Crockett@mmu.ac.uk]

What if I have a complaint about the study?

If you wish to make a complaint about this study, then please contact:
The Research Ethics and Governance Team at Manchester Metropolitan University (ethics@mmu.ac.uk, 0161 247 2853)

Further information and contact details:
Name of researcher: *Naomi Adel*
Telephone: *0161 247 6790*
Email: *N.Adel@mmu.ac.uk*
School of Computing, Maths and Digital Technology
Faculty of Science and Engineering, Manchester Metropolitan University
John Dalton Building, Chester Street, Manchester, M1 5GD

Appendix G - Background Information Sheet for FUSION_V1

Version: [Background Information Sheet V2.0 July 2019]
Date: 01/07/2019
Name: Naomi Adel
Course: PhD
Title of Project: Fuzzy natural language similarity measures through CWW
Department: Science and Engineering
Building: School of Computing, Maths and Digital Technology,
Faculty of Science and Engineering, Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester, M1 5GD
Tel: 0161 247 6790

Participant No: ____________________

In order to help with the results, I need to collect some information about you. All information collected is private and confidential and solely for the purpose of analysing the results.

What age range do you fall under? (*please tick one*)

- 18 – 29 □
- 30 – 41 □
- 42 – 53 □
- 54 and above □

What is your highest qualification to date? (*please tick one*)

- Below GCSE (or equivalent) □
- GCSE (or equivalent) □
- A-Levels (or equivalent) □
- Undergraduate (or equivalent) □
- Postgraduate (or equivalent) □
- PhD □
- Post-Doctoral □
- Other □

Appendix H - Consent Form for FUSION_V1

Version: [Consent Form V2.0 July 2019]
Date: 01/07/2019
Name: Naomi Adel
Course: PhD
Title of Project: Fuzzy natural language similarity measures through CWW
Department: Science and Engineering
Building: School of Computing, Maths and Digital Technology,
Faculty of Science and Engineering, Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester, M1 5GD

Manchester
Metropolitan
University

Participant No: ____________________

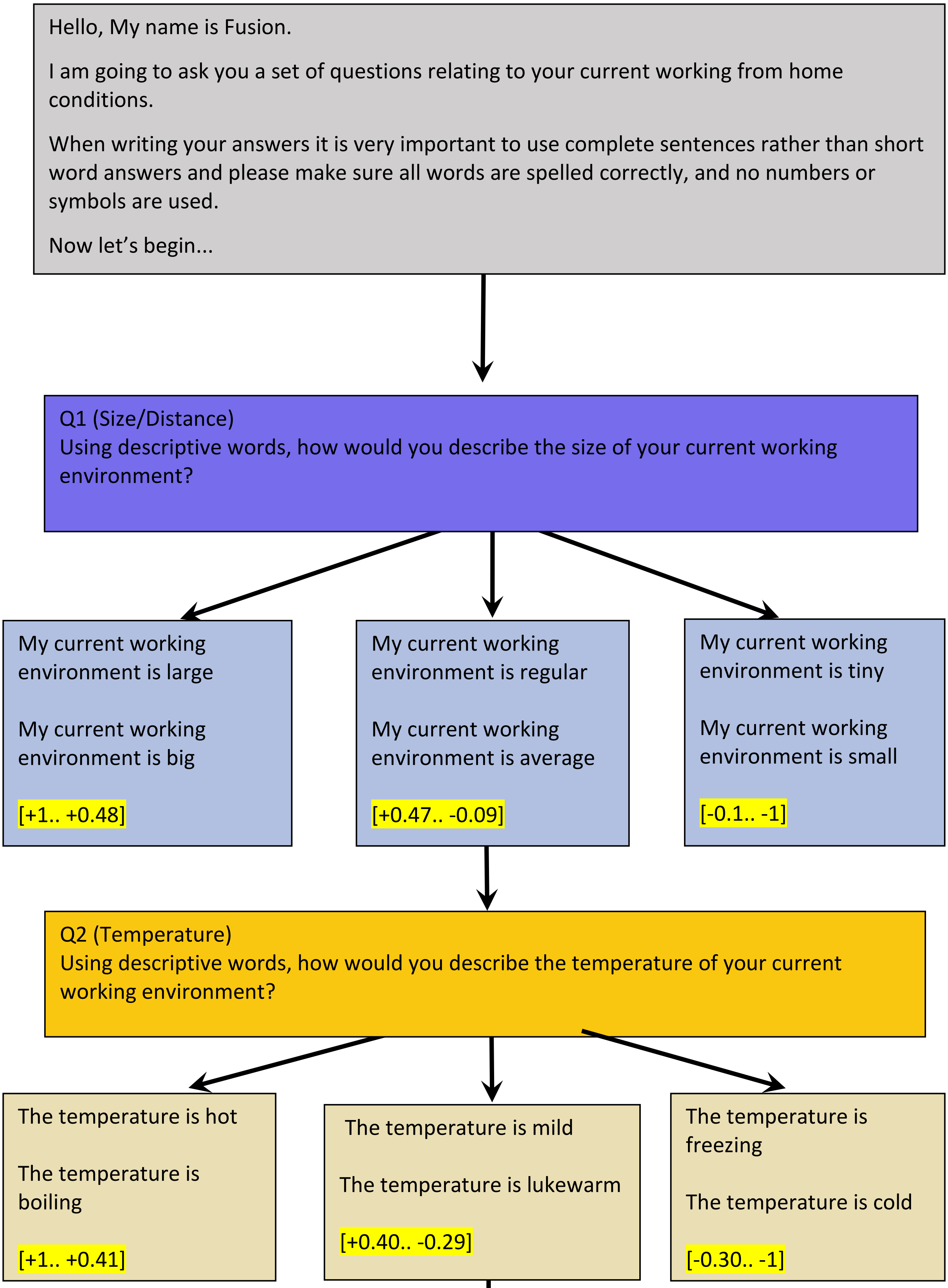
Please initial box

1. I confirm that I am eligible to participate in this study □
2. I agree that my responses given through the conversational agent may be quoted (anonymously) in publications, reports and other research outputs. □
3. I confirm that I have read and understood the information sheet dated [____/____/______] for the above project and have had the opportunity to ask questions about the experiment. □
4. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason to the named Researcher. □
5. I understand that my identifiable data which exists on the consent form only, will be destroyed no later than 7 years from the start date of this study. □

___________________________	_________________	____________________
Participant Name [printed]	*Date*	*Signature*
___________________________	_________________	____________________
Researcher Name [printed]	*Date*	*Signature*

Once this has been signed, you will receive a copy of your signed and dated consent form and participant information sheet.

Appendix I - Thresholds for FUSE_4.0 Fuzzy Categories Used in FUSION_V2 Set 1



Hello, My name is Fusion.

I am going to ask you a set of questions relating to your current working from home conditions.

When writing your answers it is very important to use complete sentences rather than short word answers and please make sure all words are spelled correctly, and no numbers or symbols are used.

Now let's begin...

Q1 (Size/Distance)
Using descriptive words, how would you describe the size of your current working environment?

My current working environment is large

My current working environment is big

[+1.. +0.48]

My current working environment is regular

My current working environment is average

[+0.47.. -0.09]

My current working environment is tiny

My current working environment is small

[-0.1.. -1]

Q2 (Temperature)
Using descriptive words, how would you describe the temperature of your current working environment?

The temperature is hot

The temperature is boiling

[+1.. +0.41]

The temperature is mild

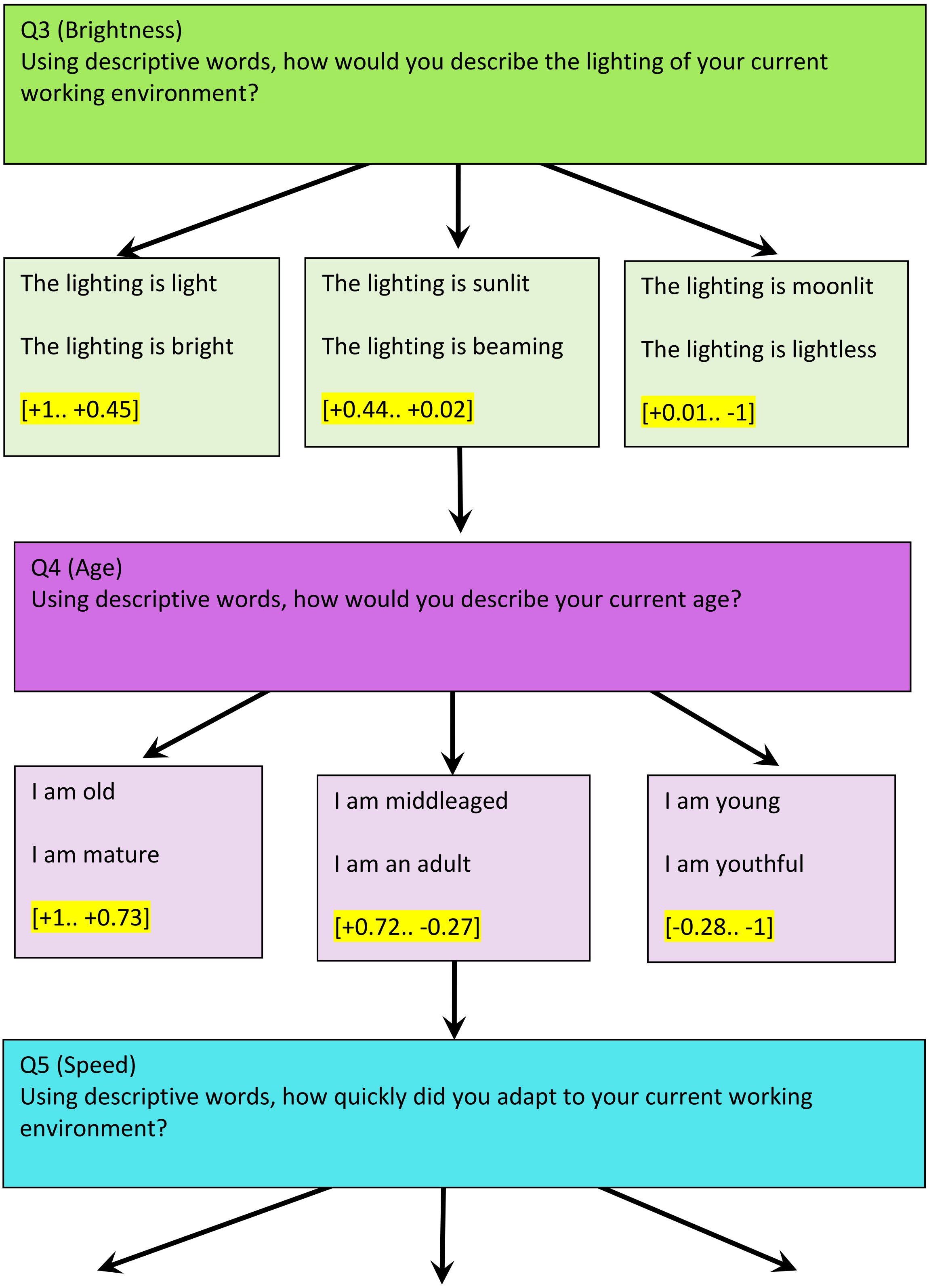
The temperature is lukewarm

[+0.40.. -0.29]

The temperature is freezing

The temperature is cold

[-0.30.. -1]



Q3 (Brightness)
Using descriptive words, how would you describe the lighting of your current working environment?

The lighting is light

The lighting is bright

[+1.. +0.45]

The lighting is sunlit

The lighting is beaming

[+0.44.. +0.02]

The lighting is moonlit

The lighting is lightless

[+0.01.. -1]

Q4 (Age)
Using descriptive words, how would you describe your current age?

I am old

I am mature

[+1.. +0.73]

I am middleaged

I am an adult

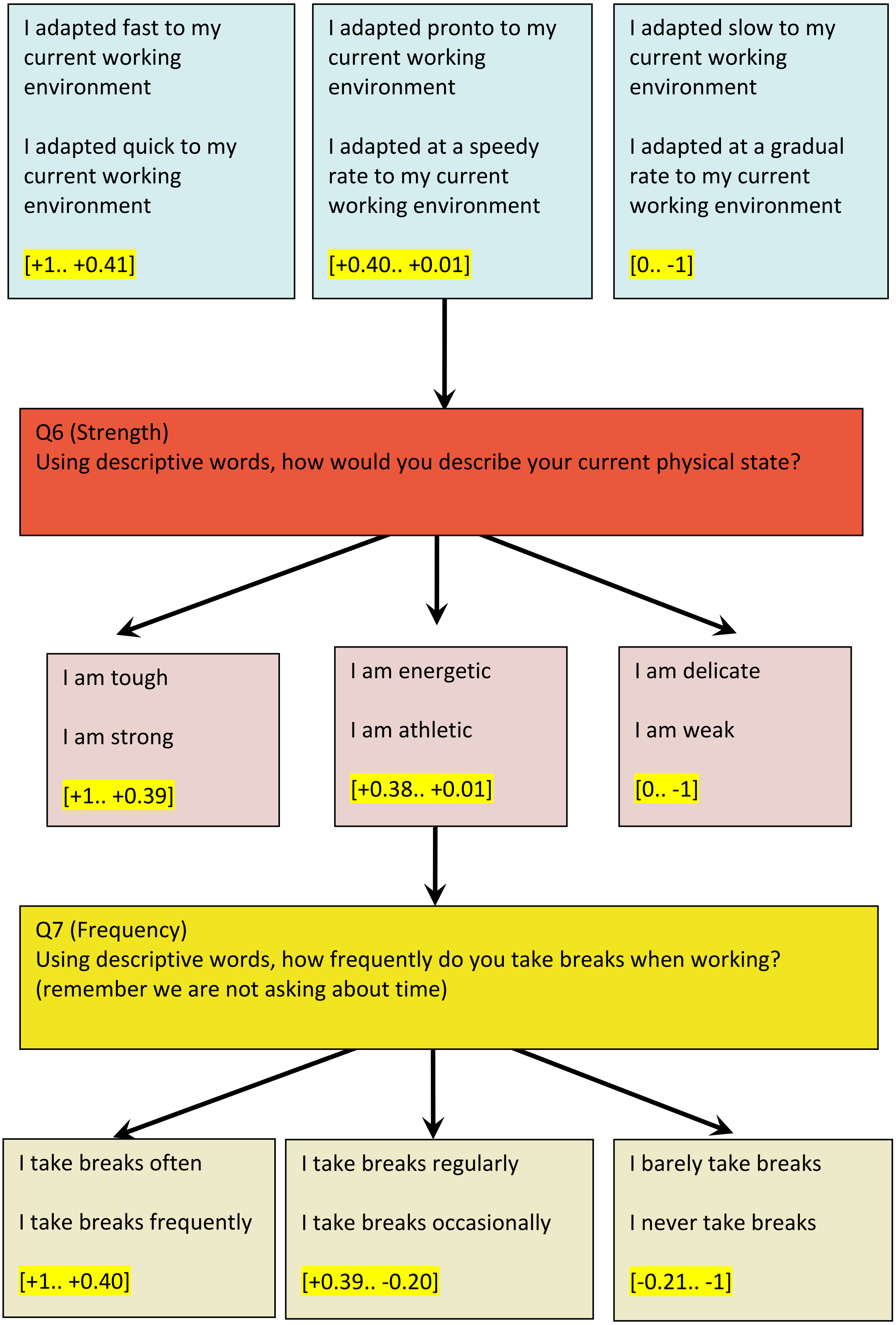
[+0.72.. -0.27]

I am young

I am youthful

[-0.28.. -1]

Q5 (Speed)
Using descriptive words, how quickly did you adapt to your current working environment?



I adapted fast to my current working environment

I adapted quick to my current working environment

[+1.. +0.41]

I adapted pronto to my current working environment

I adapted at a speedy rate to my current working environment

[+0.40.. +0.01]

I adapted slow to my current working environment

I adapted at a gradual rate to my current working environment

[0.. -1]

Q6 (Strength)
Using descriptive words, how would you describe your current physical state?

I am tough

I am strong

[+1.. +0.39]

I am energetic

I am athletic

[+0.38.. +0.01]

I am delicate

I am weak

[0.. -1]

Q7 (Frequency)
Using descriptive words, how frequently do you take breaks when working?
(remember we are not asking about time)

I take breaks often

I take breaks frequently

[+1.. +0.40]

I take breaks regularly

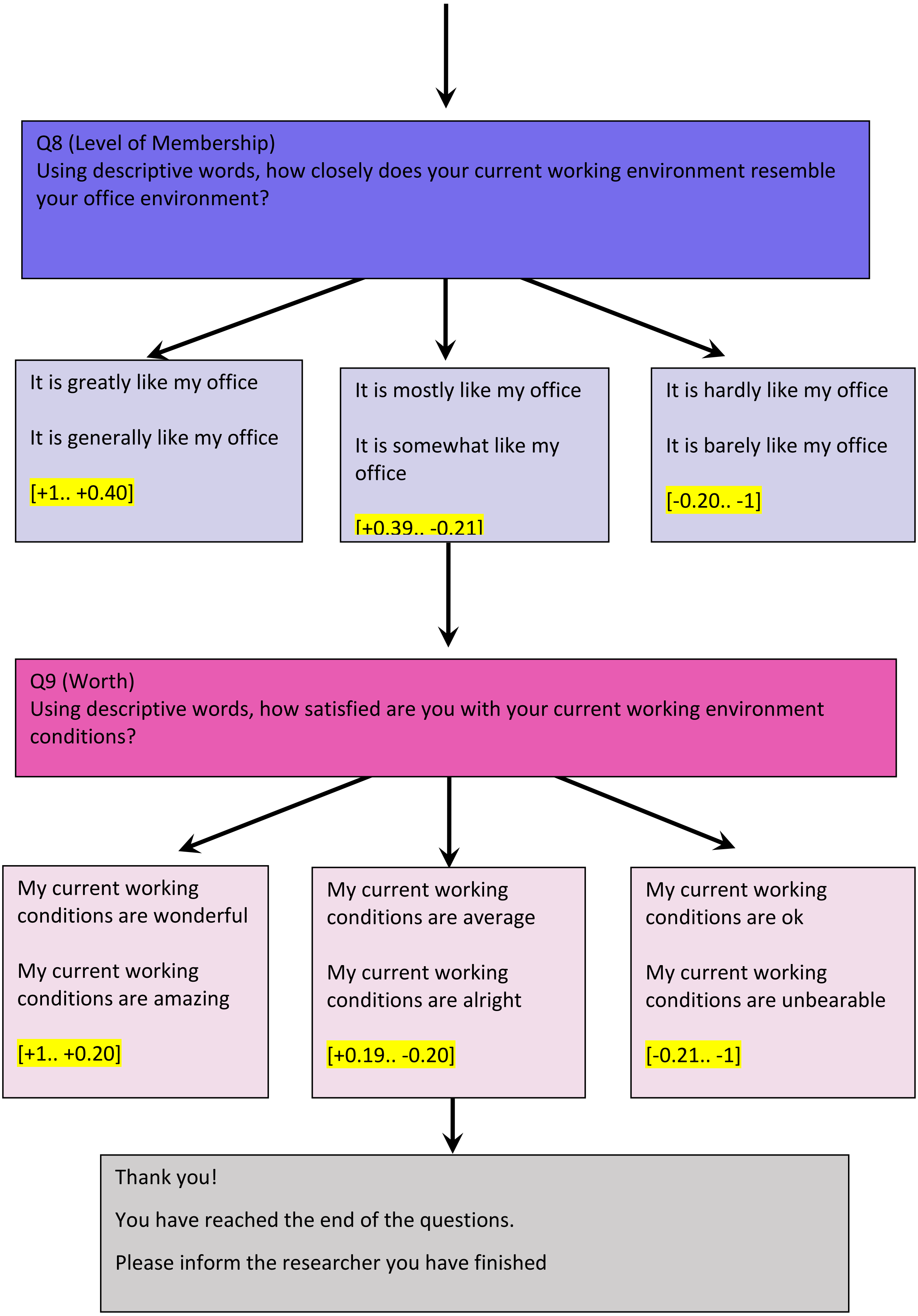
I take breaks occasionally

[+0.39.. -0.20]

I barely take breaks

I never take breaks

[-0.21.. -1]



Q8 (Level of Membership)
Using descriptive words, how closely does your current working environment resemble your office environment?

It is greatly like my office

It is generally like my office

[+1.. +0.40]

It is mostly like my office

It is somewhat like my office

[+0.39.. -0.21]

It is hardly like my office

It is barely like my office

[-0.20.. -1]

Q9 (Worth)
Using descriptive words, how satisfied are you with your current working environment conditions?

My current working conditions are wonderful

My current working conditions are amazing

[+1.. +0.20]

My current working conditions are average

My current working conditions are alright

[+0.19.. -0.20]

My current working conditions are ok

My current working conditions are unbearable

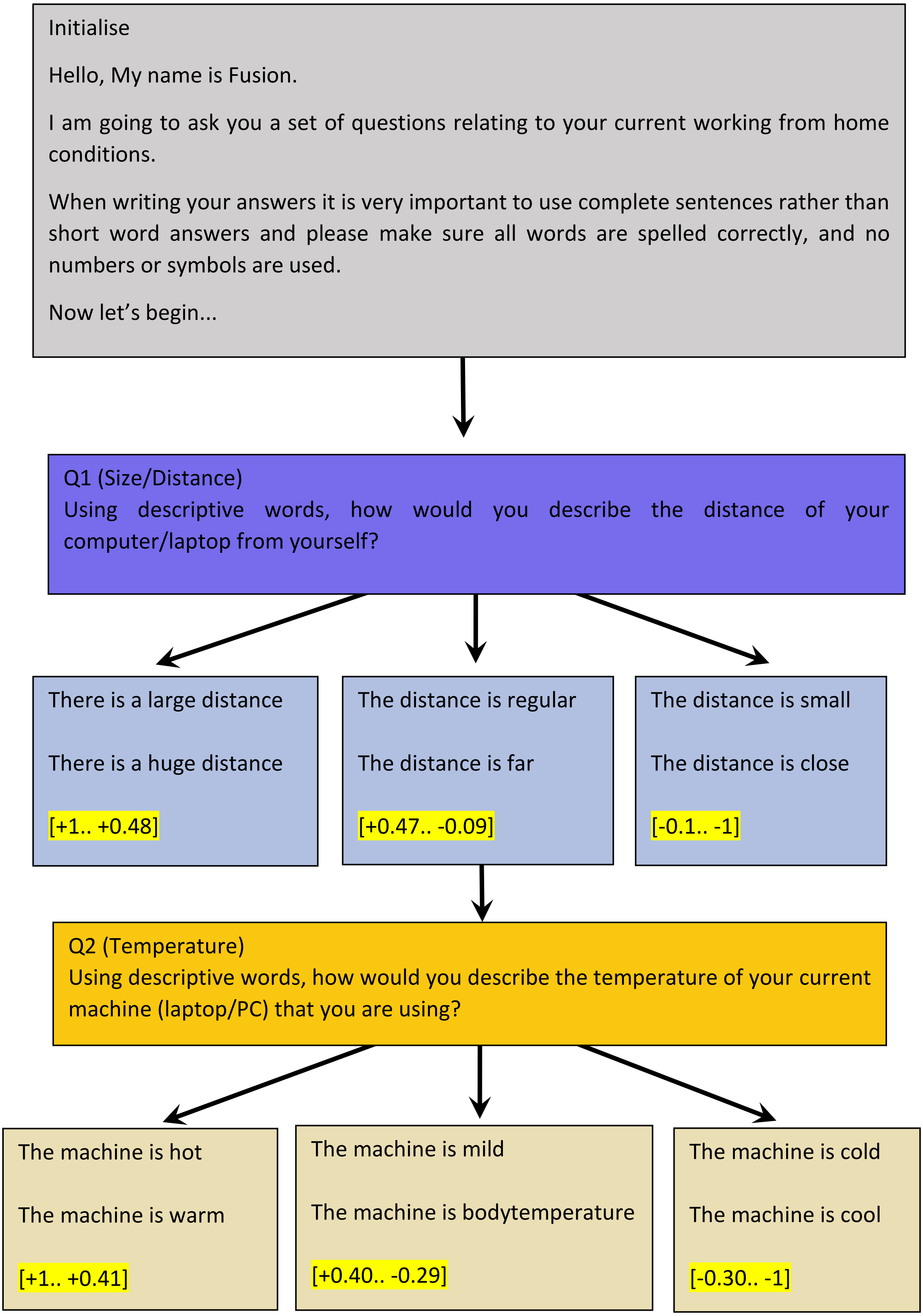
[-0.21.. -1]

Thank you!

You have reached the end of the questions.

Please inform the researcher you have finished

Appendix J - Thresholds for FUSE_4.0 Fuzzy Categories Used in FUSION_V2 Set 2



Initialise

Hello, My name is Fusion.

I am going to ask you a set of questions relating to your current working from home conditions.

When writing your answers it is very important to use complete sentences rather than short word answers and please make sure all words are spelled correctly, and no numbers or symbols are used.

Now let's begin...

Q1 (Size/Distance)
Using descriptive words, how would you describe the distance of your computer/laptop from yourself?

There is a large distance

There is a huge distance

[+1.. +0.48]

The distance is regular

The distance is far

[+0.47.. -0.09]

The distance is small

The distance is close

[-0.1.. -1]

Q2 (Temperature)
Using descriptive words, how would you describe the temperature of your current machine (laptop/PC) that you are using?

The machine is hot

The machine is warm

[+1.. +0.41]

The machine is mild

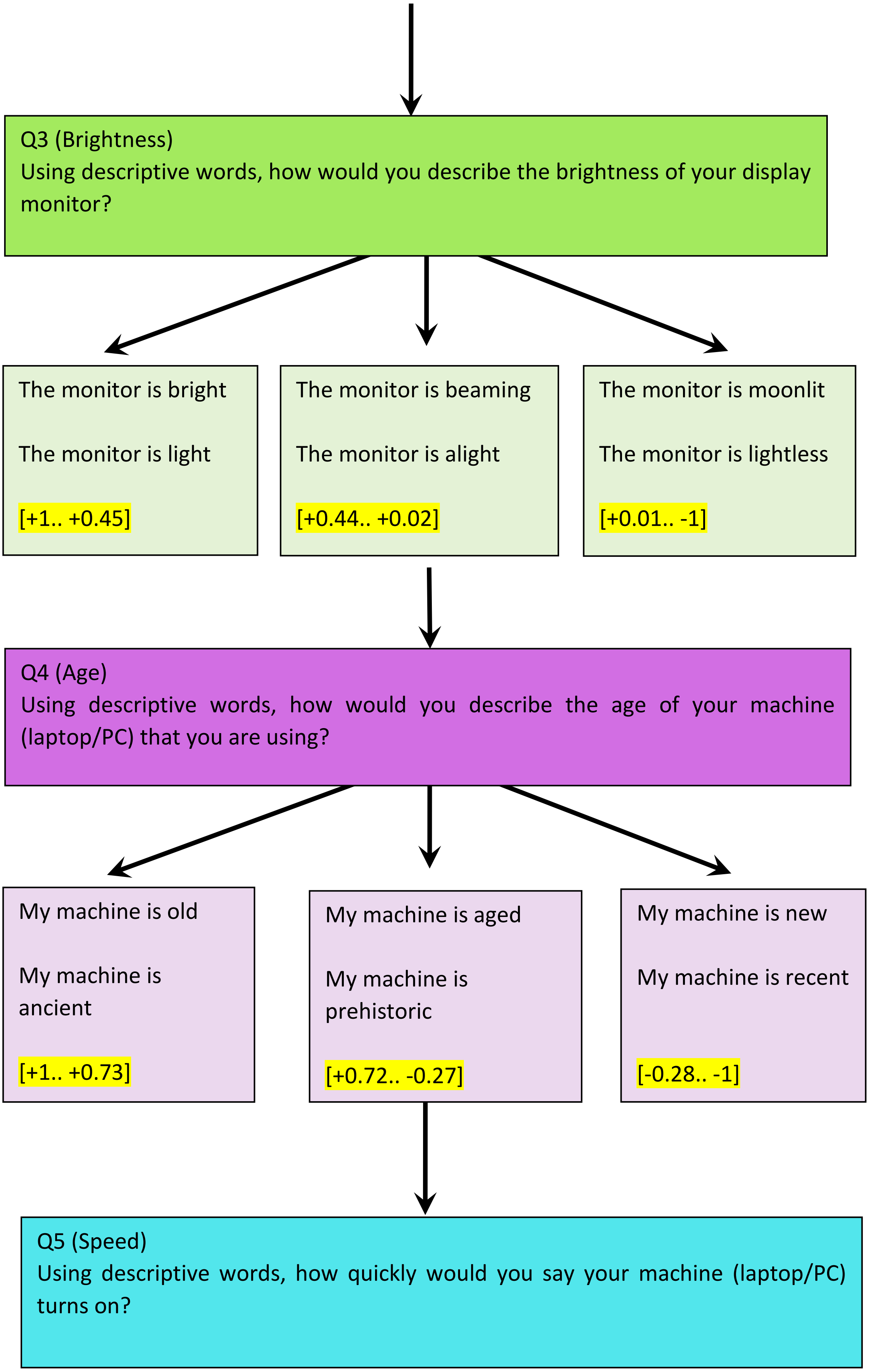
The machine is bodytemperature

[+0.40.. -0.29]

The machine is cold

The machine is cool

[-0.30.. -1]



Q3 (Brightness)
Using descriptive words, how would you describe the brightness of your display monitor?

The monitor is bright

The monitor is light

[+1.. +0.45]

The monitor is beaming

The monitor is alight

[+0.44.. +0.02]

The monitor is moonlit

The monitor is lightless

[+0.01.. -1]

Q4 (Age)
Using descriptive words, how would you describe the age of your machine (laptop/PC) that you are using?

My machine is old

My machine is ancient

[+1.. +0.73]

My machine is aged

My machine is prehistoric

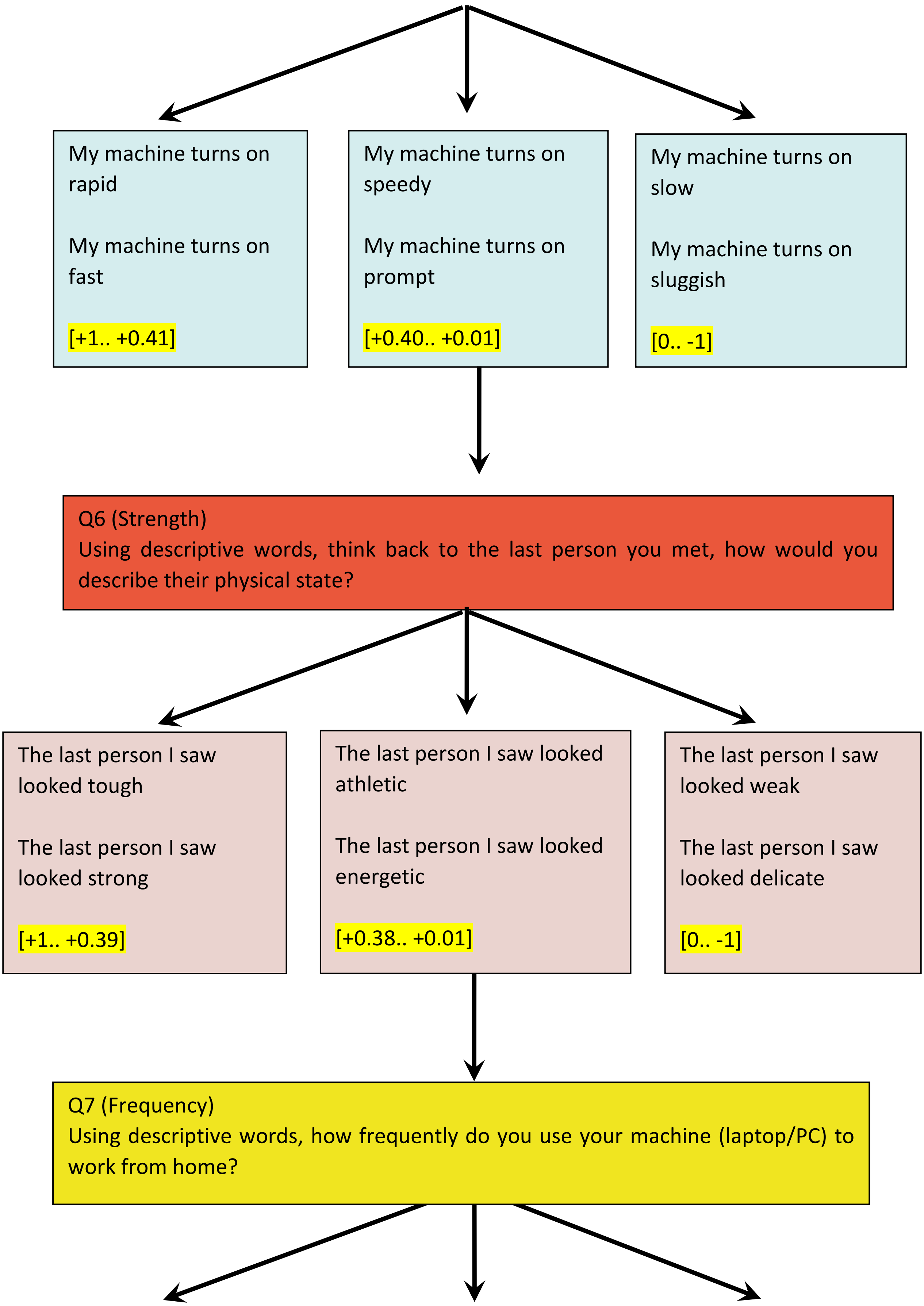
[+0.72.. -0.27]

My machine is new

My machine is recent

[-0.28.. -1]

Q5 (Speed)
Using descriptive words, how quickly would you say your machine (laptop/PC) turns on?



My machine turns on rapid

My machine turns on fast

[+1.. +0.41]

My machine turns on speedy

My machine turns on prompt

[+0.40.. +0.01]

My machine turns on slow

My machine turns on sluggish

[0.. -1]

Q6 (Strength)
Using descriptive words, think back to the last person you met, how would you describe their physical state?

The last person I saw looked tough

The last person I saw looked strong

[+1.. +0.39]

The last person I saw looked athletic

The last person I saw looked energetic

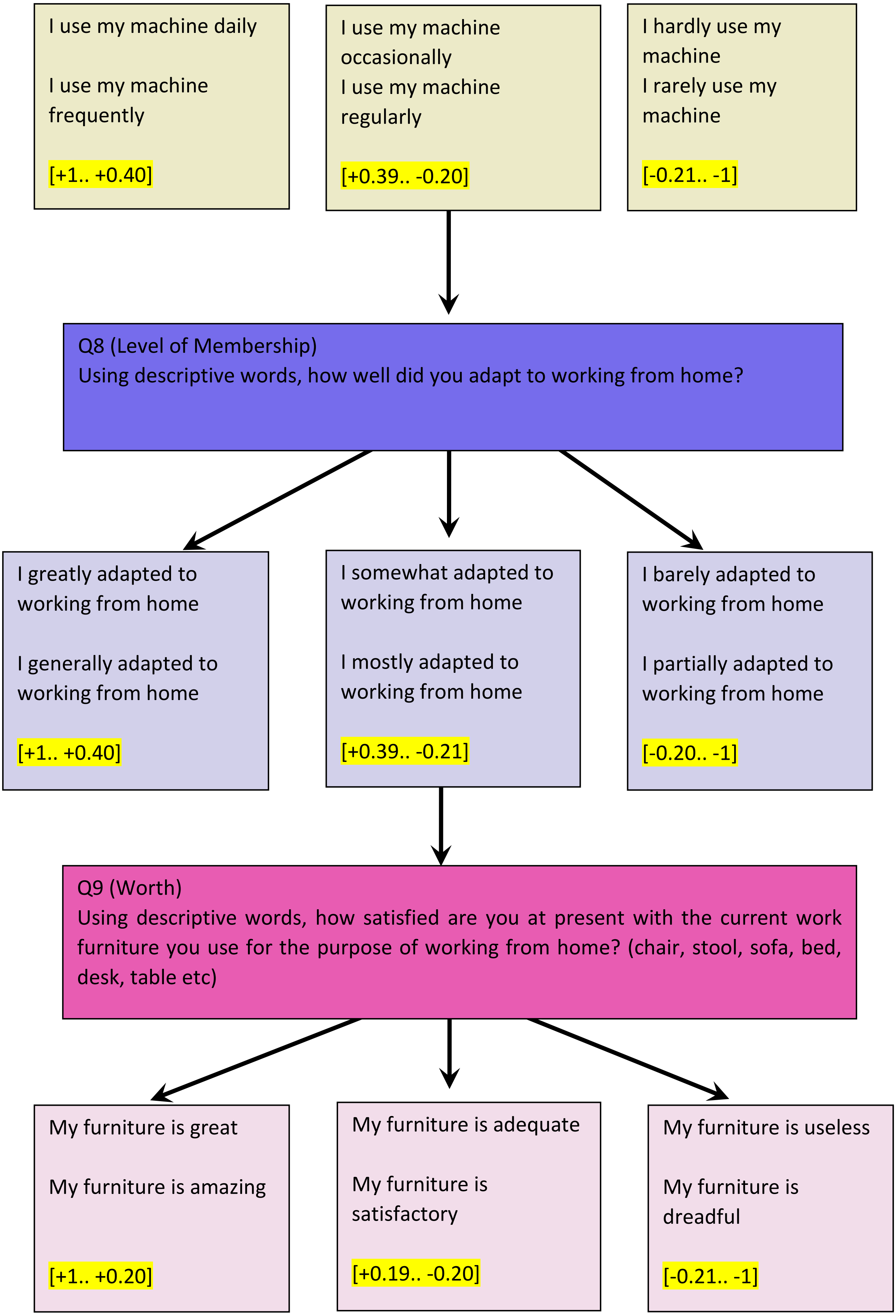
[+0.38.. +0.01]

The last person I saw looked weak

The last person I saw looked delicate

[0.. -1]

Q7 (Frequency)
Using descriptive words, how frequently do you use your machine (laptop/PC) to work from home?



I use my machine daily

I use my machine frequently

[+1.. +0.40]

I use my machine occasionally

I use my machine regularly

[+0.39.. -0.20]

I hardly use my machine

I rarely use my machine

[-0.21.. -1]

Q8 (Level of Membership)

Using descriptive words, how well did you adapt to working from home?

I greatly adapted to working from home

I generally adapted to working from home

[+1.. +0.40]

I somewhat adapted to working from home

I mostly adapted to working from home

[+0.39.. -0.21]

I barely adapted to working from home

I partially adapted to working from home

[-0.20.. -1]

Q9 (Worth)

Using descriptive words, how satisfied are you at present with the current work furniture you use for the purpose of working from home? (chair, stool, sofa, bed, desk, table etc)

My furniture is great

My furniture is amazing

[+1.. +0.20]

My furniture is adequate

My furniture is satisfactory

[+0.19.. -0.20]

My furniture is useless

My furniture is dreadful

[-0.21.. -1]

Thank you!

You have reached the end of the questions.

Please inform the researcher you have finished.

Appendix K - Participant Information Sheet for FUSION_V2

Version: [Participant Information Sheet V3.0 January 2021]
Date: 01/01/2021
Name: Naomi Adel
Course: PhD
Title of Project: Fuzzy natural language similarity measures through CWW
Department: Science and Engineering
Building: School of Computing, Maths and Digital Technology,
Faculty of Science and Engineering, Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester, M1 5GD
Tel: 0161 247 6790

Thank you for volunteering to take part in this scientific study, in the field of semantic sentence similarity. You may still withdraw before starting the task or at any point while doing it. Before you proceed, you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully. Ask questions if anything you read is not clear or you would like more information. Take time to decide whether or not to take part. Should you have any questions please do not hesitate to contact me using the information below:

Name: Naomi Adel
Email: N.Adel@mmu.ac.uk
Telephone: 0161 247 6790

What is the purpose of the study?

The purpose of this study is to collect your response based on your current working from home conditions. You will be asked to complete a short Q&A with a conversational agent who will ask you about your current working from home conditions. You will then complete a short usability questionnaire. The study should last no more than 30 minutes.

Why have I been invited?

Because you are a Native English speaker.

Do I have to take part?

The study will be explained to you after which you are asked to sign a consent form to show you agreed to take part. You are free to withdraw at any time, without giving a reason.

What will happen to me if I take part?

If you agree to take part, you will be asked by the researcher to complete the relevant consent forms and then you will be asked to answer a set of questions using a computerised conversational agent (CA). Due to current Covid conditions, you will complete this task via Microsoft Teams, where the researcher will grant you control of the machine and you will be able to type your answers when asked questions by the CA. The questions will relate to your current working from home conditions. Finally, you will be asked to complete a short usability questionnaire rating your overall experience with the CA.

No personal data is stored electronically. The only personal information stored will be your name and signature on the consent form which you sign digitally. Electronic consent forms will be stored by the researcher on an encrypted MMU laptop used by the researcher only.

What are the possible disadvantages and risks of taking part?
None

What are the possible benefits of taking part?
Help contribute towards computer systems that will help understand the English language and enjoy a free drink.

What if there is a problem?
If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions.

Will my taking part in the study be kept confidential?
Yes. When you sign the consent form you will provide your name which is kept on a soft copy consent form. You will also fill a background information sheet which will hold your age range and qualification level. Both these forms will be stored by the researcher on an encrypted MMU laptop used by the researcher only. You will then be allocated a participant number, but this will not be linked to your consent form. No personal information is recorded from participants electronically, and responses cannot be traced back to you. All information collected is private and confidential and solely for the purpose of this study.

What will happen if I don't carry on with the study?
If you withdraw from the study all the information and data collected from you, to date, will be destroyed.

What will happen to the results of the research study?
Your responses to the CA Q&A and the information you provide in the background information sheet will be analysed by the researchers and used to develop new algorithms that can determine the similarity of English language phrases. Analysis of results may be used in academic publications.

What if I have concerns about the study?
If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions:

Researchers:
[**Naomi Adel** - N.Adel@mmu.ac.uk]
[**Dr Keeley Crockett** - K.Crockett@mmu.ac.uk]

What if I have a complaint about the study?

If you wish to make a complaint about this study, then please contact:
The Research Ethics and Governance Team at Manchester Metropolitan University (ethics@mmu.ac.uk, 0161 247 2853)

Further information and contact details:
Name of researcher: *Naomi Adel*
Telephone: *0161 247 6790*
Email: *N.Adel@mmu.ac.uk*
School of Computing, Maths and Digital Technology
Faculty of Science and Engineering, Manchester Metropolitan University
John Dalton Building, Chester Street, Manchester, M1 5GD

Appendix L - Background Information Sheet for FUSION_V2
Version: [Background Information Sheet V3.0 January 2021]
Date: 01/01/2021
Name: Naomi Adel
Course: PhD
Title of Project: Fuzzy natural language similarity measures through CWW
Department: Science and Engineering
Building: School of Computing, Maths and Digital Technology,
Faculty of Science and Engineering, Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester, M1 5GD
Tel: 0161 247 6790

Manchester Metropolitan University

Participant No: ____________________

In order to help with the results, I need to collect some information about you. All information collected is private and confidential and solely for the purpose of analysing the results.

What age range do you fall under? (*please tick one*)

- 18 – 29 □
- 30 – 41 □
- 42 – 53 □
- 54 and above □

What is your highest qualification to date? (*please tick one*)

- Below GCSE (or equivalent) □
- GCSE (or equivalent) □
- A-Levels (or equivalent) □
- Undergraduate (or equivalent) □
- Postgraduate (or equivalent) □
- PhD □
- Post-Doctoral □
- Other □

Appendix M - Consent Form for FUSION_V2

Version: [Consent Form V3.0 January 2021]
Date: 01/01/2021
Name: Naomi Adel
Course: PhD
Title of Project: Fuzzy natural language similarity measures through CWW
Department: Science and Engineering
Building: School of Computing, Maths and Digital Technology,
Faculty of Science and Engineering, Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester, M1 5GD

Participant No: ____________________

Please initial box

I confirm that I am eligible to participate in this study □

2. I agree that my responses given through the conversational agent may be quoted (anonymously) in publications, reports and other research outputs. □

3. I confirm that I have read and understood the information sheet dated [_01_/_01_/_2021_] for the above project and have had the opportunity to ask questions about the experiment. □

4. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason to the named Researcher. □

5. I understand that my identifiable data which exists on the consent form only, will be destroyed no later than 7 years from the start date of this study. □

__________________________ ________________ ____________________
Participant Name [printed] *Date* *Signature*

__________________________ ________________ ____________________
Researcher Name [printed] *Date* *Signature*

Once this has been signed, you will receive a copy of your signed and dated consent form and participant information sheet.

Author Publications

Fuzzy Influence in Fuzzy Semantic Similarity Measures

IEEE International Conference on Fuzzy Systems

2021

Fuzzy Influence in Fuzzy Semantic Similarity Measures

Naeemeh Adel, Keeley Crockett
Department of Computing and Mathematics
Manchester Metropolitan University,
Chester Street,
Manchester, M1 5GD.
United Kingdom
N.Adel@mmu.ac.uk;
K.Crockett@mmu.ac.uk

Joao P. Carvalho
INESC-ID
Instituto Superior Tecnico,
Universidade de Lisboa.
Portugal
joao.carvalho@inesc-id.pt

Valerie Cross
Computer Science and Software Engineering
Miami University,
Oxford, OH.
USA
crossv@miamioh.edu

2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) | 978-1-6654-4407-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/FUZZ45933.2021.9494535

***Abstract*: The field of Computing with Words has been pivotal in the development of fuzzy semantic similarity measures. Fuzzy semantic similarity measures allow the modelling of words in a given context with a tolerance for the imprecise nature of human perceptions. In this work, we look at how this imprecision can be addressed with the use of fuzzy semantic similarity measures in the field of natural language processing. A fuzzy influence factor is introduced into an existing measure known as FUSE. FUSE computes the similarity between two short texts based on weighted syntactic and semantic components in order to address the issue of comparing fuzzy words that exist in different word categories. A series of empirical experiments investigates the effect of introducing a fuzzy influence factor into FUSE across a number of short text datasets. Comparisons with other similarity measures demonstrates that the fuzzy influence factor has a positive effect in improving the correlation of machine similarity judgments with similarity judgments of humans.**

Keywords: computing with words, natural language processing, FUSE, semantic similarity

I. INTRODUCTION

Similarity measures combine semantic and syntactic features of natural language to determine a similarity measure of two short texts. Short texts are typically 25 words or less in length [1] and include structured (sentences) and unstructured (tweets) [2, 3, 4, 5]. Substantial research has been undertaken in the field of traditional semantic similarity [6], with methods typically grouped into corpus-based [7], string-based [8], knowledge-based [9], and hybrid [1]. Applications cover a wide area, including tweet similarity [3 and 4], fake news detection [10], spam email classification [11], Radicalization Detection Based [12] and determining effective shilling attack strategies in recommendation systems [8]. Traditional similarity measures did not calculate the impact of fuzzy words in the content of the short text.

Zadeh's early work on Computing with Words (CWW) looked at the "exploitation of the tolerance for imprecision" [13] through a methodology designed to bridge the gap between human natural language and logical computation and reasoning. More recent work by Mendel recommended that since "words mean different things to different people", Type-2 and Interval Type-2 fuzzy sets should be used to model their meaning [14] in order to capture word uncertainties. Within natural language processing applications, such as dialogue systems [15], Type-2 and Interval Type-2 fuzzy sets have allowed for improved understanding of how humans use words in different contexts to elicit better machine responses to human utterances. In this work, we define a fuzzy word as a word that has a subjective meaning, is often considered ambiguous, and is based on an individual's perception, within a given context and at a given time. Adopting a hybrid approach based on crisp and fuzzy ontologies and a corpus, FAST [16] was the first fuzzy similarity measure to be developed and evaluated specifically on datasets containing fuzzy words [16]. In FAST, human perception based words were modelled using Type-1 fuzzy sets. The FUSE measure tackled the issue of uncertainty of human judgement [17] by modelling fuzzy words using Interval Type-2 fuzzy sets, originally proposed by Hao and Mendel [17]. FUSE was successfully evaluated and extended to include hedge words in [18].

A weakness of FUSE was that to obtain a similarity measurement of a fuzzy word within short texts, the words had to be within the same fuzzy category in order to determine their distance within the fuzzy category ontology. The category ontologies that were used catered for synonyms of the English language and were extensive, it was found through empirical experimentation that it was not able to measure word similarity directly between words like '*hot*' in the Temperature category and '*large*' in the Size/Distance category. In this case, the word pair (*hot* and *large*) were passed to the generalised WordNet ontology to compute the word pair similarity. Effectively, the fuzziness of the words was lost and not included in the final short text similarity calculation.

The contribution of this paper is to propose an extension to the FUSE measure [19] by the inclusion of a Fuzzy Influence (FI) factor (defined in Section III) into the short text overall similarity calculation. The aim is to ensure that each fuzzy word has an impact on each text, regardless of whether it has a matched pair word in the same fuzzy ontology or not.

Typically, semantic measures usually comprise of two weighted elements; the semantic part and the syntactic part which are optimised against human ratings of similarity. The aim on a training dataset is to obtain a machine based method with the highest correlation to human ratings. In this paper we report the summarised results from a number of empirical experiments where we determine the effect and interaction of FI with the semantic and syntactic features in short texts. We examine the effects across three datasets containing fuzzy words, were human similarity ratings have been obtained. We show that the introduction of a fuzzy influence factor can have a positive effect on the overall sentence similarity of fuzzy sentence similarity measures (FSSM) leading to better

978-1-6654-4407-1/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

correlation of human ratings when using the Pearson's correlation coefficient.

This paper is organised as follows: Section II briefly describes relevant work in fuzzy semantic similarity measures. Section III describes the Fuzzy Influence factor and how the FUSE algorithm was extended. The experimental methodology along with datasets is described in Section IV. Results and analysis are presented in Section V.

II. Related Work

A. Fuzzy Semantic Similiarty Measures

Fuzzy semantic similarity measures calculate the semantic and syntactic similarity of a short text pair through combining both the syntactic and semantic features of a short text which are weighted. Fuzzy human perception based words were first incorporated into semantic similarity measures in the FAST algorithm. Words (selected through human experimentation) were first selected for 6 categories originally proposed by Zadeh, and were modelled using Type-1 fuzzy sets [16]. Whilst FAST captured the fuzziness of words in a sentence, the modelling of them was still subjective and opinion based. Since FAST, research in the field of Computing with Words, first advocated the use of Type-2 [20] and then later the use of Interval Type-2 fuzzy models in order to model first-order word uncertainties [21].

FUSE_1.0, a more recent fuzzy measure [19], models words using Mendel's Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets [17]. Utilising the same 6 categories as FAST, each category was firstly expanded with the number of fuzzy words and 32 English speaking participants were used to score the words in each category on a scale of [0-10]. The data was then cleaned [17], and the footprint of uncertainty (FOU's) for each word was determined. Fuzzy ontologies where then constructed for each category of fuzzy words before being applied in the FUSE measure. These category ontologies were used to compute the similarity of fuzzy word pairs. Non-fuzzy word pairs were passed to the Princeton WordNet – a lexical database of English words, comprising of sets of cognitive synonyms, each related to a distinct concept [22]. FUSE_1.0 was extended further (FUSE_2.0) to include 9 fuzzy categories and applied within a dialogue system [15]. These categories are Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership, Strength, Brightness, Speed. An issue with WordNet is that it is continually updated, and this can effect results generated by any short text similarity measure that uses it. Thus, FUSE versions have evolved over the years. In this paper, we incorporate the proposed Fuzzy Influence Factor into FUSE_4.0 which models words in 9 fuzzy categories and uses the December 2020 version of WordNet [15]. The full pseudo code for the FUSE_1.0 algorithm can be found in [19] and a revised version of the algorithm is currently under review.

B. Evaluation of Semantic Similiarty

Measures that compute semantic similarity of short texts usually require correlations with ratings of similarity given by humans. Over the years, a number of datasets have been published [2, 7] which have adopted methodologies designed to capture unbiased human ratings [2, 7]. Semantic similarity measure results can also be compared against other measures. In this work, we evaluate FUSE_4.0 against 3 other measures, STASIS [23], SEMILAR [24] and the commercial Dandelion API [25]. STASIS measures similarity using an ontological approach based on a taxonomy of words achieved by calculating the distance between words in an ontology, using WordNet, as well as the distance of words to their closest subsumer. SEMILAR [24] (SEMantic simILARity toolkit) utilises the word-to-word semantic similarity measures in the WordNet Similarity library [26] as well as using Latent Semantic Analysis [27]. Dandelion API is a commercial sentence similarity measure which computes the semantic and syntactic components separately [25]. One successful use of Dandelion API is in an Automated Short Answer Scoring within knowledge-based systems [28].

III. Fuzzy Influence

Currently the FUSE_1.0 algorithm calculates the semantic and syntactic similarity of a sentence pair through a weighted combination of analysis on both the syntactic and semantic elements of a short text. A weakness of the approach used in FUSE_1.0 is that it does not take into consideration sentence pairs where fuzzy words are not in the same category; for example comparing the word "*slow*" to "*normal*". While both these words do belong to fuzzy categories (*Speed* and *Worth* respectively), they do not fall in the same fuzzy category and so WordNet is used to derive their values. Several variants of FUSE have been developed; for example, FUSE_3.0 uses 9 categories of fuzzy words and the WordNet 2019 version [15].

A. Fuzzy Infuence

In this work, we propose the addition of a fuzzy influence factor (FI) within the FUSE algorithm. FI overcomes a weakness of FUSE by ensuring fuzzy words not in the same fuzzy categories but within the same sentence have a human associated impact on determining the sentence's similarity. The *FI* for a sentence pair *sn*, can be defined as:

$$FI_{sn} = \frac{1}{n-i} \tag{1}$$

where n is the number of all the words in the sentence pair sn; and $n > 0$, and i is the count of all the fuzzy words in sn. If all the words in the sentence pair are fuzzy, i.e. $n = i$, we set $FI_{sn} := 1$, and so FI_{sn} takes values between 0 and 1. FI is applied to all sentence pair calculations within FUSE, regardless of whether fuzzy words are in the same category or not. In [19], the FUSE algorithm was first proposed to calculate the overall similarity between two fuzzy utterances, U_1 and U_2. , through the weighted addition of syntactic and semantic components. In FUSE_4.0, the overall similarity of S(U1, U2) is then calculated as:

$$S(U_1, U_2) = sem\_sim * w1 + syn\_sim * w2 + FIsn * w3 \tag{2}$$

where $w1, w2, w3 \in [0..1]\ and\ \sum w1..w3 = 1$, and *sem_sim* and *syn_sim* are calculated using pairs of semantic and syntactic similarity vectors which were determined by a word similarity measure and a short joint word vector set comprising of word frequency information and word order. See [19] for full definitions.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

IV. Experimental Methodology

To investigate the relationships between the semantic, syntactic and FI components, an empirical experiment was conducted for FUSE_2.0 to see if the introduction of a fuzzy influence factor will affect the overall sentence similarity rating to give a value closer to that of the human ratings (HR). The hypothesis for this experiment is given below:

H_0 = The inclusion of a fuzzy influence factor (FI) in the calculation of the overall semantic similarity of a sentence improves the overall correlation when compared to human ratings.

A. Metrics

In each set of experiments, three metrics (semantic, syntactic and fuzzy influence factor) are used to measure the effectiveness of variants in the FI factor within FUSE_4.0. The Pearson's correlation coefficient is used to show statistical evidence for a linear relationship between two variables *x* and *y* in this work between the human ratings and those generated by FUSE_4.0 and is defined as [29]:

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\,.\sqrt{\mathrm{var}(y)}} \tag{3}$$

where r_{xy} is the correlation coefficient, *cov(x, y)* is the sample covariance of *x* and *y*; *var(x)* is the sample variance of *x*; and *var(y)* is the sample variance of *y*.

The reliability of inter-rater agreement of human ratings of short text pairs across a population is conducted using Cicchetti's approach [30] which uses the intra-class correlation coefficient (*a-value*). The following guidelines are followed for the interpretation of inter-rater agreement measure (*a-value*) [30]:

- *a-value* < 0.40 - Poor.
- 0.40 >= *a-value* <= 0.59 - Fair.
- 0.60 >= *a-value* <= 0.74 - Good.
- 0.75 >= *a-value* <= 1.00 - Excellent.

The *a-value* is important as it shows the extent to which the data, that is collected for this study, is a correct representation of the variables measured; therefore, the aim is to achieve an *excellent* rating to maximise reliability of the human ratings of the short text pairs [31, 32].

B. DataSets

In this work three datasets, FI-25, 62-SP, and the Multiple Fuzzy Word Dataset (MFWD) were used to investigate the fuzzy influence factor. Initial work was undertaken on a dataset known as FI-25 which comprised of a set of 25 test sentences (with inclusion criteria defined below) which have been randomly sampled from 3 existing datasets, STSS-131 [2], 62-SP [15] and MFWD [34]. SP1 – SP15 of the Sentence Pairs (SP) in FI-25 consisted of poor human rating correlations when run on FUSE_1.0. These poor human rating correlations imply that automated semantic similarity measurement of FUSE_1.0 were far different than that of the average human ratings. Ideally we would like similarity values derived from the measure to be as close to the human ratings as possible. The remaining 10 pairs (SP16 – SP25) gave high correlations with human ratings by FUSE_1.0. This means that the ratings were close to that given by the human raters. This dataset was created to ensure that the impact of the fuzzy influence factor was assessed against both high and low correlations. The general methodology for collecting human ratings can be found in [19]. The 62-SP dataset was specifically designed by English language experts to contain fuzzy words from all 9 categories from the FUSE fuzzy dictionary. The origin of the sentences came from a gold standard dataset STSS-131 [2] which contained 131 crisp sentence pairs. 62 random sentence pairs were extracted from this dataset and fuzzy words from each of the 9 fuzzy categories were placed in each sentence pair using English language experts to ensure the sentences were still meaningful. A constraint on the randomisation was to ensure there were an equal number of sentence pairs in the low, medium and high categories, as identified by human participants in previous published studies [2]. This meant that each sentence had at least 2 fuzzy words. The reader's age for this dataset has been calculated as 14-15 years old (Ninth to Tenth graders) using the Automatic Readability Checker [33]. Finally, the Multiple Fuzzy Word Dataset (MFWD) [34] contains 30 sentence pairs where each sentence contains more than one fuzzy word.

V. Experimental Results and Discussion

For each of the experiments in this section the following experimental methodology was followed. The semantic, syntactic and FI weights were each separately changed using increments of 0.05 between the ranges of 0 and 1. In each case one of the weights was fixed, whilst the other pair were changed to ensure the sum of all weights was always 1. At each iteration, Pearson's correlation was recorded each time to see which values gave the best results.

A. Experiment 1 - FI on FI-25

The FI factor within FUSE_4.0 was used with a range of different empirical weighting values for the semantic, syntactic and fuzzy influence factor to see which gave the sub-optimal results. Due to space, only a range of empirical values are reported in this paper. Optimal results are calculated by comparing Pearson's correlation (*r-value*) with human ratings. The higher the *r-value,* the closer the ratings to those of humans. In F1-25, the correlation was calculated for both the "bad" performing sentence pairs (NPW) (Table I) as well as the "good" performing sentence pairs (PW) (Table III), where bad and good results were generated by FUSE_1.0. Pearson's correlation of FUSE_FI is also compared with those of several earlier versions of FUSE as well as 4 other similarity algorithms that do not cater for fuzzy words: STASIS, which is a similarity measure using WordNet [23], SEMILAR [24], Dandelion API Semantic [25] and Dandelion API syntactic [25].

Table I shows the correlation findings for experiments 1.1-1.5 ran on the sentences that were not performing well under FUSE_1.0. Results from Table I show that the measures (Sem 0.5, Syn 0.2, FI 0.3) from experiment 1.5 gave the best overall correlation and the highest correlation, beating the other algorithm measures with the exception of API Syn for SP's that did not perform well originally with FUSE_1.0 as shown in Table II. The higher the correlation, the closer the similarity ratings are to those of the human ratings (HR).

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

Figure 1 shows a scatter plot for the relationship between the two variables [35]. In this instance, the two variables are the human ratings (HR) and the correlation following the fuzzy influence factor (FI) experiment. Each dot on the scatter plot shows the values for each sentence pair on the X and Y axis, with *x* being FI and *y* being HR. The scatter plot in Figure 1 shows the positive correlation of the human ratings (HR) with the fuzzy influencer factor (FI), each on a scale of [0..1], where 0 represents no similarity and 1 represents maximum similarity. For experiment 1.5 for the 15 sentence pairs that did not perform well under FUSE_1.0, the line of best fit shows the mathematically best fit for the data; also referred to as the 'trendline'. This line shows the behaviour of a set of data, when the line goes up, this shows a positive linear relationship between the variables.

SP15 where SP15a = *"The little village of Resina is also situated near the spot"* and SP15b = *"He seems an excellent man and I think him uncommonly pleasing"*, is a clear outlier with the average human rating being 0.075, where FUSE_4.0 coming close to 0.206. SP15 contains fuzzy words {*little*, *near*, *excellent* and *uncommonly*}, *little* and *near* belong to the *Size/Distance* category, *excellent* belongs to *Worth* category and *uncommonly* belongs to *Frequency* category.

Table III shows the correlation findings for experiments 1.6-1.10 ran on the sentences that performed well under FUSE_1.0. Results from Table III show that the component weightings (Sem 0.7, Syn 0.05, FI 0.25) from experiment 1.9 gave the best overall correlation and the highest correlation, beating the other algorithm measures for SP's that performed well originally with FUSE_1.0 as shown in Table IV. The higher the correlation, the closer the similarity ratings are to those of the human ratings (HR). Figure 2 shows the scatter plot for the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for experiment 1.9 for the 10 sentence pairs that performed well under FUSE_1.0. The trendline shows a positive linear relationship between the variables. The results from these experiments on the FI-25 dataset gave positive indicators that H_0 would be accepted.

B. *Experiment 2 - FI on 62-SP*

FI-25 was a limited dataset, so a series of further empirical experiments were undertaken using a similar range of semantic, syntactic and FI factor weights using the 62-SP dataset. 62-SP consisted of 62 sentence pairs specifically designed by English language experts to contain at least 2 fuzzy words per sentence from all 9 categories [15]. Table V shows the correlation findings for experiments 2.1-2.5 ran on the 62-SP dataset. Results from Table V show that the measures (Sem 0.5, Syn 0.2, FI 0.3) from experiment 2.5 gave the best overall correlation with human ratings and also higher than competing measures as shown in Table VI. The scatter plot in Figure 3 shows the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for experiment 2.5 for the 62-SP dataset.

C. *Experiment 3 - FI on MFWD*

The same 5 experiments were also conducted on the published MFWD dataset [34]. This dataset consisted of 30 sentence pairs specifically designed by English language experts to contain at least 2 fuzzy words per sentence. Table VII shows the correlation findings for experiments 3.1-3.5 for the MFWD dataset. Results shown in Table VII show that the measures (Sem 0.8, Syn 0.1, FI 0.1) from experiment 3.1 gave

TABLE I. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR FI-25 SP'S NOT PERFORMING WELL UNDER FUSE_1.0

Pearson Correlation	r Value	r Value	r Value	r Value	r Value
	Exp. 1.1 **Sem 0.8** **Syn 0.1** **FI 0.1**	**Exp. 1.2** **Sem 0.7** **Syn 0.1** **FI 0.2**	**Exp. 1.3** **Sem 0.75** **Syn 0.15** **FI 0.1**	**Exp. 1.4** **Sem 0.7** **Syn 0.05** **FI 0.25**	**Exp. 1.5** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.695292	0.706837	0.717356	0.687299	0.771050

TABLE II. COMPARISON OF SSM BEST RESULTS FROM TABLE I

SSM	r Value
HR vs FUSE_4.0	0.771050
HR vs FUSE_2.0	0.681673
HR vs FUSE_3.0	0.706030
HR vs STASIS	0.712598
HR vs API Semantic	0.495320
HR vs API Syntactic	0.883992
HR vs SEMILAR	0.765862

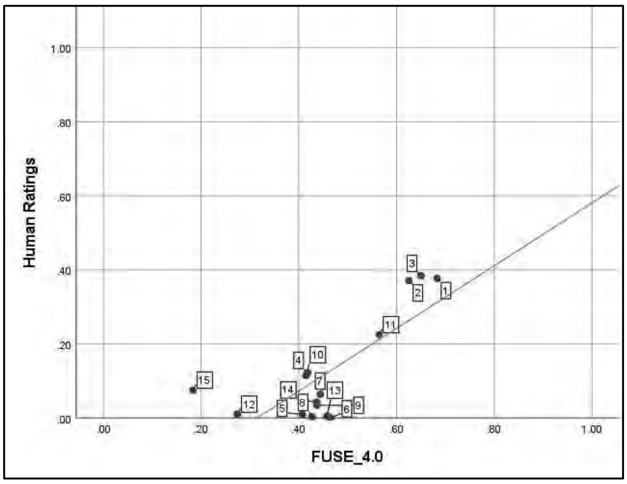


Fig. 1. FI-25 Scatter Plot NPW **(Sem 0.5, Syn 0.2, FI 0.3)**

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

TABLE III. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR FI-25 SP'S PERFORMING WELL UNDER FUSE_1.0

Pearson Correlation	**r Value**	**r Value**	**r Value**	**r Value**	**r Value**
	Exp. 1.6 **Sem 0.8** **Syn 0.1** **FI 0.1**	**Exp. 1.7** **Sem 0.7** **Syn 0.1** **FI 0.2**	**Exp. 1.8** **Sem 0.75** **Syn 0.15** **FI 0.1**	**Exp. 1.9** **Sem 0.7** **Syn 0.05** **FI 0.25**	**Exp. 1.10** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.249668	0.233447	0.187771	0.299713	0.082649

TABLE IV. COMPARISON OF SSM BEST RESULTS FROM TABLE III

SSM	**r Value**
HR vs FUSE_4.0	0.299713
HR vs FUSE_2.0	0.191413
HR vs FUSE_3.0	0.205204
HR vs STASIS	0.167745
HR vs API Semantic	0.051874
HR vs API Syntactic	0.073051
HR vs SEMILAR	0.128564

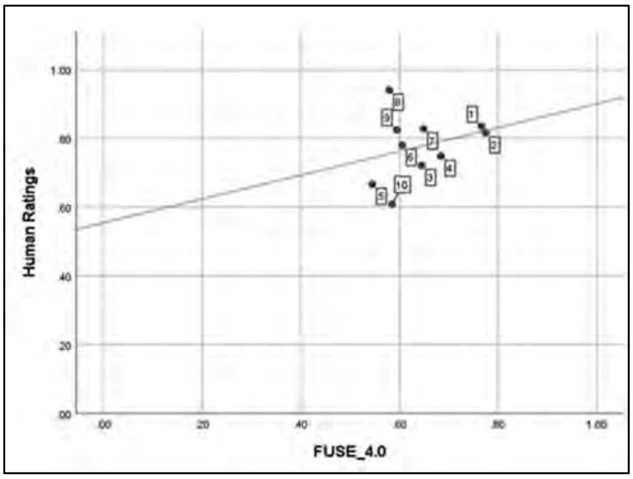


Fig. 2. FI-25 Scatter Plot PW **(Sem 0.7, Syn 0.05, FI 0.25)**

TABLE V. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR 62-SP

Pearson Correlation	**r Value**	**r Value**	**r Value**	**r Value**	**r Value**
	Exp. 2.1 **Sem 0.8** **Syn 0.1** **FI 0.1**	**Exp. 2.2** **Sem 0.7** **Syn 0.1** **FI 0.2**	**Exp. 2.3** **Sem 0.75** **Syn 0.15** **FI 0.1**	**Exp. 2.4** **Sem 0.7** **Syn 0.05** **FI 0.25**	**Exp. 2.5** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.622094	0.642160	0.646160	0.625525	0.702729

TABLE VI. COMPARISON OF SSM BEST RESULTS FROM TABLE V

SSM	**r Value**
HR vs FUSE_4.0	0.702729
HR vs FUSE_2.0	0.555268
HR vs FUSE_3.0	0.626043
HR vs STASIS	0.592999
HR vs API Semantic	0.526305
HR vs API Syntactic	0.671170
HR vs SEMILAR	0.664572

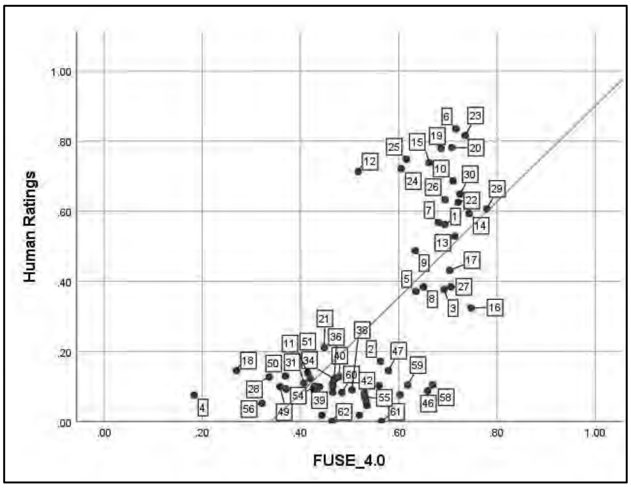


Fig. 3. 62-SP Scatter Plot **(Sem 0.5, Syn 0.2, FI 0.3)**

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

TABLE VII. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR MFWD

	r Value	**r Value**	**r Value**	**r Value**	**r Value**
Pearson Correlation	**Exp. 3.1** **Sem 0.8** **Syn 0.1** **FI 0.1**	**Exp. 3.2** **Sem 0.7** **Syn 0.1** **FI 0.2**	**Exp. 3.3** **Sem 0.75** **Syn 0.15** **FI 0.1**	**Exp. 3.4** **Sem 0.7** **Syn 0.05** **FI 0.25**	**Exp. 3.5** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.758884	0.741024	0.755944	0.734019	0.693317

TABLE VIII. COMPARISON OF SSM FROM BEST RESULTS FROM TABLE VII

SSM	**r Value**
HR vs FUSE_4.0	0.758884
HR vs FUSE_2.0	0.753772
HR vs FUSE_3.0	0.768331
HR vs STASIS	0.745248
HR vs API Semantic	0.700868
HR vs API Syntactic	0.393033
HR vs SEMILAR	0.730265

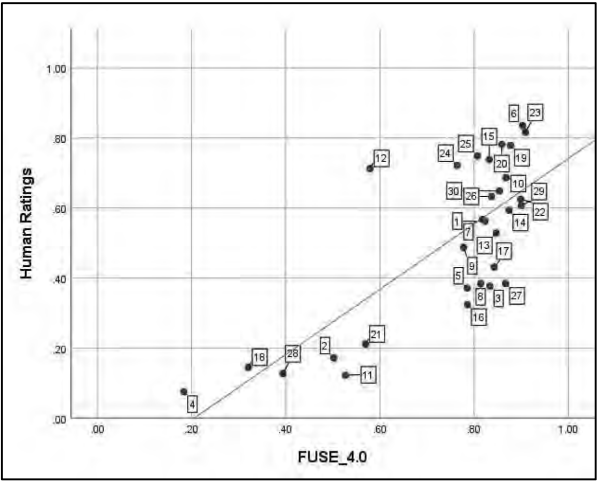


Fig. 4. MFWD Scatter Plot **(Sem 0.8, Syn 0.1, FI 0.1)**

TABLE IX. (A-VALUE) AND (P-VALUE) FOR EACH DATASET

Datasets	FI25_NPW	FI25_PW	FI25	62-SP	MFWD
a-value	0.998	0.953	0.997	0.987	0.999
p-value	.000	.000	.000	.000	.000

the best overall correlation and the highest correlation beating the other algorithm measures with the exception of FUSE_3.0 which was slightly higher as shown in Table VIII. The scatter plot in Figure 4 shows the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for experiment 3.1 for the MFWD dataset.

VI. DISCUSSION

Table IX shows information with regards to the datasets that were used in the FI experiment. The *a-value* shows the Intra-class Correlation Coefficient (ICC) for each of the datasets that we experimented on across the different algorithms. Since the *a-value* results are between 0.75 and 1.00 for each dataset, it is deemed that the inter-rater agreement of human ratings are *excellent* according to Cicchetti [30]. Table IX also shows that the *p-value* for each dataset is less than 0.05 for a confidence level of 95% and thus provides support for our research hypothesis H_0.

The snapshot of empirical experiments conducted on several datasets indicated that the inclusion of a FI factor in a FSSM can improve the performance of the algorithm in terms of its correlation with human ratings. The interaction of the FI factor with both the semantic and syntactic components of FUSE_4.0 must be kept to a minimum, in order to preserve the importance of the word order and ontological path length in calculating the overall similarity. This work, whilst accepting H_0, recognizes that more work needs to be done in determining a more generalizable FI factor.

VII. CONCLUSION AND FUTURE WORK

In closing, this work has shown that a fuzzy influence factor has a positive impact on the correlation of human ratings in a FSSM. Experimental results in this paper have shown that the FI factor must be empirically determined. The results across 3 datasets have shown an *excellent* rating for ICC. Although this FI is relatively simple, it has to a degree been able to model the uncertainty of human perception-based words which have already been modelled using Interval Type-2 fuzzy sets. The FUSE algorithm can show distinct benefits over crisp semantic similarity algorithms only when there is at least one fuzzy word in the short text pair. Therefore, the FUSE algorithm is recommended when it is important to assess the similarity of fuzzy words in a given context.

Further work includes investigating the generalisability of the FI factor and modelling fuzzy logic operators, such as NOT within the context of fuzzy short text similarity.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

ACKNOWLEDGMENT

This work was partially supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

REFERENCES

[1] Y. Li, Z.A. Bandar, J.D. O'Shea, D. Mclean and K. Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics", *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp.1138-1150, 2006.

[2] J.D. O'Shea, Z.A. Bandar and K. Crockett, "A new benchmark dataset with production methodology for short text semantic similarity algorithms", *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, no. 4, p.19, 2013.

[3] P. Zhang, X. Huang and L. Zhang, "Information mining and similarity computation for semi- / un-structured sentences from the social data", *Digital Communications and Networks*, 2020.

[4] C. Little, D. Mclean, K. Crockett and B. Edmonds, "A Semantic and Syntactic Similarity Measure for Political Tweets". *IEEE Access*, vol. 8, pp.154095-154113, 2020.

[5] N. Alnajran, K. Crockett, D. McLean and A. Latham, "An Empirical Performance Evaluation of Semantic-Based Similarity Measures in Microblogging Social Media" In *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, pp.126-135, Dec. 2018.

[6] D.W. Prakoso, A. Abdi, and C. Amrit, "Short text similarity measurement methods: a review", Soft Computing, pp.1-25, 2021.

[7] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2. no. 2, pp.1-25, 2008.

[8] V.W. Anelli, Y. Deldjoo, T. Di Noia, E. Di Sciascio and F.A. Merra, Sasha: Semantic-aware shilling attacks on recommender systems exploiting knowledge graphs. In *European Semantic Web Conference* ,pp. 307-323, May. 2020.

[9] H.J.P.J. Dong-hong, "Convolutional Network-Based Semantic Similarity Model of Sentences", *Journal of South China University of Technology (Natural Science)*, vol. 45, no. 3, p.68, 2017.

[10] X. Zhang, and A.A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion" *Information Processing & Management*, vol. 57, no. 2, p.102025, 2020.

[11] S. Venkatraman, B. Surendiran and P.A.R. Kumar, "Spam e-mail classification for the Internet of Things environment using semantic similarity approach", *The Journal of Supercomputing*, vol. 76, no. 2, pp.756-776, 2020.

[12] O. Araque and C.A. Iglesias, "An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity", *IEEE Access*, vol. 8, pp.17877-17891, 2020.

[13] L.A. Zadeh, "Fuzzy logic = computing with words", *Fuzzy Systems, IEEE Transactions on*, vol. 4, no. 2, pp.103-111, 1996.

[14] J.M. Mendel, "Type-2 Fuzzy Sets as Well as Computing with Words", In *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp.82-95, Feb. 2019.

[15] N. Adel, K. Crockett, D. Chandran and J.P. Carvalho, "Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures", *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp.1-8, 2020.

[16] K. Crockett, D. Chandran and D. Mclean, "On the Creation of a Fuzzy Dataset for the Evaluation of Fuzzy Semantic Similarity Measures", *IEEE WCCI – FUZZ*, China, pp.752-759, 2014.

[17] M. Hao and J.M. Mendel, "Encoding words into normal interval type-2 fuzzy sets: HM approach", *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp.865-879, 2016.

[18] N. Adel, K. Crockett, A. Crispin, J.P. Carvalho and D. Chandran, Human Hedge Perception–and its Application in Fuzzy Semantic Similarity Measures. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-7, 2019.

[19] N. Adel, K.A. Crockett, A. Crispin, D. Chandran and J. Carvalho, "FUSE (Fuzzy Similarity Measure) - A Measure for Determining Fuzzy Short Text Similarity Using Interval Type-2 Fuzzy Sets", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp.1-8, 2018.

[20] A. Bilgin, H. Hagras, A. Malibari, M.J. Alhaddad, and D. Alghazzawi, "Towards a general type-2 fuzzy logic approach for computing with words using linear adjectives", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* pp.1-8, 2012.

[21] J.M. Mendel and D. Wu, "Perceptual Computing: Aiding People in Making Subjective Judgments", John Wiley & Son, 2010.

[22] Princeton University, "About Wordnet", [Online], Available: https://wordnet.princeton.edu/ [Accessed 13 Jun. 2014].

[23] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources", *IEEE Transactions on knowledge and data engineering*, vol. 15, no. 4, pp.871-882, 2003.

[24] V. Rus, M. Lintean, R. Banjade, N.B. Niraula and D. Stefanescu, "Semilar: The semantic similarity toolkit", In *Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations*, pp.163-168, Aug. 2013.

[25] SpazioDati, "Dandelion API", [Online], Available: https://dandelion.eu/ [Accessed 24 Jan. 2020].

[26] T. Pedersen, S. Patwardhan and J. Michelizzi, "WordNet:Similarity - Measuring the Relatedness of Concepts", In *The Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pp.1024-1025, Jul. 2004.

[27] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman. "Indexing by latent semantic analysis", *Journal of the American society for information science*, vol. 41, no. 6, pp.391-407, 1990.

[28] T. Luchoomun, M. Chumroo and V. Ramnarain-Seetohul, "A knowledge based system for automated assessment of short structured questions", In *2019 IEEE Global Engineering Education Conference (EDUCON)*, pp.1349-1352, Apr. 2019.

[29] Kent State University, "SPSS Tutorials: Pearson Correlation", [Online] Available: https://libguides.library.kent.edu/SPSS/PearsonCorr [Accessed 18 Sept. 2020].

[30] D.V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology", *Psychological Assessment*, vol. 6, no. 4, p. 284, 1994.

[31] M.L. McHugh, Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), pp.276-282, 2012.

[32] Statistics How To, "Intraclass Correlation", [Online] Available https://www.statisticshowto.com/intraclass-correlation/#:~:text=Intraclass%20correlation%20measures%20the%20reliability,groups%20or%20sorted%20into%20groups.&text=A%20high%20Intraclass%20Correlation%20Coefficient,values%20from%20the%20same%20group, [Accessed 11 Jan. 2021].

[33] Readability Formulas, "Automatic Readability Checker", [Online] Availble https://readabilityformulas.com/free-readability-formula-tests.php, [Accessed 15 Jan. 2018].

[34] D. Chandran, "The development of a fuzzy semantic sentence similarity measure", *Doctorate of Philosophy*, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2013.

[35] Chartio | Data Tutorials, "A Complete Guide to Scatter Plots", [Online] Available https://chartio.com/learn/charts/what-is-a-scatter-plot/, [Accessed 11 Jan. 2021].

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:39:17 UTC from IEEE Xplore. Restrictions apply.

Using Fuzzy Set Similarity in Sentence Similarity Measures.

IEEE International Conference on Fuzzy Systems

2020

Using Fuzzy Set Similarity in Sentence Similarity Measures

Valerie Cross
Computer Science and Software Engineering
Miami University
Oxford, OH USA
crossv@miamioh.edu

Valeria Mokrenko
Computer Science and Software Engineering
Miami University
Oxford, OH USA
mokrenvi@miamioh.edu

Keeley Crockett
Computational Intelligence Lab
Manchester Metropolitan University
Manchester, UK
K.Crockett@mmu.ac.uk

Naeemeh Adel
Department of Computing and Maths, Manchester Metropolitan University
Manchester, UK
N.Adel@mmu.ac.uk

***Abstract*— Sentence similarity measures the similarity between two blocks of text. A semantic similarity measure between individual pairs of words, each taken from the two blocks of text, has been used in STASIS. Word similarity is measured based on the distance between the words in the WordNet ontology. If the vague words, referred to as fuzzy words, are not found in WordNet, their semantic similarity cannot be used in the sentence similarity measure. FAST and FUSE transform these vague words into fuzzy set representations, type-1 and type-2 respectively, to create ontological structures where the same semantic similarity measure used in WordNet can then be used. This paper investigates eliminating the process of building an ontology with the fuzzy words and instead directly using fuzzy set similarity measures between the fuzzy words in the task of sentence similarity measurement. Their performance is evaluated based on their correlation with human judgments of sentence similarity. In addition, statistical tests showed there is not any significant difference in the sentence similarity values produced using fuzzy set similarity measures between fuzzy sets representing fuzzy words and using FAST semantic similarity within ontologies representing fuzzy words.**

Keywords—ontology, semantic similarity, fuzzy set similarity measures, human perception, sentence similarity measures

I. Introduction

Humans often find it easier to express domain knowledge using inexact, vague terms, or fuzzy words. Such words challenge the communication between humans and machines. Determining how similar two blocks of text are also faces the challenge of dealing with fuzzy words. Measuring the similarity between crisp words has typically been handled using semantic similarity measures. Much research has examined the use of semantic similarity measures within the context of an ontology, a knowledge structure containing concepts and defining the relationships between these concepts.

STASIS [1] is a system that produces sentence similarity measures between blocks of text. It measures the similarity between pairs of individual words, one from each block. The semantic similarity measure proposed in [2] is used within the WordNet ontology. Although the STASIS work made progress in measuring text similarity, it failed to address the occurrence of imprecise and vague words, i.e., fuzzy words that occur extensively in natural language. This capability is needed in order to advance conversational understanding between humans and machines.

Fuzzy sets can serve as a means of representing fuzzy words. A framework for handling fuzzy words is the computing with words (CWW) [3] methodology by which fuzzy words can be quantified, scaled against each other and then become machine representable. The quantifying and scaling steps require that humans provide their perception of fuzzy words. Once fuzzy words are machine representable, then similarity measurement between the words can be performed. This fuzzy word similarity measurement is a necessary task in defining fuzzy sentence similarity measures (SSMs).

Since STASIS does not handle fuzzy words, additional research pursued improvements to SSMs by addressing this limitation. The FAST (Fuzzy Algorithm for Similarity Testing) [4] system uses CWW methods to develop a SSM that incorporates the similarity measurement between fuzzy words found in sentences or pieces of short text. The additional work in FAST to handle fuzzy words showed an improvement in its SSM as compared to that of STASIS when evaluated based on their correlations with human judgments of sentence similarities. This experimental study required the creation of datasets containing quantified fuzzy words. The quantification was based on surveying humans on their perceived numerical evaluation of the fuzzy words. These fuzzy words are organized into structured ontologies where the semantic similarity found in [2] can be used to measure the similarity between the fuzzy words.

Fuzzy sets can be modelled as type-1 or type-2 fuzzy sets. Further research has explored the use of type-2 fuzzy sets for fuzzy words in the FUSE (FUzzy Similarity mEasure) system [5]. It extends FAST by replacing the type-1 fuzzy set representations with type-2 fuzzy sets. The rationale was that type-1 fuzzy sets could not reflect the subjective nature of the

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

human evaluators and capture the uncertainty of humans [6]. Interval sets are used to represent the type-2 membership functions since they are simpler to use.

As done in FAST, the fuzzy words are arranged into ontologies. The same semantic similarity [2] used in STASIS and FAST is used in FUSE. Both FUSE and FAST require this step of transforming the fuzzy sets representing the fuzzy words into ontologies so that a semantic similarity measure can be used within the constructed ontologies. The major difference between the FAST and FUSE ontologies is in the level of detail considered in their construction. FAST with its type-1 fuzzy sets uses only 5 nodes with a depth of 2 in its ontologies. FUSE with its type-2 interval fuzzy sets uses ontologies with 11 nodes with a depth of 5. Building these ontologies based on the developed fuzzy sets for the fuzzy words is a required step to use the semantic or ontological similarity measure.

A previous paper [7] focused on determining if fuzzy set similarity measures might be used directly on the fuzzy sets with the goal of eliminating the ontology construction step. The measurement of similarity between fuzzy words represented as type-1 fuzzy sets used the following three existing fuzzy set similarity measures [8] of Zadeh's sup-min, Jaccard, and GeoSim. A fourth similarity measure referred to as Type-2 Dist uses a scaled COG for the type-2 fuzzy sets and the distance between their scaled COGs. The fuzzy set definitions for the fuzzy words for both type-1 and type-2 were obtained from the authors in [5]. The paper [7] reports on how well these simpler fuzzy set similarity measures correlated with the semantic similarity measure used in FAST and FUSE.

All of the fuzzy set similarity measures had a much higher correlation with FUSE's semantic similarity results based on its more sophisticated 11 node ontologies than FAST's semantic similarity results correlated with those of FUSE. The study showed that the results from the FAST and FUSE semantic similarity measures are very much dependent on the structure of the ontologies that have been developed from the type-1 fuzzy sets and type-2 interval fuzzy sets. Evaluating the use of these fuzzy set similarity measure in the computation of the FAST and FUSE sentence similarity measures, however, was not undertaken in that work.

The objective of this paper is investigate the performance of the fuzzy similarity measures when they replace the semantic similarity measure of FAST and FUSE in their SSMs. Their performance in the task of sentence similarity measurement is evaluated based on how well the resulting SSMs correlate with that of human judgments of sentence similarity.

The paper organization is as follows: Section II reviews from [8] the fuzzy word representation used and the four fuzzy set similarity measures. Section III describes the software that was available for this study and the modifications made to use the fuzzy set similarity measures in the SSMs. Section VI explains the experiments and how the evaluation of the fuzzy set similarity measures is performed. The results from the experiments and analysis are presented. Finally, Section V presents the conclusions and future work.

II. Fuzzy Words and Similarity Between Them

This section is a summary of the description found in [7]. To use fuzzy set similarity measures the fuzzy words must have a fuzzy set representation. The FAST research used questionnaires with human evaluators to develop a defuzzified value or mean and the standard deviation for each of the fuzzy words. These values for type-1 fuzzy sets were acquired from the FAST researchers. With these values, a pseudo triangular fuzzy set is created where the membership degree at the mean value is 1.0. A normal probability density distribution is used and values ±3 standard deviations away from the mean were used for the end points of the triangular fuzzy set since 99.7% of the data is within three standard deviations of the mean. Fig. 1 shows the triangular membership function for *centre* with a mean of 4.93 and a standard deviation of 0.5. The simplest approach to building fuzzy sets for fuzzy words is used since the hypothesis is to determine if these sets based on human judgment might be used with well-known fuzzy set similarity measures to eliminate the need to build ontologies.

The same twenty word pairs and the triangular membership type-1 fuzzy sets created for them in [7] are used for this research along with the associated pairs of sentences containing those fuzzy words. The first three fuzzy set similarity measures discussed below can simply be used on the triangular membership functions. Because for FUSE it was thought that type-2 fuzzy sets may better represent the subjective nature of a fuzzy word. Type-2 interval fuzzy sets were created for fuzzy words and a center of gravity (COG) was determined using the upper and lower footprints of uncertainty for the type 2 fuzzy sets. These COG values were acquired from the FUSE research and used in [7] as well as in this current research. The fourth fuzzy set similarity measure type-2 distance uses distance between COGs in determining the similarity between fuzzy words. Both the sup-min and the Jaccard measures produce a 0 similarity when the two fuzzy sets do not overlap. GeoSim and the COG type-2 similarity measures, however, produce a non-zero value even when the fuzzy sets do not overlap since both are based on distance.

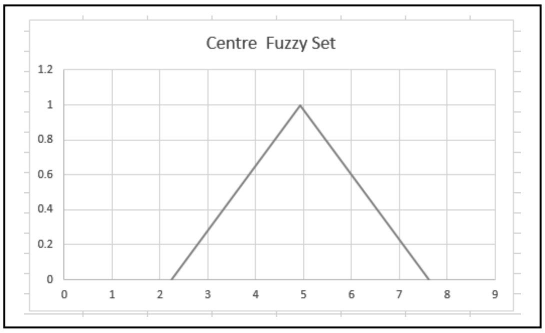


Fig. 1 Centre fuzzy set

A. Sup-Min

In [8] a detailed and thorough review of a variety of fuzzy set similarity measures is provided. Zadeh's consistency index also known as the sup-min or partial matching index, falls into the set-theoretic category of fuzzy similarity measures. It

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

roughly estimates the similarity between two fuzzy sets by finding at what domain values they intersect and determines their similarity by taking the highest membership degree among their intersection points. Given two fuzzy sets A and A', similarity between the two is determined as

$$S_{Zadeh}(A, A') = \sup_{u \in U} T(A'(u), A(u)) \qquad (1)$$

where T can be any t-norm, but usually the minimum is used for the t-norm. It is referred to as a partial matching index since it only provides an estimated similarity value between the two fuzzy sets.

B. Jaccard

The fuzzy Jaccard similarity measure is defined as a fuzzy extension of the Jaccard index [14] between two crisp sets by replacing set cardinality with fuzzy set cardinality. This fuzzy set similarity measure is also in the set theoretic category but provides a more comprehensive view of similarity between the two fuzzy sets since all elements in both fuzzy sets are considered and not just the intersection points as in sup-min. Given two fuzzy sets A and A', similarity between the two is determined as

$$S_{Jaccard}(A, A') = |\, A \text{⍰} A'| \; / \; |\, A \cup A '| \qquad (2)$$

so the similarity is measured by the proportion of the area of the intersection of the two fuzzy sets to the area of the union of the two fuzzy sets.

C. Geometric Fuzzy Similarity Based on Dissemblance Index

Set theoretic fuzzy set similarity measures do not consider the distance of the fuzzy set A' from A. With the geometric fuzzy similarity measure [9], the distance between the two sets is the basis for determining their similarity. This distance is based on the dissemblance index that measures the distance between two real intervals. If V = [v_1, v_2] and W = [w_1, w_2], then

$$DI(V,W) = (|v_1 - w_1| + |v_2 - w_2|) / [2(\beta_2 - \beta_1)] \qquad (3)$$

where [β_1, β_2] is an interval that contains both V and W. The factor $2(\beta_2 - \beta_1)$ is necessary to produce a normalized degree of dissemblance such that $0 \leq D(V, W) \leq 1$. The dissemblance index consists of two components, the left and right sides of each interval and may be generalized to fuzzy intervals.

A fuzzy interval N is defined by a pair of boundary functions L and R and parameters (r_1, r_2, λ, ρ). The core of N, the values for which $\mu_N(r)=1.0$ is the interval [r_1, r_2]. Parameters λ and ρ are used to define the left L and the right R boundary functions and the support of N, the values for which $\mu_N(r) \geq 0$, which is [$r_1 - \lambda$, $r_{2} + \rho$]. The L function and the R function define the membership functions for elements in the intervals [$r_1 - \lambda$, r_1] and [r_2, $r_2 + \rho$], respectively. If L is positively sloping and linear and R is negatively sloping and linear then the interval N is a trapezoidal fuzzy membership function. Calculating the fuzzy dissemblance index between A and A' is done as an integration over α in the range 0 to 1 as

$$fDI(A'(u),A(u))=[\int||L_{A'}(\alpha)-L_A(\alpha)|+|R_{A'}(\alpha)-R_A(\alpha)|d\alpha] \;/\; (2(\beta_2-\beta_1)) \qquad (4)$$

where [β_1, β_2] is an interval that contains both A' and A. It can be converted into a similarity measure between the fuzzy intervals as

$$S_{GeoSim}(A, A') = 1 - fDI(A(u), A'(u)) \qquad (5)$$

With this similarity measure, even though A and A' may not overlap, a nonzero similarity value is produced since the distance between the two sets is used.

D. Similarity on Type-2 Defuzzified Values Distance

As previously explained in [5], type-2 interval fuzzy sets were used and then defuzzified into a single value by adapting Mendel's footprint of uncertainty (FOU) method [6]. For each word in the six categories, the COG was determined using the lower FOU and upper FOU. The COGs were then scaled into the range [-1, +1]. To see how well a measure based solely on the distance between these scaled COG values worked, the following simple similarity measure is also used in this study:

$$S_{Type2\text{-}Dist}(A, A') = 1 - |\, COG_{Scaled}(A) - COG_{Scaled}(A')| / 2 \qquad (6)$$

The distance between the two centers of gravity is normalized by the size of the scaled interval [-1, +1].

III. Software Used and Modified in Experiments

In [7], the study focused on determining how well fuzzy set similarity measures correlated with the semantic similarity measure proposed in [2] which were used on the 5 node ontologies in FAST and the 11 node ontologies in FUSE. In that study, the effectiveness of the fuzzy set similarity measures in the overall task of determining sentence similarity was not investigated; only how well the different fuzzy words similarity measures correlated with each other is reported.

Here our research requires that the previous sentence similarity measurement systems, such as STASIS and FAST, be modified to use the fuzzy set similarity measures between fuzzy words in place of the semantic similarity measure. This modification is needed to determine the performance of fuzzy set similarity measures within a sentence similarity measurement. First STASIS, FAST, and FUSE are briefly described and then the modifications made first to STASIS and then to FAST are presented.

A. STASIS, FAST, amd FUSE

STASIS [1] determines the degree of similarity between sentences or short blocks of texts by using both semantic information and word order information implied in the sentences. The semantic similarity calculation relies on the vocabulary in the WordNet ontology and on corpus statistics found in the Brown Corpus [10]. The semantic similarity between pairs of words *w_1* and *w_2*, one taken from each sentence, is determined as

$$\boxed{S(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}}$$

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

where *l* represents the length of the path between the two words in WordNet and *h* represents the depth of their common subsumer.

STASIS uses this word similarity measure between all possible pairs of words from the two texts. A semantic vector is created for the sentence that weights the similarity based on the importance of the words where importance is derived from the Brown Corpus statistics. A syntactic vector is created using the positions of words in the texts. These two vectors are combined to produce an overall level of similarity for the two sentences.

FAST was developed to address the limitations of STASIS since STASIS is not able to determine semantic similarity between fuzzy words in judging the similarity between pairs of sentences. A major task to accomplish this is the creation of fuzzy set representations for the fuzzy words. A dataset containing quantified fuzzy words is organized hierarchically into six different categories [7]: *age, size/distance, frequency, goodness, membership level* and *temperature*. As previously explained, the fuzzy words in each category were quantified by human subjects. FAST used this quantification to create a 5 node ontology for each of the categories. FAST can be seen as an extension of STASIS since in order to use the same semantic similarity measure S(w_1, w_2), FAST required these ontologies. FAST follows the same approach to measuring sentence similarity except fuzzy words that are not found in WordNet do not receive a similarity of 0. Instead the fuzzy words are found in the appropriate category ontology and their semantic similarity can be used in determining the overall sentence similarity.

FUSE is an extension of FAST in that it uses the same approach with organized category ontologies but they are built from type-2 fuzzy sets. These type-2 fuzzy sets were based on questionnaires completed by human subjects. FUSE also increases the fuzzy word vocabulary by 57%. The S(w_1, w_2) measure is used in FUSE as done with STASIS and FAST. The only difference is the ontologies being used to determine path length *l* and the depth *h* of the common subsumer.

Both the STASIS and the FAST code are written in Python and were acquired from the first two authors in [5]. The provided FAST code was stated to be an earlier version of the code on which FUSE development was initiated. An effort was made to acquire more recent FAST code by contacting the first author of [4] in hopes of getting the most recent FAST. This effort was not successful. All results using this FAST code or modifications of FAST are from the use of this acquired FAST code; therefore, the SSM results reported in this paper could deviate from those reported in [4]. The FUSE code, also written in Python, was not made available, but the sentence similarities produced by FUSE were provided for the pairs of sentences used in this research. These sentence similarity values are given in the FUSE Reported column in Table VI.

B. Software Modifications

The objective was to modify STASIS so that instead of producing a fuzzy word similarity of 0 when the fuzzy words could not be found in WordNet, the fuzzy set similarity measure between the fuzzy words is used. WordNet is used to check for synsets for a word in the sentence. If no synsets are returned, then the word cannot be found. The word may also have multiple synsets associated with it since a word could have more than one “sense” such as the word *bat*, i.e., the animal bat and a bat used in baseball. As implementation and testing progressed, it was discovered that some fuzzy words might be found in the WordNet ontology so that they could have a semantic similarity with another non-fuzzy word, i.e., a fuzzy and non-fuzzy word pair or may be paired with another fuzzy word. The implementation decision was made to only replace the fuzzy and non-fuzzy semantic similarity value with the fuzzy set similarity value for the fuzzy-fuzzy word pair if the fuzzy set similarity is greater.

A study of the FAST code revealed that the FAST developer made the same decision to replace the WordNet semantics similarity with the similarity produced by using the category ontologies for fuzzy words only if it was greater than the WordNet semantic similarity. This finding confirmed the decision that had been made in the modifications to STASIS to add fuzzy set similarity measurement between fuzzy words. The FAST code is very similar to the STASIS code except for the use of the category ontologies when determining similarity between fuzzy words. Here the modification required replacing the use of semantic similarity measures with fuzzy set similarity measures between the fuzzy words.

In both the obtained STASIS and FAST code, identical words are only assigned a similarity value of 1 if no synsets exist for the identical word. If a word has multiple synsets, as previously explained, it has multiple senses. If this is the case, then the assigned semantic similarity is the maximum between all the different senses for the word. As implementation progressed, a decision was made to create another version that also assigns a similarity value of 1 if there is only one synset for the word since that means there is only one “sense” for the word.

IV. Experimental Results and Analysis

The FAST research developed pairs of sentences for evaluating the FAST sentence similarity measures. First a list of 30 sentence pairs from the dataset in [11] was generated with 20 having a high level of similarity, 5 of medium level and 5 of low. The sentence pairs were split and sentences randomly divided among three English language experts, who added a fuzzy word to each sentence and enhanced or reduced a particular attribute from it. From the three versions of the sentence, two were randomly paired together. Then the similarity of the sentences pairs were determined by surveying 18 people using questionnaires that asked them to rate how similar the sentences were on a scale of 0 to 10. These numbers were summarized as an average human rating score (AHR). Of the 30 sentence pairs, 20 contain fuzzy words in the same fuzzy category. The fuzzy category in some senses specifies the context of the same word which may exist in different fuzzy categories. These pairs with their italicized fuzzy words are listed in Table I and used in this study. Fuzzy similarity measure between fuzzy words only is performed for those word pairs in the same fuzzy category. In these experiments it was decided that it did not make sense to

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

measure the similarity between fuzzy words used in different contexts. No fuzzy rules are used; only fuzzy similarity measures are used between the fuzzy words.

TABLE I. 20 SENTENCE PAIRS WITH FUZZY WORD

Sentence Pair	**Sentence 1** **Sentence 2**
P1	When I was going out to meet my friends there was a *short* delay at the train station. The train operator announced to the passengers on the train that there would be a *massive* delay.
P5	Sometimes in a *large* crowd accidents may happen, which can cause life threatening injuries. There was a *small* heap of rubble left by the builders outside my house this morning.
P7	If you continuously use these products, I guarantee you will look very *young*. I assure you that, by using these products over a long period of time, you will appear almost *youthful*.
P8	I always like to have a *tiny* slice of lemon in my drink, especially if it's coke. I like to put a *large* wedge of lemon in my drinks, especially cola.
P9	I dislike the word quay, it confuses me every time, I *always* think of the thing for locks, there's another one. I dislike the word quay, it confuses me every time, I *always* think of the thing for locks, there's another one.
P10	Though it took many hours travel on the extremely *long* journey, we finally reached our house safely. We got home safely in the end, though it was a *mammoth* journey.
P11	The man presented a *minuscule* diamond to the woman and asked her to marry him. A man called Dave gave his fiancée an *enormous* diamond ring for their engagement.
P13	The *tiny* ghost appeared from nowhere and frightened the old man. The *diminutive* ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.
P15	Midday is 12 o'clock in the *midpoint* of the day. Midday is 12 o'clock in the *centre* of the day.
P16	The first thing I do in a morning is make myself a *lukewarm* cup of coffee. The first thing I do in the morning is have a cup of *hot* black coffee.
P18	This is a terrible noise level for a new car, I expected it to be of *good* quality. That's a very good car, on the other hand mine is *great*.
P19	Meet me on the *huge* hill behind the church in half an hour. Join me on the *small* hill at the back of the church in 30 minutes.
P20	It gives me *immense* pleasure to announce the winner of this year's beauty pageant. It's a *great* pleasure to tell you who has won our annual beauty parade
P22	Will I have to drive a *great* distance to get to the nearest petrol station? Is it a *long* way for me to drive to the next gas station?
P23	You have a very familiar face; do I know you from somewhere *nearby*? You have a very familiar face; do I know you from somewhere where I used to live *faraway*.
P24	I have invited a *great* number of different people to my party so it should be interesting. A *small* number of invitations were given out to a variety of people inviting them down the pub.
P25	I am sorry but I can't go out as I have *loads* of work to do. I've a *gargantuan* heap of things to finish so I can't go out I'm afraid.
P27	Will you drink a glass of *excellent* wine while you eat? Would you like to drink this *wonderful* wine with your meal?
P29	*Large* boats come in all shapes but they all do the same thing. *Oversized* chairs can be comfy and not comfy, depending on the chair.
P30	I am so hungry I could eat a whole *big* horse plus desert. I could have eaten another *massive* meal, I'm still starving.

In Table II and Table III, several different sentence similarity results are given for STASIS and FAST. The second column in these two tables is the AHR for the pair of sentences. The Reported Results are those provided directly from the FAST and FUSE researchers. The Obtained Code Zero Synset column shows the results produced by running the code provided by the FAST and FUSE researchers. In this code identical words must have zero synsets to be assigned a similarity value of one. The last column shows the results for the modified STASIS and FAST code that uses a test checking if identically spelled words have only one or zero synsets.

Table II and Table III also show both the Pearson and Spearman correlation of the various STASIS and FAST versions. A substantial difference in the correlation with the similarities of these versions to the AHR exists. For STASIS, the Obtained Code results highest correlation. A possible explanation for the difference in the Reported versus the Run results is the difference in the Natural Language Took Kit (NLTK) [12] versions from when STASIS originally produced the results and the current version (3.4.5). It is unlikely WordNet versions caused the difference because although synset offsets change, synsets are stable with few splits and merges between the versions [13]. Similarly for FAST, the Obtained Code results produce a higher correlation than the Reported results. ANOVA analysis performed on the three different versions of STASIS indicates the means for three STASIS versions are not significantly different at the 0.05 level. The same outcome occurred for the three FAST versions.

FAST correlations are higher than the STASIS correlations. These higher correlations are expected since FAST handles fuzzy words that STASIS is not able to. To determine if a statistically significant difference exists in the means of the STASIS and FAST Obtained Code results, a two tail t-test was performed. The outcome of the t-test shows that there is a significant difference between STASIS Obtained Code Results and FAST Obtained Code Results at the 0.05 level with p-value of 0.0073. For the t-test between STASIS Modified Zero or One Synset and FAST Modified Zero or One Synset produced a significant difference with a p-value of 0.0013. These t-tests verify that handling fuzzy words in FAST does significantly improve correlation with human judgments of sentence similarity over that of STASIS.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

TABLE II. TASIS RESULTS

Sentence Pairs	Human Judgment (AHR)	Reported Results	Obtained Code Zero Synset	Modified to Zero or One Synset
P1	3.833	0.74688	0.74688	s 0.746352
P5	1.281	0.553945	0.553945	0.543552
P7	7.095	0.854431	0.854431	0.806182
P8	6.719	0.90160	0.779976	0.763487
P9	0.952	0.68323	0.615611	0.619075
P10	8.248	0.707534	0.707534	0.742276
P11	4.957	0.531135	0.465735	0.449818
P13	3.286	0.533853	0.564408	0.577580
P15	9.138	0.999921	0.999926	0.999889
P16	6.781	0.844044	0.844044	0.844044
P18	2.11	0.475089	0.496757	0.496756
P19	6.757	0.779292	0.779292	0.732798
P20	8.986	0.758728	0.823382	0.793999
P22	8.852	0.882129	0.882129	0.881933
P23	7.043	0.858609	0.858609	0.858609
P24	3.833	0.707128	0.707128	0.707051
P25	8.857	0.626350	0.742006	0.693284
P27	8.919	0.707795	0.707795	0.614119
P29	1.295	0.389489	0.389489	0.268906
P30	6.624	0.508935	0.534416	0.529974
Pearson with AHR		**0.631977**	**0.724682**	**0.6775326**
Spearman with AHR		**0.628056**	**0.717563**	**0.642347**

TABLE III. FAST RESULTS

Sentence Pairs	Human Judgment (AHR)	Reported Results	Obtained Code Results	Modified to Zero or One Synset
P1	3.833	0.716059	0.766476	0.76603505
P5	1.281	0.553945	0.554011	0.54362499
P7	7.095	0.848375	0.837837	0.83783737
P8	6.719	0.896886	0.772687	0.77268639
P9	0.952	0.681290	0.613872	0.61774001
P10	8.248	0.824531	0.822187	0.83205864
P11	4.957	0.517416	0.489134	0.4891348
P13	3.286	0.583988	0.608148	0.60813224
P15	9.138	0.999921	0.999890	0.99989027
P16	6.781	0.897493	0.861690	0.86169039
P18	2.11	0.498348	0.498887	0.49888599
P19	6.757	0.782346	0.779151	0.77914815
P20	8.986	0.782177	0.831654	0.83164902
P22	8.852	0.901850	0.900176	0.89987811
P23	7.043	0.891414	0.872690	0.87269049
P24	3.833	0.712779	0.713812	0.71373476
P25	8.857	0.664910	0.758325	0.75585767
P27	8.919	0.794916	0.792803	0.7927075
P29	1.295	0.372960	0.477078	0.37296983
P30	6.624	0.563401	0.567695	0.56340074
Pearson with AHR		**0.718752**	**0.782514**	**0.785599**
Spearman with AHR		**0.686724**	**0.786762**	**0.780745**

The previous discussion compared the results of STASIS and FAST without using fuzzy set similarity measures. FAST handles fuzzy words by using semantic similarity within its category ontologies. Table IV show the sentence similarity values produced after modifying the obtained STASIS code to use each of the four fuzzy set similarity measures between fuzzy words that cannot be found in WordNet. The fuzzy set similarity value may also replace the semantic similarity measure from WordNet when the fuzzy set similarity value is greater than that of the semantic similarity within WordNet. The values in Table IV are based on using a check for Zero or One Synset when identical words are found. That version produces slightly higher correlations with human judgments of sentence similarities than just checking for zero synsets.

Since STASIS GeoSim has the highest Pearson correlation with human judgments, a t-test on the SSM values between it and those of FAST Modified Zero or One Synset was performed. The purpose of the t-test is to determine if the modified STASIS using fuzzy set similarity measures differs significantly from FAST using semantic similarity within its category ontologies. There is no statistically significant difference between their SSM values with a p-value of 0.06

Table V shows the SSM values produced after modifying the obtained FAST code to use the four fuzzy set similarity measures between fuzzy words instead of the FAST semantic similarity measure within its category ontologies. Again, the SSM values in Table V are based on using Zero or One Synset check when identical words are found. Since the FAST Zadeh SSM values in Table V have the highest correlations with those of human judgments, a t-test between the SSM values of FAST Zadeh and those of FAST Modified Zero or One Synset found in Table III was performed. There is no statistically significant difference between the FAST Modified Zero or One Synset SSM values and the FAST Zadeh SSM values with a p-value of 0.30.

An ANOVA test was performed on the STASIS SSM values for the four fuzzy set similarity measures in Table IV and showed no statistically significant difference among the four fuzzy set similarity measures with a p-value of 0.98.

An ANOVA test was also performed on the FAST SSM values for the four fuzzy set similarity measures in Table V and showed no statistically significant difference among the four fuzzy set similarity measures with a p-value of 0.91.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

TABLE IV. STASIS RESULTS WITH FUZZY SET SIMILARITY MEASURES

Sentence Pairs	GeoSim	Zadeh	Jaccard	Type2-Dist
P1	0.787064	0.763680	0.754448	0.775012
P5	0.544365	0.543552	0.543552	0.543552
P7	0.846477	0.845563	0.82348 S7	0.850371
P8	0.792185	0.779492	0.768859	0.787140
P9	0.623408	0.623408	0.623408	0.623408
P10	0.834697	0.833500	0.810405	0.836665
P11	0.498857	0.449818	0.449818	0.468732
P13	0.602700	0.611673	0.557758	0.612318
P15	0.999889	0.999899	0.999889	0.999903
P16	0.898757	0.903868	0.878825	0.886171
P18	0.512070	0.511553	0.504335	0.514108
P19	0.790848	0.770028	0.742340	0.760205
P20	0.834148	0.834457	0.828792	0.834934
P22	0.901838	0.901839	0.899851	0.901530
P23	0.914425	0.873909	0.866268	0.912640
P24	0.717015	0.711358	0.707051	0.710872
P25	0.795177	0.799607	0.758452	0.777154
P27	0.798399	0.803035	0.765686	0.798276
P29	0.443928	0.452315	0.413302	0.435904
P30	0.559292	0.559306	0.559001	0.558381
Pearson with AHR	0.780550	0.7769880	0.766168	0.771685
Spearman with AHR	0.807823	0.805566	0.803310	0.792779

TABLE V. FAST RESULTS WITH FUZZY SET SMILARITY MEASURES

Sentence Pairs	GeoSim	Zadeh	Jaccard	Type2-Dist
P1	0.77846842	0.73734862	0.71299107	0.75310801
P5	0.53784191	0.53702486	0.53702486	0.53702486
P7	0.77651531	0.77548316	0.75288337	0.78118585
P8	0.79134429	0.77900059	0.76880401	0.78646828
P9	0.6249823	0.6249823	0.62498230	0.62498230
P10	0.78147147	0.77990048	0.75305702	0.78415594
P11	0.49646667	0.48913428	0.48913428	0.46645995
P13	0.59921112	0.60815523	0.55496662	0.60880177
P15	0.99989028	0.99990028	0.99989028	0.99990456
P16	0.89211791	0.86003185	0.76476191	0.91034563
P18	0.53525255	0.53306474	0.50409044	0.54414839
P19	0.79157590	0.77103232	0.74372152	0.76134248
P20	0.83366025	0.83397266	0.82830291	0.83445445
P22	0.83924665	0.83910880	0.83429783	0.84016202
P23	0.85950594	0.79375393	0.77808485	0.85468583
P24	0.69718287	0.69171208	0.68753593	0.69124231
P25	0.75802306	0.76293300	0.72051962	0.73925663
P27	0.79671784	0.80137271	0.76401735	0.79659489
P29	0.44978550	0.45828609	0.41897634	0.44168157
P30	0.56064827	0.56055748	0.54297103	0.56498333
Pearson with AHR	0.76525447	0.78809245	0.77830218	0.75649209
Spearman with AHR	0.7589320	0.823618	0.8115834	0.782249

Table VI shows the Reported Results for FUSE on the 20 pairs of sentences. The FUSE code could not be obtained so only the reported results are provided. FUSE's Pearson correlation with the human judgments of SSM values is greater than 0.631977 of the STASIS Reported Results. This result is to be expected since STASIS does not handle fuzzy words. It is, however, slightly lower than the 0.718752 of the FAST Reported results. A t-test was performed between Reported FUSE SSM values and Reported FAST SSM values to see if this difference is statistically significant. The t-test result indicates their SSM values for the two are not statistically significant with a p-value of 0.74.

. The SSM values of the Fuse Reported Results are compared to those of STASIS GeoSim using a two tailed t-test. The result showed no statistically significant difference in their SSM values with a p-value of 0.55. The SSM values of the Fuse Reported Results are also compared to those of FAST Zadeh using a two tailed t-test. The result showed no statistically significant difference in their SSM values with a p-value of 0.21.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

TABLE VI. FUSE RESULSTS

Sentence Pairs	FUSE REPORTED
P1	0.736759
P5	0.553945
P7	0.802018
P8	0.896688
P9	0.674952
P10	0.781539
P11	0.530601
P13	0.590667
P15	0.999921
P16	0.892987
P18	0.480010
P19	0.753917
P20	0.784651
P22	0.886215
P23	0.908339
P24	0.707962
P25	0.656508
P27	0.792183
P29	0.538658
P30	0.582188
Pearson with AHR	0.691786
Spearman with AHR	0.68317

V. CONCLUSIONS AND FUTURE WORK

FAST [4] developed a method to handle fuzzy words in the measurement of sentence similarity, and FUSE [5] later made enhancements for handling fuzzy words. Both of these approaches arrange fuzzy words into category ontologies and use semantic similarity measures within the ontologies to determine the similarity between fuzzy words. This paper presents a study on the use of fuzzy set similarity measures in place of semantic similarity measures within ontologies.

The results of the experiments with the modified STASIS and FAST code show that the modifications to these two to use fuzzy set similarity measures produce SSMs with correlations very close to and just as good as those correlations that STASIS and FAST produced using their semantic similarity measures. In particular, Zadeh's fuzzy set similarity measure, when used in the modified FAST code, produced both the greatest Pearson and Spearman correlations as see in Table V. Using a t-test between the two approaches for the SSM values that produced the highest correlations for each approach showed there is no statistically significant differences in the SSM values between the FAST obtained code and the FAST code modified to use fuzzy set similarity measures.

The main advantage of using fuzzy set similarity measures is that building the category ontologies in order to use the semantic similarity measure is not necessary. Fuzzy set similarity measures can be used directly on the fuzzy set representations of the fuzzy words. In the future, modifications to the FUSE code may also determine if using fuzzy set similarity measures could improve correlation with human judgment and my include an investigation into other type-2 fuzzy set similarity measures. In addition, a study of the different NLTK versions used to perform the SSM calculation may also be undertaken.

REFERENCES

[1] Y. Li, D. Mclean, Z. Bandar, J. O'Shea, K. Crockett, "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 8, pp.1138-1150, 2006.

[2] Y. Li, Z. Bandar, D. McLean, "An approach for measuring semantic similarity between words using multiple information sources". IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp.871-882, 2003.

[3] L. Zadeh, "From Computing with Numbers to Computing with Words—from Manipulation of Measurements to Manipulation of Perceptions. Logic, Thought and Action," International Journal of Applied Math. Comput. Sci., vol.12, no.3, pp. 307–324, 2002.

[4] D. Chandran, K. A. Crockett, D McLean, Z. Bandar, "FAST: A fuzzy semantic sentence similarity measure," International Conference on Fuzzy Systems, FUZZ-IEEE, 2013.

[5] N. Adel, K. A. Crockett, A. Crispin, D. Chandran, J. P. Carvalho, "FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets," International Conference on Fuzzy Systems, FUZZ-IEEE pp. 1 -8 2018:

[6] Mendel, J. "Computing with words and its relationships with fuzzistics", Information Sciences vol. 177, no. 4, pp.988-1006, 2007.

[7] V. Cross, V. Mokrenko, K. Crockett, N. Adel, "Ontological and Fuzzy Set Similarity between Perception-Based Words," International Conference on Fuzzy Systems, FUZZ-IEEE, 2019.

[8] V. Cross, An Analysis of Fuzzy Set Aggregators and Compatibility Measures, Ph.D. Dissertation, Computer Science and Engineering, March 1993, Wright State University, Dayton, OH, 264 pages.

[9] V. Cross, T. Sudkamp, "Geometric compatibility modification," Fuzzy Sets and Systems, vol. 84, no. 3, pp. 283-299, 1996.

[10] Brown Corpus Information, http://www.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

[11] Li, Yuhua & McLean, David & Bandar, Zuhair & O'Shea, James & Crockett, Keeley, "Sentence Similarity Based on Semantic Nets and Corpus Statistics," IEEE Transactions on Knowledge and Data Engineering, vol 18, pp. 1138-1150, 2006.

[12] https://www.nltk.org

[13] E. Kafe, "How Stable are WordNet Synsets?." *LDK Workshops*. 2017.

[14] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines", Bulletin de la Société Vaudoise des Sciences Naturelles, vol. 37, pp. 241–272, 1901.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:38:52 UTC from IEEE Xplore. Restrictions apply.

Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures

IEEE International Conference on Fuzzy Systems

2020

Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures

Naeemeh Adel, Keeley Crockett
School of Computing, Mathematics and Digital Technology,
Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
N.Adel@mmu.ac.uk

David Chandran
Institute of Psychiatry, Psychology & Neuroscience, Kings
College London, 16 De Crespigny Park, London,
SE5 8AF, UK
Joao P. Carvalho
INESC-ID / Instituto Superior Tecnico, Universidade de
Lisboa, Portugal

Abstract— **Dialogue systems are automated systems that interact with humans using natural language. Much work has been done on dialogue management and learning using a range of computational intelligence based approaches, however the complexity of human dialogue in different contexts still presents many challenges. The key impact of work presented in this paper is to use fuzzy semantic similarity measures embedded within a dialogue system to allow a machine to semantically comprehend human utterances in a given context and thus communicate more effectively with a human in a specific domain using natural language. To achieve this, perception based words should be understood by a machine in context of the dialogue. In this work, a simple question and answer dialogue system is implemented for a café customer satisfaction feedback survey. Both fuzzy and crisp semantic similarity measures are used within the dialogue engine to assess the accuracy and robustness of rule firing. Results from a 32 participant study, show that the fuzzy measure improves rule matching within the dialogue system by 21.88% compared with the crisp measure known as STASIS, thus providing a more natural and fluid dialogue exchange.**

Keywords— dialogue systems, conversational agents, fuzzy semantic similarity measures, fuzzy natural language

I. Introduction

Dialogue Systems (DS) are applications, which effectively replace human experts by interacting with users through natural language dialogue to provide a type of service or advice [1]. In order for a DS to engage with humans, they must be able to handle extended natural language dialogue relating to complex tasks and potentially engage in decision-making. In this sense, agents are helpful tools for human-machine interaction, allowing the input of data via natural language, processing sentences, and returning answers appropriately through text. DS, sometimes known as conversational agents, have been used in a wide range of applications such as customer service [1], help desk support [2], Educational [3,4,5,6], Cognitive Behavioural Therapy for young adults [7], insurance [8] and healthcare [9]. Dialogue understanding has become more valuable to companies with the easier ability to gain insights from unstructured text through Google's AutoML and natural language API [10], to Amazon's use of supervised machine learning to allow correct interpretation of natural language vocabulary reducing, for example, the detection of false positive responses [11]. For spoken DS, task based systems which utilise deep reinforcement learning techniques in their dialogue management systems are also becoming more available to industry [12]. What makes a successful DS is the ability for the machine to understand and interpret the human's natural language response in the context of the conversation.

Traditionally, DS used a pattern matching method to determine the most suitable response through computation of rule strengths for all matched occurrences of scripted patterns in the context of the system. The pattern matching approach has shown effectiveness and flexibility to develop extended dialogue applications [1, 13, 14] especially when coupled with ruled based matching algorithms to produce controlled responses and offer flexibility to sustain dialogues with users. However, scripting patterns is known as a laborious and time-consuming task with many flaws. More recently, some DS have opted to use short text semantic similarity measures (STSM) in place of pattern matching [6, 14, 15]. Utilising STSM within a DS is more effective than other techniques because it replaces the scripted patterns by a few natural language sentences in each rule. Evaluation of STSM based systems has been shown to improve the robustness of the system in terms of increasing the number of correctly fired rules, thus maintaining the conversational flow and increasing usability [15, 16]. However, when traditional STSM are used, they do not sufficiently match the fuzziness of natural language i.e. the human perception-based words, leading to a fundamental meaning of the human utterance in the dialogue context being misunderstood, causing incorrect firing of a rule, leading to incorrect flow of conversation and even wrong tasks being suggested.

Fuzzy Sentence Similarity Measures (FSSM) are algorithms that can compare two or more short texts or phrases which contain human perception-based words, and will return a numeric measure of similarity (composed of both semantic and syntactic elements) of meaning between them. This paper utilises one such measure known as FUSE (FUzzy Similarity mEasure) [17] which uses both WordNet [18] and a series of fuzzy ontologies which have been modelled from human representations using Interval Type-2 fuzzy sets [17]. FUSE has

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

been shown to model *intra-personal* and *inter-personal* uncertainties of fuzzy words representative of natural language.

This paper describes the creation and evaluation of a simple DS which utilises the FUSE measure to match human utterances to a set of fuzzy phrases with a rule-based system. The aim is to improve the robustness of rule matching within the DS compared with the use of a crisp similarity measure in a market research scenario where the capture of rich descriptive dialogue is important in gaining customer insight. A fuzzy DS can be used to automate the analysis of unstructured answers given to open ended questions, allowing for richer insight when collecting survey data. For example, an understanding of the dialogue, can lead to further probing to obtain more descriptive answers that provide greater insight into why a particular answer was given. This paper aims to address the following research question:

Can a Fuzzy Sentence Similarity Measure (FSSM) be incorporated into a dialogue system to improve rule matching ability from user utterance compared with a traditional STSM?

This paper is organised as follows; Section II provides a brief overview of dialogue systems and illustrates the differences between the use of traditional pattern matching and semantic similarity measures with the management of the human-machine conversation. Section III describes the design of a simple dialogue system that comprises of an FSSM, for collating human responses for evaluating customer feedback in a café and section IV describes the experimental methodology and results. Finally, section V presents the conclusions and future work.

II. Dialogue Systems

In this section, we briefly examine the dialogue engine within the DS, which is used to maintain conversational flow. We review and highlight typical problems associated with pattern matching and outline why the use of STSS overcomes some of the problems.

```
rule <tle-help-desk>
a:0.01
c:%att_name%
p:50 * something wrong * pc*
p:50 * something wrong * pc
p:50 * something wrong * computer*
p:50 * computer* * faulty*
p:50 * pc* faulty*
p:50 * computer* broken*
p:50 * pc* broken*
p:50 * computer *nt work*
p:50 * pc* *nt work*
p:50 * curing * fault * computer*
p:50 * curing * fault * pc*
p:50 * fault* * pc*
p:50 * fault* computer*
p:50 * pc * fault*
p:50 * computer * fault*
p:50 * problem * pc*
p:50 * problem * computer*
r: Please can you explain what the problem is? *<set
att_service_type PC_fault>
```

Fig. 1 Pattern matching rule

A) Strengths and Weaknesses of Pattern Matching

A dialogue system, sometimes referred to as a conversational agent (CA) is a computer program which interacts with a user through natural language dialogue and provides some form of service [1, 2, 19, 20, 21], however, they typically suffer from high maintenance in updating dialogue patterns for new scenarios due to the huge number of language patterns within the scripts. Typically DS work off scripts, which are organized into contexts, consisting of hierarchically organized rules with combining patterns and associated responses (see Figure. 1 for an example of a pattern matching rule). Scripts need to capture a wide variety of inputs and hence many rules are required, each of which deals with an input pattern and the possible variations and an associated response [5, 14, 16]. InfoChat is one such pattern matching system which utilises the sophisticated PatternScript scripting language [22] and has been adapted over the years for use in intelligent conversational tutorial systems [6]. Figure. 1 shows an example of a pattern matching rule, *<tle-help-desk>* which has been encoded using the scripting language provided with the agent InfoChat. The rule uses default values for (*a*)ctivation and (*p*)attern matching strength, has a (*c*)ondition (that the variable *att_name* has a value) and a response consisting both of a text and the setting of a variable *<set att_service_type PC_fault>*. Figure. 1 illustrates that scripting patterns is inefficient, results in domain instability and high maintenance costs. Whilst pattern matching scripting engines are a mature technology and robust, to some degree to expected user input, scripting is an art form and requires good knowledge of the language and the ability to perform in-depth knowledge engineering of the domain [1, 4, 16].

B) Semantic Similarity Measures

In a Semantic Dialogue System, each rule is matched in accordance with a pre-determined semantic similarity threshold, which is set initially through empirical evaluation and depends upon the sensitivity of rules within a context. A simple rule (Figure. 2) comprises of a set of prototypical sentences, (*s*), where the similarity with the user utterance is calculated using a STSM. Each rule has a series of responses, (*r*), which are provided to the user and can be randomly selected. Each rule also has an associated default rule, which would fire if the user utterance failed to match any prototypical sentences within the rule. O'Shea et al [15] devised a semantic scripting language which incorporated an STSS through adapting the pattern matching language of InfoChat [16] which encompasses the

```
rule <tle-help-desk>
c:%att_name%
s: There is a problem with my computer
r: Please can you explain what the problem is? *<set
att_service_type PC_fault>
```

Fig 2. Semantic rule

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

ability to extract patterns to set variables, set rule conditions and freeze, promote and demote rules.

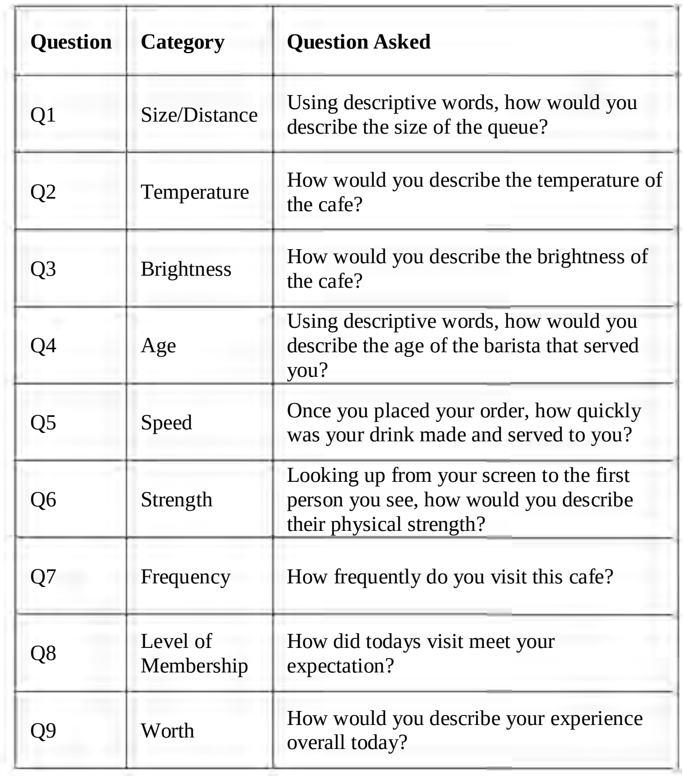
In a semantic system, prototypical sentence rules are compared with user utterances using a pre-selected STSS algorithm and the rule with the highest similarity match would fire. The most obvious benefit of using semantic rules is that no patterns are required and more importantly the semantic meaning of the utterance can be captured and acted upon within the dialogue context. Aljameel [4] used a hybrid similarity approach, combining an STSM with limited patterns, to construct an Arabic conversational intelligent tutoring system for the education of autistic children. The conversational agent processed Arabic utterances using a novel crisp STSM which utilised the cosine similarity measure to solve the word order issue associated with the Arabic language. Consequently, this reduced the number of scripts and rules required. Through empirical evaluation of two versions of the system, the use of an STSM reduced the number of unrecognised human utterances to 5.4% compared to 38% in the pattern scripted version and, hence, the systems incorrect responses were reduced to 3.6% compared to 10.2% in the pattern scripted version [4]. Similar improvements on the benefits of utilising a STSM within DS are also reported in [23]. In this paper, we will replace the traditional semantic similarity measure with a Fuzzy semantic similarity measure to evaluate the effectiveness of a DS through a reduction in the incorrect responses and unrecognised human utterances compared with using an STSM.

III. A SIMPLE DIALOGUE SYSTEM FOR COLLATING USER RESPONSES

A) Overview

In this section, we describe a simple question and answer dialogue system that utilises the FUSE semantic similarity measure [17], to match user utterances to different categories of responses to each question. The dialog structure is therefore a linear sequence of questions, where each question response has three possible branches. The aim is to distinguish between human perceptions of fuzzy words in nine categories to assess if the correct rule fires in response to natural language used within the human utterance. FUSE [17] is an ontology based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception based words. The FUSE algorithm identifies fuzzy words in a human utterance and determines their similarity in context of both the semantic and syntactic construction of the sentence. Currently FUSE consists of nine fuzzy categories each containing a series of fuzzy words. These categories are Size/Distance, Age, Temperature, Worth, Level of Membership, Frequency, Brightness, Strength and Speed. Initial selection and methodology for word population can be found in [17]. An experiment originally described in [17] was used to capture human ratings to create the fuzzy ontology for these categories where words were modelled based on Mendel's Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets [24]. A full description of the FUSE algorithm and the general approach on how the fuzzy word models and measures in each category were derived is given in [17].

TABLE I: CAFÉ FEEDBACK DIALOGUE QUESTIONS MAPPED TO FUZZY CATEGORIES



Question	Category	Question Asked
Q1	Size/Distance	Using descriptive words, how would you describe the size of the queue?
Q2	Temperature	How would you describe the temperature of the cafe?
Q3	Brightness	How would you describe the brightness of the cafe?
Q4	Age	Using descriptive words, how would you describe the age of the barista that served you?
Q5	Speed	Once you placed your order, how quickly was your drink made and served to you?
Q6	Strength	Looking up from your screen to the first person you see, how would you describe their physical strength?
Q7	Frequency	How frequently do you visit this cafe?
Q8	Level of Membership	How did todays visit meet your expectation?
Q9	Worth	How would you describe your experience overall today?

B) Design of a Dialogue System for Café Feedback

In order to establish if a FSSM could be used in a dialogue system, a simple question and answer system was designed to obtain feedback from participants who visited a local café. This was done using a knowledge engineering approach and involved gathering information about typical questions asked in a customer satisfaction online questionnaire concerning customer satisfaction levels in high street cafes. Existing survey questions were either a mixture of dichotomous questions, multiple choice, Likert scale questions or free text. Within the proposed Café feedback DS, each question selected had to be transformed into one which would allow the user to provide descriptive textual answers in order to gather as much data as possible to evaluate the impact of the fuzzy semantic measure. To ensure all the categories in FUSE were covered, nine questions where created (Table I), each one covering responses that would contain words or synonyms of words from each fuzzy category. Each question formulates a question-rule within the DS where each rule can have three responses which represent full coverage of the categories as defuzzified word values obtained through human experts and Type-II modelling using HMA approach [17].

The rule responses were divided into three thresholds of high, medium and low, and words (and word synonyms) within each category would fall under each threshold. The threshold for each category varies as the number of words and measurements in each category varies (dependent on human perceptions [17]). The thresholds in each of the nine categories were selected based

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

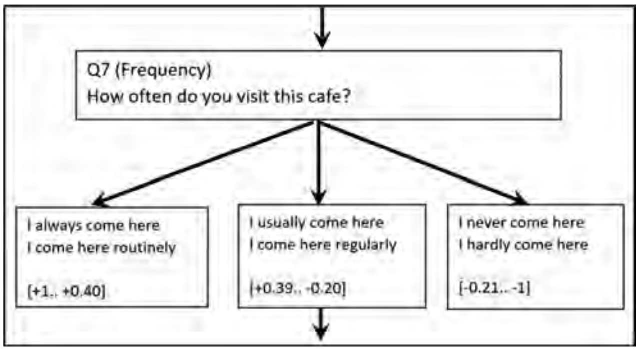


Fig 3. Frequency threshold

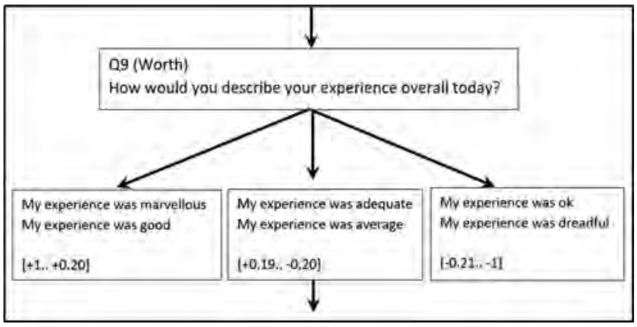

Q9 (Worth)
How would you describe your experience overall today?
My experience was marvellous
My experience was good
[+1.. +0.20]
My experience was adequate
My experience was average
[+0.19.. -0.20]
My experience was ok
My experience was dreadful
[-0.21.. -1]

Fig 4. Worth threshold

on the words in that specific category. An example is shown in Figures 3 and 4 for the two categories of *Frequency* and *Worth.*

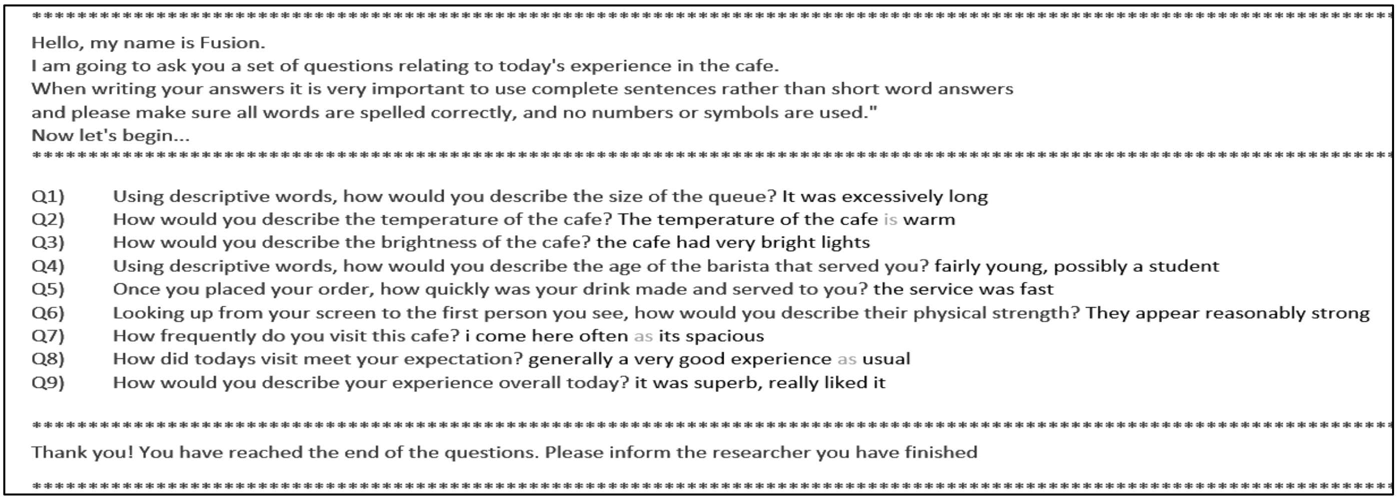
Considering Figure. 3, for the category *Frequency*, the low threshold begins at [-1] and ends at [+0.40], with the last word to fall in this threshold being *Everytime*, and the next word after this which begins the mid threshold is *Occasionally* at [+0.39], and this threshold continues up to [-0.20], and even though this is now a negative value, it still falls in the mid threshold for this category, and the low threshold starts at [-0.21] and ends at [-1]. Examining Figure. 4 for category *Worth*, the high threshold starts at [+1] and ends at [+0.20], the mid threshold begins at [+0.19] and ends at [-0.20], and the low threshold begins at [-0.21] and ends at [-1]; thus there was not a single fixed threshold for all nine categories, as the words and there values varied in each category. In order to determine the specific high, medium and low thresholds for each fuzzy category, two English language experts independently grouped the words for each category. In the case of disagreement, a third expert was asked to cast the deciding vote.

```
<default-rule1><size/distance>
s: It was long
s: It was huge
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-high>
c: temperature_context

<Default-rule2><size/distance>
s: It was average
s: It was regular
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-medium>
c: temperature_context

<Default-rule3><size/distance>
s: It was tiny
s: It was small
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-low>
c: temperature_context
```

Fig 5. Sample Rules for Size/ distance category

C) Scripting

Each question (Table I) was scripted into a context which represented a category. Three English prototypical sentences were used in each rule to enable coverage of either the high, medium or the low thresholds. In addition, there were initialisation and conclusion contexts. Figure. 5 shows three rules (*r*) from the *Size/Distance* category. Each dialogue exchange between human and machine generated a human utterance that was compared to the prototypical sentences in each rule. In each context, the rule where the (*s*)sentence gave the highest similarity score compared with the human utterance, was analysed and fired through FUSE. An attribute is set i.e. *att_size-distance-high* becomes true if *default-rule1* fires and a

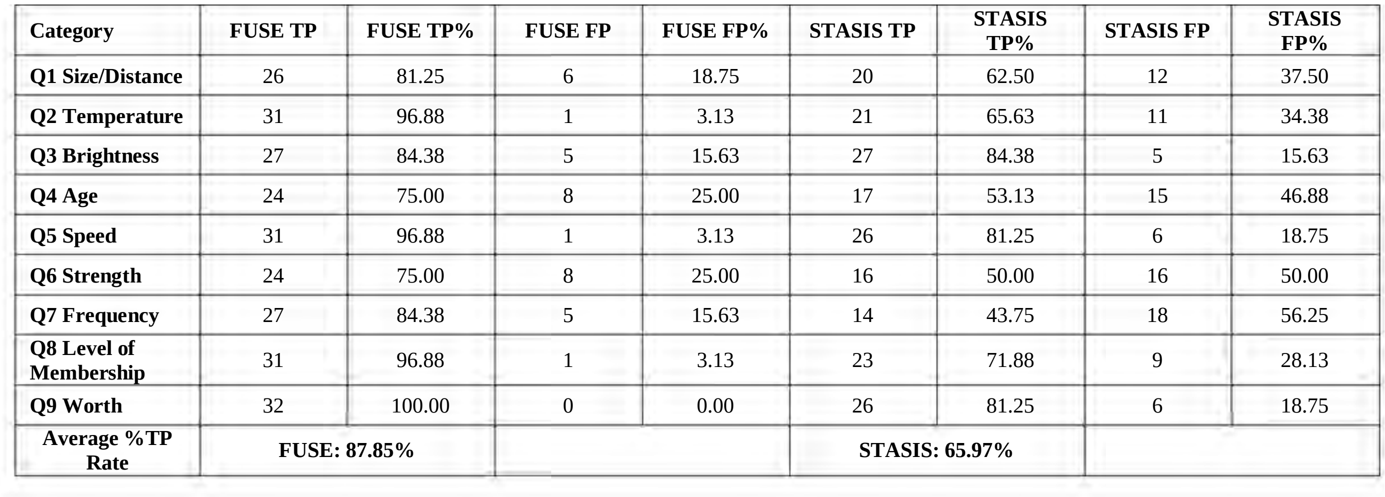


```
****************************************************************************************************
Hello, my name is Fusion.
I am going to ask you a set of questions relating to today's experience in the cafe.
When writing your answers it is very important to use complete sentences rather than short word answers
and please make sure all words are spelled correctly, and no numbers or symbols are used."
Now let's begin...
****************************************************************************************************

Q1)    Using descriptive words, how would you describe the size of the queue? It was excessively long
Q2)    How would you describe the temperature of the cafe? The temperature of the cafe is warm
Q3)    How would you describe the brightness of the cafe? the cafe had very bright lights
Q4)    Using descriptive words, how would you describe the age of the barista that served you? fairly young, possibly a student
Q5)    Once you placed your order, how quickly was your drink made and served to you? the service was fast
Q6)    Looking up from your screen to the first person you see, how would you describe their physical strength? They appear reasonably strong
Q7)    How frequently do you visit this cafe? i come here often as its spacious
Q8)    How did todays visit meet your expectation? generally a very good experience as usual
Q9)    How would you describe your experience overall today? it was superb, really liked it

****************************************************************************************************
Thank you! You have reached the end of the questions. Please inform the researcher you have finished
****************************************************************************************************
```

Fig 6. Simple Interface Design

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

TABLE II: RESULTS OF FUSION DS WITH FUSE VS STASIS SSM



Category	FUSE TP	FUSE TP%	FUSE FP	FUSE FP%	STASIS TP	STASIS TP%	STASIS FP	STASIS FP%
Q1 Size/Distance	26	81.25	6	18.75	20	62.50	12	37.50
Q2 Temperature	31	96.88	1	3.13	21	65.63	11	34.38
Q3 Brightness	27	84.38	5	15.63	27	84.38	5	15.63
Q4 Age	24	75.00	8	25.00	17	53.13	15	46.88
Q5 Speed	31	96.88	1	3.13	26	81.25	6	18.75
Q6 Strength	24	75.00	8	25.00	16	50.00	16	50.00
Q7 Frequency	27	84.38	5	15.63	14	43.75	18	56.25
Q8 Level of Membership	31	96.88	1	3.13	23	71.88	9	28.13
Q9 Worth	32	100.00	0	0.00	26	81.25	6	18.75
Average %TP Rate	**FUSE: 87.85%**				**STASIS: 65.97%**			

change in context will occur, denoted by the '*c:*' identifier. As this is a simple linear DS the change in context is always set to the context of the next question until all questions have been asked. Figure. 6 shows an example of a participants answers.

On initiation of the system, the DS begins with the simple message:

"Hello, My name is Fusion. I am going to ask you a set of questions relating to today's experience in the cafe. When writing your answers it is very important to use complete sentences rather than short word answers and please make sure all words are spelled correctly, and no numbers or symbols are used. Now let's begin...".

After all questions were asked the final message was "*Thank you! You have reached the end of the questions. Please inform the researcher you have finished.*"

A log file recorded all dialogue, including the semantic similarity score for each rule during the completion of the survey. In this version of the system, all human utterances were recorded, with incorrect utterances failing to match any rules in each context also being recorded.

IV. EXPERIMENTAL DESIGN

A) Experimental Methodology

Following Manchester Metropolitan Universities ethical approval process (Ethos number: 11759), 32 participants were recruited through an advertising campaign through the University. After agreeing to take part, and agreeing a suitable time, participants were given a voucher to purchase a drink at one of two cafes within the University. On purchasing a beverage, the participant was asked to sit down and observe their environment for 10-15 minutes. Once finished, they notified the researcher (who was sat independently) and began to complete the café feedback survey using the DS about their experience and visit to the café. During this interaction, the typed user utterances for each answer is run through the DS and compared with the thresholds for the corresponding category. For analysis purposes, each user utterance was taken and compared with the two sentences for each of the high, medium and low threshold sentences. The similarity is calculated for each sentence pair using FUSE and the results are recorded and the highest similarity rating is noted for each interaction. All dialogue exchanges are recorded in a log for analysis. Once completed, the participants completed a short usability questionnaire, with questions comparable to those used to typically assess usability of DS [25, 26].

To analyse the results, a dataset consisting of 288 rows was compiled of all user responses to all questions, along with the semantic similarity measurement for each rule calculated using FUSE. For comparison purposes, the same rules and responses were also fired through a well-established similarity measurement known as STASIS [27]. STASIS is not able to capture the meaning of fuzzy words. STASIS only caters for crisp values and uses WordNet and Browns Corpus to find similarity rating for sentence pairs [27].

B) Results

Table II shows the results from all 32 participants for the TP and FP values run for both FUSE and STASIS and shows the percentage of correct TP for FUSE compared with that of STASIS. The fuzzy words assigned to each of the thresholds are examined and if the DS has picked up the correct sentence match then this is counted as a True Positive (TP) and given a score of 1. If the highest similarity rating has not fallen under the correct threshold of words, then it is classed as a False Positive (FP) and given a score of 0.

As can be seen from the results in Table II, FUSE has an average TP rating of 87.85% and STASIS has an average TP rating of only 65.97%. The average TP rating represents the total number of correctly fired rules that are also correctly matched with the user utterances and are therefore a true positive. These results show that the fuzzy dictionary of words modelled within the FUSE categories increases the similarity

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

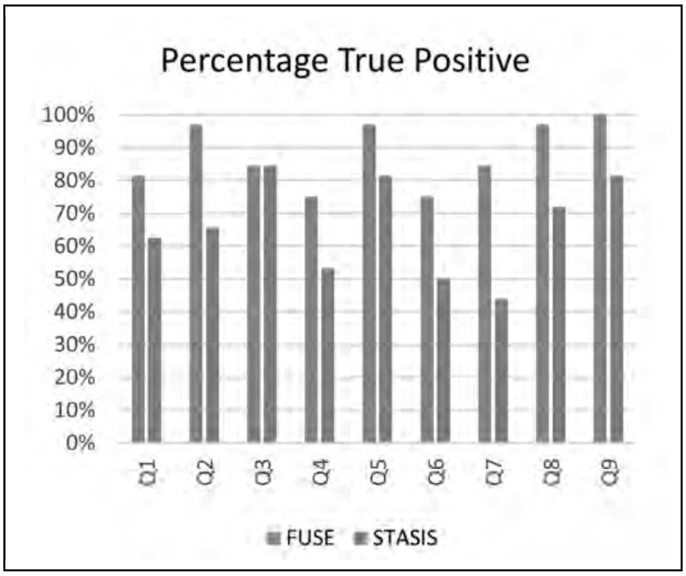


Fig 7. Percentage of TP values for FUSE vs STASIS

rating when compared with that of human utterances as opposed to just crisp values.

Figure. 7 shows the percentage of correctly matched user utterances using FUSE and STASIS. Each question is designed to represent a separate category for comparison purposes, therefore even though STASIS does not have a fuzzy dictionary and only uses WordNet it can still be used in this scenario to compare the effects of fuzzy words vs crisp values. It can be seen in Figure. 7 that for all the nine categories, with the exception of *Brightness* (Q3), FUSE always has a higher TP rating then STASIS, meaning it has a higher number of true positive matches that fired under the correct threshold. For Q3 (*Brightness*), both FUSE and STASIS scored the same, meaning they both fired the same correct thresholds.

C) Discussion

Overall, the results have shown that a DS that utilises the FUSE measure to determine which rule fires, provides a higher average TP rating using fuzzy words as opposed to STASIS that only uses crisp values. There was an improvement of 21.88% in the average TP rating as can be seen in Table II when compared with STASIS, where fuzzy words are not taken into consideration. There were however, some rules that did not fire correctly and this section provides some in-depth analysis of those rules to feed into future work on the system.

In total, 8 (out of 288) of the user utterances contained some numerical responses as well as just words; an example is shown below of an instance where the DS asked the question relating to the category *Age*:

Q4) Using descriptive words, how would you describe the age of the barista that served you?

User Utterance: The physical appearance of the barista tells that she was in her 30's

Both FUSE and STASIS picked this up as belonging to the low category, consisting of words such as *baby, young, child*, etc; when according to the two English language experts, it should be in the mid threshold containing words such as *adult*, *middle-aged*, *grownup* etc. On the other hand, when the DS asked the question relating to the category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of the queue?

User Utterance: The size of the queue was 2-3 people long with a wait time of no longer than 1 minute.

Both FUSE and STASIS picked this up as being in the mid threshold, containing words such as *average*, *standard*, *middle*, and the two English language experts agreed that this can be classed as a TP and is in the correct threshold.

Neither FUSE nor STASIS was able to deal with the effect of the inclusion of negation words within utterances. For example, when the DS asked the question relating to the category *Brightness*:

Q3) How would you describe the brightness of the cafe?

User Utterance: The light level of the cafe is not bright

Both FUSE and STASIS picked this up as the high threshold because of the word *bright*, when in effect due to the use of the word *not*, it actually means it was dark. Therefore is this case, the correct rule category did not fire (i.e. *bright* was identified as being in the high threshold by the English language experts, but the presence of the word *not* would contradict this and it should be the in the low threshold).

An additional example of negations leading to an incorrect rule firing was when the DS asked the question relating to the category *Strength*:

Q6) Looking up from your screen to the first person you see, how would you describe their physical strength?

User Utterance: I would describe them as lean and not very strong.

Both FUSE and STASIS picked this up as belonging to the high threshold due to the word *strong (*and had an increased intensity in FUSE to the hedge word very*)*, when in fact because of the use of the word *not* it actually should belong to the low or mid thresholds and this was also confirmed by the two English language experts.

There were some instances where FUSE correctly matched a rule and STASIS did not. One example of this is when the DS asked the question relating to the category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of the queue?

User Utterance: The size of the queue was huge.

FUSE picked this up as belonging to the high threshold with a similarity value of ((D1) It was long: 0.57554), and STASIS picked this up as belonging to the low threshold, with a similarity value of ((D3) It was small: 0.53459). The high threshold is correct, since it holds words such as *big*, *massive* and *huge.* Although the difference in the two similarity ratings are small, it is down to the fact that the high threshold actually

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

holds the word *huge* therefore this is the threshold it must fall under for it to be a TP [17].

An instance when STASIS correctly matched a rule and FUSE did not is when the DS asked the question relating to the category *Brightness*:

Q3) How would you describe the brightness of the cafe?

User Utterance: It was fairly bright

STASIS picked this up as belonging to the high threshold with a similarity value of ((D1) The cafe was bright: 0.36442), and FUSE picked this up as belonging to the mid threshold with a similarity value of ((D2) The cafe was luminous: 0.67367). The high threshold is correct as it holds words such as *sunny*, *radiant* and *bright*.

D) Effect on Usability

All participants completed a short usability survey comprising of 13 Likert scale questions with allowable free text, following completion of the task. A full in-depth usability analysis is beyond the scope of this paper, but it is important to highlight that the inclusion of a FSSM into the DS did not appear to negatively affect the usability of the system. In summary, 94% agreed or strongly agreed that a DS could be used as a mechanism to answer survey questions in the future. 90% of participants reported no inconsistences when using the system and 91% found the system easy to interact with and intuitive to use.

V. Conclusion And Further Work

This paper has described the development of a simple linear DS that incorporated the FUSE semantic similarity algorithm. The semantic similarity of user utterances and rules was compared using both FUSE and STASIS in order to determine which of the three rules in each category would fire. The results show that the average TP of FUSE is 87.85% which is an improvement of 21.88% when compared with STASIS rule firing rating (65.97%). Given the original research question, we conclude that a Fuzzy Sentence Similarity Measure (FSSM) can be incorporated into a dialogue system to improve rule matching ability from a user utterance compared with a traditional STSM. A weakness of utilising FUSE was its inability to deal with the word “Not” within the dialogue, which caused misfiring of rules. Future work will address this issue by looking at ways to apply the fuzzy NOT operator to the associated word.

Despite the simplicity of the DS, a number of issues have been recognised. Firstly, neither measure (STASIS or FUSE) were able to produce correct rule firings when a negation word was used to form part of the utterance. All though hedges had been considered as an addition to the FUSE fuzzy dictionary [17], negation words were not included in the similarity calculation within FUSE. Secondly, FUSE is very much dependent on the fuzzy dictionary created in previous work, which were generated from many empirical experiments [17] where humans rated words within categories and then within the context of general sentences. In this paper, it is clear that the context of perception-based words does matter when used by a FSSM in a DS. Further work will include the evaluation of a second, more substantial prototype DS, which will incorporate other fuzzy similarity measures [28] and revisit the impact of hedge words.

Acknowledgment

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

References

[1] J. D. O’Shea, Z. Bandar, K. Crockett, “Systems Engineering and Conversational Agents”, In *Intelligence-Based Systems Engineering*, Intelligent Systems Reference Library, Springer, Berlin, Heidelberg, 2011, vol. 10, pp. 201-232.

[2] L. Ozaeta, M. Graña, 2018. A View of the State of the Art of Dialogue Systems. In: de Cos Juez F. et al. (eds) *Hybrid Artificial Intelligent Systems*. *HAIS*. 2018. Lecture Notes in Computer Science, vol. 10870. Springer, Cham https://doi.org/10.1007/978-3-319-92639-1_59

[3] L. Lin, P. Ginns, T. Wang, P. Zhang, 2020. Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain?. *Computers & Education*, 143, p.103658.

[4] S.S. Aljameel, “Development of an Arabic conversational intelligent tutoring system for education of children with autism spectrum disorder”, PhD dissertation, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2018.

[5] S.S. Aljameel, J.D. O’Shea, K. Crockett, A. Latham, and M Kaleem, 2019. LANA-I: an Arabic conversational intelligent tutoring system for children with ASD. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 498-516. Springer, Cham.

[6] A. Latham, K. Crockett, D. McLean, 2014. An adaptation algorithm for an intelligent natural language tutoring system. *Computers & Education*, 71, pp. 97-110.

[7] K.K. Fitzpatrick, A. Darcy, M. Vierhile, 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, vol. 4, no. 2, p.e19.

[8] F. Koetter, M. Blohm, J. Drawehn, M. Kochanowski, J. Goetzer, D. Graziotin and S. Wagner, 2019, February. Conversational Agents for Insurance Companies: From Theory to Practice. In *International Conference on Agents and Artificial Intelligence*, pp. 338-362. Springer, Cham.

[9] J.L.Z. Montenegro, C.A. da Costa, R. da Rosa Righi, 2019. Survey of Conversational Agents in Health. *Expert Systems with Applications*, vol. 129, pp. 56-67, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.03.054.

[10] Google Natural Language, Jan. 01, 2020. [Online]. Available: https://cloud.google.com/natural-language/#overview. [Accessed Jan. 3, 2020].

[11] Alexa and Alexa Device FAQs, Feb. 23, 2016. [Online] Available: https://www.amazon.com/gp/help/customer/display.html?tag=skim1x169757-20&nodeId=201602230. [Accessed Jan. 3, 2020].

[12] J. He, B. Wang, M. Mingming Fu, T. Yang and X. Zhao, Hierarchical attention and knowledge matching networks with information enhancement for end-to-end task-oriented dialog systems, *IEEE Access*, vol. 7, pp. 18871–18883, 2019.

[13] R.R.A. Pazos, B.J.J. González, L.M.A. Aguirre, F.J.A. Martínez and H.H.J Fraire, 2013. Natural Language Interfaces to Databases: An Analysis of the State of the Art. In: *Recent Advances on Hybrid Intelligent Systems*, O. Castillo, P. Melin and J. Kacprzyk, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 463-480.

[14] J.D. O'Shea, “A framework for applying short text semantic similarity in goal-oriented conversational agents”, PhD dissertation, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2010.

[15] K. O'Shea, K. Crockett, Z. Bandar, J.D. O'Shea, Erratum: An approach to conversational agent design using semantic sentence

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

similarity (Appl Intelligence) *Applied Intelligence*, vol. 40, no. 1, pp. 199-199, 2014.

[16] C. Curry, "A framework for developing a conversational agent to improve normal age-associated memory loss and increase subjective wellbeing", PhD dissertation, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2019.

[17] N. Adel, K. Crockett, A. Crispin, D. Chandran and J.P. Carvalho, Jul. 2018, FUSE (Fuzzy Similarity Measure)-A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets. In *2018 IEEE International Conference on Fuzzy Systems* (*FUZZ-IEEE*) pp. 1-8, IEEE.

[18] Princeton University, "About Wordnet". [Online]. Available: http://wordnet.princeton.edu/ [Accessed Jun. 13, 2014].

[19] J.G. Harms, P. Kucherbaev, A. Bozzon and G.J. Houben. 2018. Approaches for Dialog Management in Conversational Agents. *IEEE Internet Computing*, vol. 23, no. 2, pp.13-22.

[20] J.B. Aujogue, A. Aussem, 2019. Hierarchical Recurrent Attention Networks for Context-Aware Education Chatbots. In *2019 International Joint Conference on Neural Networks* (*IJCNN*) pp. 1-8, IEEE.

[21] J. Lester, K. Branting, B. Mott, "Conversational Agents". CRC Press LLC. [Online]. Available: https://www.ida.liu.se/~729A15/mtrl/Lester_et_al.pdf [Accessed Jun. 16, 2015].

[22] D. Michie, C. Sammut, Infochat Scripter's Manual, Convagent Ltd, Manchester, UK, 2001.

[23] M. Kaleem, J.D. O'Shea, K. Crockett, 2014. Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents. In *2014 14th UK Workshop on Computational Intelligence* (*UKCI*) pp. 1-8: IEEE, ISBN: 978-1-4799-5538-1, DOI: 10.1109/UKCI.2014.6930180.

[24] M. Hao, J.M., Mendel, 2015. Encoding words into normal interval type-2 fuzzy sets: HM approach, *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp. 865-879.

[25] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre and M. Cieliebak, 2019. Survey on Evaluation Methods for Dialogue Systems. *arXiv preprint arXiv:1905.04071.*

[26] X. Chen, J. Mi, M. Jia, Y. Han, M. Zhou, T. Wu and D. Guan, October 2019. Chat with Smart Conversational Agents: How to Evaluate Chat Experience in Smart Home. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1-6, ACM.

[27] Y. Li, D. McLean, Z. Bandar, J.D. O'Shea and K. Crockett, 2006. Sentence similarity based on semantic nets and corpus statistics, *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138-1150.

[28] V. Cross, V. Morenko, K. Crockett and N Adel, 2019, June. Ontological and fuzzy set similarity between perception-based words. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* pp. 1-6, IEEE.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:33:14 UTC from IEEE Xplore. Restrictions apply.

Ontological and Fuzzy Set Similarity between Perception-Based Words

IEEE International Conference on Fuzzy Systems

2019

Ontological and Fuzzy Set Similarity between Perception-Based Words

Valerie Cross
Computer Science and Software Engineering
Miami University
Oxford, OH USA
crossv@miamioh.edu

Valeria Morenko
Computer Science and Software Engineering
Miami University
Oxford, OH USA
mokrenvi@miamioh.edu

Keeley Crockett
Computational Intelligence Lab
Manchester Metropolitan University
Manchester, UK
K.Crockett@mmu.ac.uk

Naeemeh Adel
School of Computing, Maths, and Digital Technology
Manchester Metropolitan University
Manchester, UK
N.Adel@mmu.ac.uk

***Abstract*—Fuzzy short text semantic similarity measures allow the inclusion of human perception based words to be within the similarity measurement which results in better correlation on the meaning of the short text with human understanding. Existing measures such as FUSE and FAST rely on the creation of fuzzy ontological structures from the modelling of perception words using type-1 or type-2 fuzzy sets. Due to the complex methodology of creating these ontologies, fuzzy word representation cannot be guaranteed due to language evolution. This paper presents a comparative study of simpler fuzzy set similarity measures. The results surprisingly indicate that a very simple fuzzy set similarity measure created from the center of gravity (COG) distance between type-2 fuzzy sets has a very high correlation with the FUSE semantic similarity measure.**

Keywords—ontology, semantic similarity, fuzzy set similarity, human perception

I. Introduction

A goal of artificial intelligence is to develop machines that communicate and understand natural language. Communication between machines uses crisp quantities, but an important characteristic of natural language is many words are vague or imprecise. Vagueness often exists in domain knowledge as understood by humans. Often humans communicating with each other or providing domain knowledge are more comfortable using inexact, vague terms, or perception-based, that is, fuzzy words that are subjective. For humans and machines to communicate and for machines to understand domain knowledge, a method of interpreting fuzzy words is needed. Computing with Words (CWW) [1] provides the ability to interpret these fuzzy words. Fuzzy set theory and CWW research presents essential concepts necessary to make progress towards the goal of finding representations of natural language or fuzzy words used by humans and reasoning with these representations.

Handlubg uncertainty in human language has motivated the natural language processing research community to develop sentence similarity measures. Early work focused on syntactic similarity [2]. Latent Semantic Analysis [3] brought in semantic similarity between blocks of text by producing statistics based on occurrences of the words in the blocks within a large corpus. Using statistical analysis, LSA creates semantic vectors. It calculates similarity between these vectors. Following this, STASIS [4] examined the use of semantic similarity measures within the context of an ontology, a knowledge structure containing concepts and defining relationships between these concepts. Much research exists on semantic similarity measures, also referred to as ontological similarity measures [5], between concepts in an ontology. For measuring text similarity for short pieces of text, STASIS uses the WordNet ontology and a semantic similarity measure [6] between each word pair, one word from each text, to create a semantic vector and incorporates corpus statistics in the semantic vector. STASIS integrates the early approach of measuring syntactic similarity into the final similarity measure between two pieces of text.

Although this previous research made progress in measuring text similarity, it failed to address the occurrence of imprecise and vague words, i.e., fuzzy words that occur extensively in natural language. This capability is needed in order to advance conversational understanding between humans and machines. Since different people have different interpretations or meanings for fuzzy words, singular quantities for them are not reasonable. Fuzzy sets serve as a means of representing fuzzy words. CWW provides a framework by which fuzzy words can be quantified, scaled against each other and then become machine representable. The scaling of fuzzy words through obtaining human perceptions is a critical step for creating fuzzy sentence similarity measures.

FAST (Fuzzy Algorithm for Similarity Testing) [7] was developed to measure the similarity between pairs of fuzzy words and incorporate this additional similarity evaluation into the overall sentence similarity measure between sentences or pieces of short text. To accomplish this, it was necessary to create a dataset containing quantified fuzzy words which are organized into six different categories [7]: *age, size/distance, frequency, goodness, membership level* and *temperature*. In a comparative experimental study, FAST demonstrated an improvement in measuring semantic sentence similarity over existing algorithms STASIS and LSA, which are unable to process fuzzy words in text.

978-1-5386-1728-1/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:47 UTC from IEEE Xplore. Restrictions apply.

More recent research developed FUSE (FUzzy Similarity mEasure) [8] which extends the FAST research to address the differences between modeling fuzzy words with type-1 versus type-2 fuzzy sets. In FAST, human experts were used to create type-1 fuzzy sets for the fuzzy words; however, on further consideration, it was felt that these fuzzy sets were not accurate representations because of the subjective nature of the human evaluators. Essentially, type-1 fuzzy sets could not capture the uncertainty of humans [9]. FUSE uses specifically type-2 interval fuzzy sets since they are simpler to use because the membership functions are interval sets. FUSE also has a larger vocabulary across the six categories with over 57% increased coverage of fuzzy words. Both FUSE and FAST, however rely on pre-constructed fuzzy ontologies, resulting in complex measures, which will not perform well if there is not extensive modelling of fuzzy words for any given language.

This paper focuses on the measurement of similarity between fuzzy words represented as type-1 fuzzy sets using three different existing fuzzy set similarity measures. These fuzzy sets are directly created from the data collected from the human evaluators. This approach is simpler than that of FAST and FUSE for measuring fuzzy word similarity. Because type-2 fuzzy sets may better represent the subjective nature of a fuzzy word and are used in FUSE, a fourth similarity measure using a scaled COG for the type-2 fuzzy word representations is also used in our study. The objective is to determine how well these simpler fuzzy set similarity measures correlate with the semantic similarity measure used in FAST and FUSE.

The paper organization is as follows: Section II first examines some of the difficulties when using humans to gather data for the process of defining fuzzy words as fuzzy sets and describes the approaches to representing fuzzy words to measure similarity between them. Section III describes the approach for fuzzy word representation used in this paper's research. It reviews the existing fuzzy set similarity measures and a simple similarity measure calculated from the distance between the COGs for two fuzzy words represented as type-2 interval fuzzy sets. Section IV describes the experimental design and compares the results from applying these measures to word pairs used in previous studies [4] [7] [8]. Finally, Section V presents the conclusions and future work.

II. Context of Fuzzy Words and their Representation

A. Type-1 versus Type-2 Fuzzy Sets

In [1] a fuzzy set (type-1) representation is described as a means of defining perception-based or fuzzy words. Type-2 fuzzy sets [9] were developed to address the issue of perception-based words varying from individual to individual. Instead of using a single fuzzy set, a set of fuzzy sets represents a fuzzy word; that is, a type-2 fuzzy set is a set wherein all its elements are fuzzy type-1 sets. In FAST, type-1 fuzzy sets are developed for fuzzy words but in FUSE type-2 interval fuzzy sets are used. In both of these approaches ontologies are created to represent the relationships among the fuzzy words in six different categories. The six categories are broad enough to hold a large range of fuzzy word and allow related fuzzy words to be scaled in terms of association within the category. These ontologies are created by scaling a representative value of the fuzzy set into the interval [-1, +1]. The scaled value determines into which node of the ontology the fuzzy word is placed.

B. Creating the fuzzy word representation for FAST

Two empirical experiments were undertaken with human subjects. The first required the subjects to populate the six categories with fuzzy words. Next subjects had to quantify the fuzzy words in each category. Quantification was done using a scale of 0 to 10. The subjects were asked to specify a single value, a point in the 0 to 10 scale where the membership function for a fuzzy word would be highest. For each fuzzy word the mean and standard deviation values were calculated from all the subjects' ratings for that fuzzy word. Then the relationships among the fuzzy words within a category were established by creating ontologies based on these values. These ontologies of fuzzy words are needed since the semantic similarity measure used between two fuzzy words is that in [6]. Although numerous semantic similarity measures have been proposed over the years [5], this research focuses on the specific measure used in the FAST and FUSE research which addresses some of the weaknesses of the older semantic similarity measures. The formula for the semantic similarity measure, *S* used to determine word pair similarity of words, w_1 and w_2 is

$$\boxed{S(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}}$$

where *l* represents the path length between the two words in the ontology and *h* represents the depth of their common subsumer. For FAST, the parameters α and β were set to 0.2 and 0.6, respectively and were determined empirically.

To create the category ontologies, five nodes were established for each category. The root node for each category contains those fuzzy words whose mean values were around the midpoint value (within the 0 to 10 range). This root node is labeled AVERAGE for each category. As an example, for the *size/distance* category, the five nodes are labeled {VERY SMALL, SMALL AVERAGE, LARGE, VERY LARGE}. Examples of fuzzy words in its root node include *medium* and *middle.* From the root node there are two branches. To the left are two nodes for the fuzzy words with lower mean values, {VERY SMALL, SMALL}. To the right are two nodes for the fuzzy words with higher mean values, {LARGE, VERY LARGE}. To place the fuzzy words in the appropriate nodes, the mean values were re-scaled to a range of -1 to +1 and then a range of re-scaled values was established for each node and used to determine to which node a fuzzy word should be assigned. Each category ontology was created in this manner; for example, the *temperature* category has the nodes {VERY COLD, COLD, AVERAGE, HOT, VERY HOT}. FAST uses the created category ontologies with the semantic similarity measure in [6] to determine the similarity between pairs of fuzzy words. This word pair similarity measurement is one component of the FAST algorithm that establishes a measure of text similarity between pairs of sentences or pieces of text.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:47 UTC from IEEE Xplore. Restrictions apply.

C. Creating the fuzzy word representation for FUSE

FUSE takes a similar approach to FAST in that in creates ontologies based on the six categories and the fuzzy words within those categories; however, it expanded on the number of fuzzy words since FAST had only 196 words within the six categories. It did this by taking the existing FAST words and adding only the one word synonyms for these words that could be found in a dictionary. This process resulted in a total of 309 fuzzy words over the six categories.

As in FAST, human subjects are used to construct the fuzzy sets for the fuzzy wordsT. hese fuzzy sets are based on Mendel's Hao-Mendel Approach (HMA) using type-2 interval fuzzy sets [13] to collect data from the subjects. The same 0 to 10 range is kept. The subjects are asked to provide an interval value for the fuzzy word instead of a single value as in FAST. This interval value represents the range where the subject believes the fuzzy word should be placed in the range of 0 to 10. Noise is eliminated by removing bad data and outliers.

From the cleaned up data, the center of gravity (COG) was determined using the upper and lower footprints of uncertainty. As in FAST, the COG value for a fuzzy word was scaled into the -1 to +1 range in order to create the ontology. FUSE, however, increased the number of nodes for a category ontology from 5 to 11 and the root node was an arbitrary category label node. The ontology became a binary tree with nodes containing negative values on the left side of the root node and nodes containing positive values on the right side. The fuzzy words were grouped using a 0.2 interval size. As in FAST, the similarity measure given in [6] was used with these category ontologies to determine semantic similarity between pairs of fuzzy words. The parameters α and β for FUSE were determined empirically and set to 0.15 and 0.85, respectively.

III. Fuzzy Set Similarity Measure Between Fuzzy Words

The approaches to measuring fuzzy word similarity in STASIS, FAST and FUSE have as their basis semantic or ontological similarity measures within an ontology structure. A detailed review of semantic similarity measures can be found in [5]. The FAST and FUSE approaches require creating ontologies for each of the six categories so that a semantic similarity measure can be used between the fuzzy words. The approach used in our research does not require creating ontologies, Instead three fuzzy set similarity measures are used between triangular fuzzy sets created from the FAST type-1 fuzzy sets. The fourth similarity measure uses the distance between the normalized centers of gravity (COG) for type-2 interval fuzzy sets created for FUSE.

A. Creation of Trianglar Fuzzy Sets

For purposes of the FAST experiments data from the type-1 fuzzy sets were acquired from the FAST researchers, specifically the defuzzified value or mean and the standard deviation. With these values, a pseudo triangular fuzzy set is created where the membership degree at the mean value is 1.0. A normal probability density distribution is used and values ±3 standard deviations away from the mean were used for the end points of the triangular fuzzy set since 99.7% of the data is within three standard deviations of the mean. See Fig. 1 that shows the triangular membership function for *centre* with a mean of 4.93 and a standard deviation of 0.5. The simplest approach to building fuzzy sets for fuzz words is used since the hypothesis is to determine if these sets based on human judgment might be used with well-known fuzzy set similarity measures to eliminate the need to build ontologies.

Twenty word pairs selected from those in [7] are used to compare measures. Triangular fuzzy sets are created for each fuzzy word. Fuzzy set similarity measures can simply be used between the triangular membership functions. This approach is more efficient since the category ontologies creation is eliminated. Experiments are described in the following section with the specific fuzzy set similarity measures discussed here. The fuzzy word pair similarities are produced to determine how closely the results correlate with those produced by STASIS, FAST and FUSE; all of which use the same semantic similarity measure within an ontology.

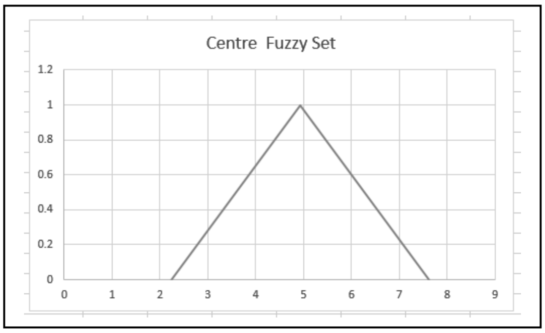


Fig. 1 Centre fuzzy set

The first three fuzzy set similarity measures described are used on the triangular membership functions. The last one uses the COGs of the type-2 interval fuzzy sets. Zadeh's sup-min is a partial matching measure [10]. The fuzzy Jaccard is a fuzzy set equality measure [11]. GeoSim uses the geometric distance between fuzzy sets to determine similarity [12]. The COG similarity measure for type-2 simply takes the distance between normalized COGs for the two fuzzy sets. It then normalizes the distance and converts it to similarity by subtracting from one. Both sup-min and the Jaccard measures produce a 0 similarity when the two fuzzy sets do not overlap. GeoSim and the COG type-2 similarity measures, however, produce a non-zero value even when the fuzzy sets do not overlap since both are based on distance.

B. Sup-Min

In [10] a detailed and thorough review of a variety of fuzzy set similarity measures is provided. Zadeh's consistency index also known as the sup-min or partial matching index falls into the set-theoretic category of fuzzy similarity measures. It roughly estimates the similarity between two fuzzy sets by finding at what domain values they intersect and determines their similarity by taking the highest membership degree among their intersection points. Given two fuzzy sets A and A', similarity between the two is determined as

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:47 UTC from IEEE Xplore. Restrictions apply.

$$S_{Zadeh}(A, A') = \sup_{u \in U} T(A'(u), A(u)) \qquad (1)$$

where T can be any t-norm, but usually the minimum is used for the t-norm. It is referred to as a partial since it only provides an estimated similarity value between the two fuzzy sets. .

C. Jaccard

The fuzzy Jaccard similarity measure is defined as a fuzzy extension of the Jaccard index [11] between two crisp sets by replacing set cardinality with fuzzy set cardinality. This fuzzy set similarity measure is also in the set theoretic category but provides a more comprehensive view of similarity between the two fuzzy sets since all elements in both fuzzy sets are taken into account not just the intersection point as in sup-min. Given two fuzzy sets A and A', similarity between the two is determined as

$$S_{Jaccard}(A, A') = | A \cap A' | \ / \ | A \cup A' | \qquad (2)$$

so the similarity is measured by the proportion of the area of the intersection of the two fuzzy sets to the area of the union of the two fuzzy sets.

D. Geometric Fuzzy Similarity Based on Dissemblance Index

Set theoretic fuzzy set similarity measures do not consider the distance of the fuzzy set A' from A. With the geometric fuzzy similarity measure [12], the distance between the two sets is the basis for determining their similarity. This distance is based on the dissemblance index that measures the distance between two real intervals. If $V = [v_1, v_2]$ and $W = [w_1, w_2]$, then

$$DI(V,W) = (|v_1 - w_1| + |v_2 - w_2|) / [2(\beta_2 - \beta_1)] \qquad (3)$$

where $[\beta_1, \beta_2]$ is an interval that contains both V and W. The factor $2(\beta_2 - \beta_1)$ is necessary to produce a normalized degree of dissemblance such that $0 \leq D(V, W) \leq 1$. The dissemblance index consists of two components, the left and right sides of each interval and may be generalized to fuzzy intervals.

A fuzzy interval N is defined by a pair of boundary functions L and R and parameters $(r_1, r_2, \lambda, \rho)$. The core of N, the values for which $\mu_N(r) = 1.0$ is the interval $[r_1, r_2]$. Parameters λ and ρ are used to define the left L and the right R boundary functions and the support of N, the values for which $\mu_N(r) \geq 0$, which is $[r_1 - \lambda, r_{2+}\rho]$. The L function and the R function define the membership functions for elements in the intervals $[r_1 - \lambda \ \ r_1]$ and $[r_2, r_2 + \rho]$, respectively. If L is positively sloping and linear and R is negatively sloping and linear then the interval N is a trapezoidal fuzzy membership function. Calculating the fuzzy dissemblance index between A and A' is done as an integration over α in the range 0 to 1 as

$$fDI(A'(u),A(u))=[\int||L_{A'}(\alpha)-L_A(\alpha)|+|R_{A'}(\alpha)-R_A(\alpha)|d\alpha] / [2(\beta_2-\beta_1)] \ (4)$$

where $[\beta_1, \beta_2]$ is an interval that contains both A' and A. fDI calculates a dissimilarity measure between the two fuzzy intervals based on a normalized distance. It can be converted into a similarity measure between the fuzzy intervals as

$$S_{GeoSim}(A, A') = 1 - fDI(A(u), A'(u)) \qquad (5)$$

With this similarity measure, even though A and A' may not overlap, a nonzero similarity value is produced since distance between the two sets is used.

E. Similarity on Type-2 Defuzzified Values Distance

As previously explained in [8] type-2 interval fuzzy sets were used and then defuzzified into a single value by adapting Mendel's footprint of uncertainty (FOU) method [13]. For each word in the six categories, the COG was determined using the lower FOU and upper FOU. The COGs were then scaled into the range [-1, +1]. To see how well a measure based solely on the distance between these scaled COG values worked, the following simple similarity measure is also used in this study:

$$S_{Type2\text{-}Dist}(A, A') = 1 - | COG_{Scaled}(A) - COG_{Scaled}(A')| / 2 \qquad (6)$$

The distance between the two centers of gravity is normalized by the size of the scaled interval [-1, +1]. Calculating this similarity measure between pairs of fuzzy words provides a means of determining how well it correlates with the ontology-based similarity measure developed for FAST and FUSE.

IV. Experimental Results and Analysis

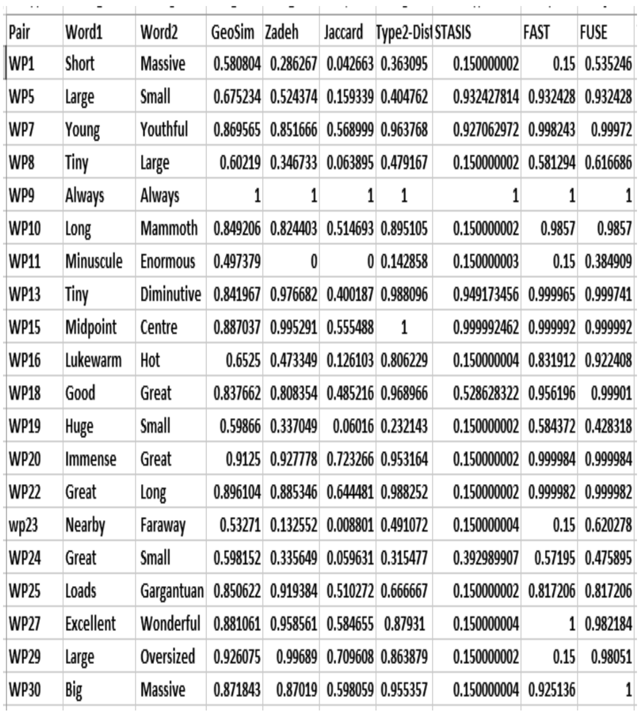
Table I shows 20 fuzzy word pairs used in the experimental investigation. These pairs were taken from the 30 sentence pairs used in the FAST study on sentence similarity [7]. Each of the sentences in the 30 pairs contain only one fuzzy word. Only 20 fuzzy word pairs are selected since 10 pairs are not both from the same category. Although a limited number of pairs, they can still provide evidence of proof of concept for the use of fuzzy set similarity measures. Once more data becomes available, more experiments can be undertaken.

Table I shows the similarity values produced by the various measures. STASIS, FAST and FUSE similarity values are all determined using the semantic similarity measure in [6] and differ because they use different ontological structures. STASIS uses WordNet. FAST uses the fuzzy category ontologies, each having five nodes in a binary tree structure and derived from the type-1 fuzzy sets created for each fuzzy word. FUSE also uses category ontologies; however, each has 11 nodes with a binary tree structure with 5 nodes on each side of the tree. Type-2 interval fuzzy sets are used to derive the FUSE category ontologies.

The correlations between the various pairs of similarity measures are presented in Table II. One can clearly see that STASIS has the lowest correlation with all the other similarity measures. That is an expected result since STASIS does not handle fuzzy words but uses the semantic similarity measure in [6] with the WordNet ontology. Its highest correlations are with FAST at over 0.46 and with FUSE at almost 0.39. Both of these use the same semantic similarity measure as STASIS, however, they use their own ontology categories instead of WordNet. The higher correlation of STASIS with FAST is most likely due to the FAST's simpler ontological structure so that the effects of fuzzy word similarity measure is not as significant as that for FUSE.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:47 UTC from IEEE Xplore. Restrictions apply.

TABLE I. SIMILARITY VALUES

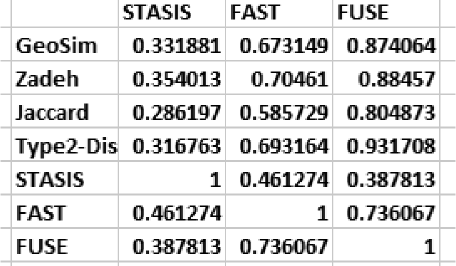


Pair	Word1	Word2	GeoSim	Zadeh	Jaccard	Type2-Dist	STASIS	FAST	FUSE
WP1	Short	Massive	0.580804	0.286267	0.042663	0.363095	0.150000002	0.15	0.535246
WP5	Large	Small	0.675234	0.524374	0.159339	0.404762	0.932427814	0.932428	0.932428
WP7	Young	Youthful	0.869565	0.851666	0.568999	0.963768	0.927062972	0.998243	0.99972
WP8	Tiny	Large	0.60219	0.346733	0.063895	0.479167	0.150000002	0.581294	0.616686
WP9	Always	Always	1	1	1	1	1	1	1
WP10	Long	Mammoth	0.849206	0.824403	0.514693	0.895105	0.150000002	0.9857	0.9857
WP11	Minuscule	Enormous	0.497379	0	0	0.142858	0.150000003	0.15	0.384909
WP13	Tiny	Diminutive	0.841967	0.976682	0.400187	0.988096	0.949173456	0.999965	0.999741
WP15	Midpoint	Centre	0.887037	0.995291	0.555488	1	0.999992462	0.999992	0.999992
WP16	Lukewarm	Hot	0.6525	0.473349	0.126103	0.806229	0.150000004	0.831912	0.922408
WP18	Good	Great	0.837662	0.808354	0.485216	0.968966	0.528628322	0.956196	0.99901
WP19	Huge	Small	0.59866	0.337049	0.06016	0.232143	0.150000002	0.584372	0.428318
WP20	Immense	Great	0.9125	0.927778	0.723266	0.953164	0.150000002	0.999984	0.999984
WP22	Great	Long	0.896104	0.885346	0.644481	0.988252	0.150000002	0.999982	0.999982
wp23	Nearby	Faraway	0.53271	0.132552	0.008801	0.491072	0.150000004	0.15	0.620278
WP24	Great	Small	0.598152	0.335649	0.059631	0.315477	0.392989907	0.57195	0.475895
WP25	Loads	Gargantuan	0.850622	0.919384	0.510272	0.666667	0.150000002	0.817206	0.817206
WP27	Excellent	Wonderful	0.881061	0.958561	0.584655	0.87931	0.150000004	1	0.982184
WP29	Large	Oversized	0.926075	0.99689	0.709608	0.863879	0.150000002	0.15	0.98051
WP30	Big	Massive	0.871843	0.87019	0.598059	0.955357	0.150000004	0.925136	1

Removing STASIS from the comparison since it does not handle fuzzy words, FUSE has the highest correlation with all the other similarity measures. Note that its correlations for all the fuzzy set similarity measures are greater than 0.80 and so greater than its correlation of about 0.74 with FAST. FAST is basically a precursor to FUSE with the noted differences for FUSE of type-2 interval fuzzy sets versus type-1 in FAST and the more complex 11 node category ontology versus only the 5 node category ontology in FAST.

It is surprising to see the simple fuzzy set similarity measure $S_{\text{Type2-Dist}}$ has the highest correlation 0.931708 with the more complex FUSE since it requires building ontologies for each of the six categories and using semantic similarity within an ontology. The $S_{\text{Type2-Dist}}$ simply takes the distance between the normalized COGs for the two fuzzy words, normalizes that distance based on the [-1, +1] interval, and converts it to a fuzzy similarity measure by subtracting it from 1.

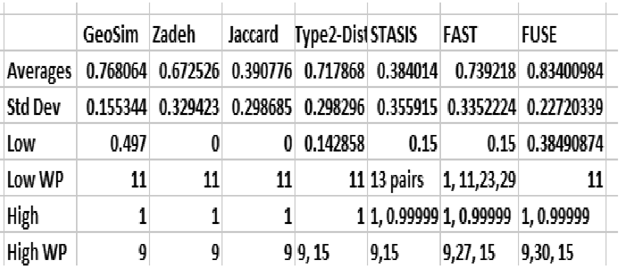
TABLE II. CORRELATIONS BETWEEN SIMILARITY VALUES



	STASIS	FAST	FUSE
GeoSim	0.331881	0.673149	0.874064
Zadeh	0.354013	0.70461	0.88457
Jaccard	0.286197	0.585729	0.804873
Type2-Dis	0.316763	0.693164	0.931708
STASIS	1	0.461274	0.387813
FAST	0.461274	1	0.736067
FUSE	0.387813	0.736067	1

Table III shows summary statistics for the similarity measures given in Table I.

TABLE III. SUMMARY STATISTICS FOR SIMILARITY VALUES



	GeoSim	Zadeh	Jaccard	Type2-Dist	STASIS	FAST	FUSE
Averages	0.768064	0.672526	0.390776	0.717868	0.384014	0.739218	0.83400984
Std Dev	0.155344	0.329423	0.298685	0.298296	0.355915	0.3352224	0.22720339
Low	0.497	0	0	0.142858	0.15	0.15	0.38490874
Low WP	11	11	11	11	13 pairs	1, 11,23,29	11
High	1	1	1	1	1, 0.99999	1, 0.99999	1, 0.99999
High WP	9	9	9	9, 15	9,15	9,27, 15	9,30, 15

As can be seen in Table I, all similarity measures agree on at least one word pair with the smallest similarity value, that is, word pair 11. However, only the Zadeh and Jaccard measures return 0 for this pair since there is no overlap between the triangular membership functions for those two fuzzy words. STASIS produces 0.15 similarity for 13 of the 20 word pairs and FAST produces 0.15 similarity for 4 of the 20 pairs and agrees with STASIS on those same 4 pairs. Since STASIS cannot handle fuzzy words, it can only use the semantic similarity measure as applied within the WordNet ontology and, therefore, cannot discriminate between these 13 pairs. FAST improves upon STASIS but still produces 4 pairs at the same similarity of 0.15. Only for word pair 11 does Type2-Dist similarity measure produce a value close to 0.15.

All similarity measures also agree on at least one word pair with the greatest similarity value, word pair 9. This word pair is somewhat of a reasonableness check since the pair has identical words. But note that Type2-Dist also produced a similarity value of 1 for word pair 15. This result is due to the defuzzified mean value of the Type 2 interval fuzzy sets being basically identical for those two words *midpoint* and *centre* based on the human evaluations. For the ontology-based similarity measures, all three produced similarity values extremely close to 1 so that this word pair is also listed for them. Both FAST and FUSE have an additional word pair that produces a value of 1, word pairs 27 and 30, respectively. These results may be attributed to the difference in the construction of the ontology structures created using the defuzzified mean values for FAST and FUSE.

For the average similarity values, STASIS has the lowest one. This result is again expected since this similarity measure does not consider fuzzy words, only a word's position in the WordNet hierarchy. The Jaccard set-based measure follows closely after STASIS with the next lowest average. With the type-1 fuzzy set creation by human experts, the experts only provided one number in the [0, 10] interval and the standard deviations were based on the set of expert evaluations. It is possible that the triangular fuzzy sets created from the mean and standard deviation values are a poorer representation that affects the set-based fuzzy similarity measure more than the distance based GeoSim and partial matching Zadeh measures. More experiments are needed to verify this possible explanation for Jaccard's lower similarity values.

From Table I, comparison for producing highest similarity values among all similarity measures shows that FUSE produces the highest or ties for highest with FAST for 12 of the

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:47 UTC from IEEE Xplore. Restrictions apply.

word pairs. FAST has the highest similarity or ties with FUSE 9 of the word pairs. Out of those word pairs with the highest similarity values, FAST and FUSE tie 6 times. FUSE produces higher similarity values because even for word pairs falling in the same node both within FAST and FUSE and, therefore, having a path length l equal to 0, the depth of the node h is typically at a higher level in FUSE than in FAST due to a maximum depth of 5 for FUSE compared to that of 2 for FAST. In addition the parameter β for FUSE is larger than that of FAST, i.e. 0.85 compared to 0.6. When FAST does produce a higher similarity, the path length l between the word pairs in FUSE's ontology is much greater than that in FAST's ontology, and with this case typically both word pairs are on different paths from the root node in both the FAST and FUSE ontologies. The depth h, therefore, would have the same value since the subsumer is the root node.

As can be seen from Table I, the fuzzy set based similarity measures rarely produce similarity measures greater than those that use the semantic similarity measure within an ontology. GeoSim and Type2-Dist have highest similarity for 2 word pairs each. Zadeh only has highest similarity once. For the lowest similarity values, Jaccard has lowest similarity for 12 of the 20 word pairs. It is to be expected that the semantic similarity measure used within the FAST and FUSE ontologies would produce higher similarity values than the fuzzy set similarity measures since there is a limit to the greatest path length of 4 and 10, and depth of 2 and 5, respectively The results from the semantic similarity measure are very much dependent on the structure of the ontologies that have been developed from the type-1 and type-2 interval fuzzy sets.

FUSE generally produces higher similarity values but both FAST and FUSE agree on numerous word pairs. This can occur when both the path distance l between the words pairs and the depth h of the subsumer of the word pairs are identical in both the FAST and FUSE ontologies.

V. Conclusions and Future work

This paper has conducted a study on fuzzy word sets derived from data collected from human participants and evaluates the performance of four simple fuzzy set similarity measures. It compares these results to the results of one semantic similarity measure as applied to two different ontologies created for FAST and FUSE from the fuzzy word sets. From the study, a very simple fuzzy set similarity measure created from COG distance between type-2 fuzzy sets has a very high correlation with the FUSE similarity results, even higher than that of FAST results with FUSE, both of which use the same semantic similarity measure. This result demonstrates that the construction of the ontology for the categories plays a significant factor in the resulting similarity values. The major difference between the two ontologies is in the level of detail considered in their construction. FAST is created using type-1 fuzzy sets and uses only 5 nodes with a depth of 2 in its ontology. FUSE is created using type-2 interval fuzzy sets and its ontology has 11 nodes with a depth of 5. Creating these ontologies is not straightforward and determining the appropriate structure for fuzzy word categories needs more investigation.

Although ontology creation for fuzzy words is challenging and it is unlikely that human perceptions of all the fuzzy words in a given language could be modelled, even with a limited number of fuzzy word models, the use of fuzzy semantic similarity measures in applications is beneficial. One aspect of future work looks at incorporating such measures into dialogue systems to replace traditional pattern matching algorithms with short text comparisons. Another area is to use the fuzzy set similarity measures instead of semantic similarity within the sentence similarity systems of FAST and FUSE to determine how well they correlate with human judgments. A hybrid of a fuzzy set similarity measure and a semantic similarity measure should be experimented with for the cases where sentence similarity does not agree with the human judgments of sentence similarity.

References

[1] L. Zadeh, "From Computing with Numbers to Computing with Words—from Manipulation of Measurements to Manipulation of Perceptions. Logic, Thought and Action," International Journal of Applied Math. Comput. Sci., vol.12, no.3, pp. 307–324, 2002.

[2] G. Salton, C. Buckle, Term-weighting approaches in automatic textretrieval", Information processing & management vol.24, no. 5, pp.513-523, 1988.

[3] T. Landauer, P. Foltz, D. Laham, "An introduction to latent semantic analysis," Discourse processes vol. 25, no 3, pp.259-284,1998..

[4] Y. Li, D. Mclean, Z. Bandar, J. O'Shea, K. Crockett, "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 8, pp.1138-1150, 2006.

[5] V. Cross, Xinran Yu, Xueheng Hu, "Unifying ontological similarity measures: A theoretical and empirical investigation," Int. J. Approx. Reasoning vol. 54 no. 7, pp. 861-875, 2013.

[6] Li, Y, Bandar, Z. McLean, D. "An approach for measuring semantic similarity between words using multiple information sources". IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp.871-882, 2003.

[7] D. Chandran, K. A. Crockett, D McLean, Z. Bandar, "FAST: A fuzzy semantic sentence similarity measure," International Conference on Fuzzy Systems, FUZZ-IEEE, 2013.

[8] N. Adel, K. A. Crockett, A. Crispin, D. Chandran, J. P. Carvalho, "FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets," International Conference on Fuzzy Systems, FUZZ-IEEE pp. 1 -8 2018:

[9] Mendel, J. "Computing with words and its relationships with fuzzistics", Information Sciences vol. 177, no. 4, pp.988-1006, 2007.

[10] V. Cross, An Analysis of Fuzzy Set Aggregators and Compatibility Measures, Ph.D. Dissertation, Computer Science and Engineering, March 1993, Wright State University, Dayton, OH, 264 pages.

[11] P. Jaccard. "The distribution of the flora in the alpine zone", *New Phytologist*, vol. 11, pp. 37–50, 1912.

[12] V. Cross, T. Sudkamp, "Geometric compatibility modification," Fuzzy Sets and Systems, vol. 84, no. 3, pp. 283-299, 1996.

[13] M. Hao and J. M. Mendel, "Encoding words into normal interval type-2 fuzzy sets: HM approach," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp. 865-879, 2016.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:47 UTC from IEEE Xplore. Restrictions apply.

Human Hedge Perception – and its Application in Fuzzy Semantic Similarity Measures

IEEE International Conference on Fuzzy Systems

2019

Human Hedge Perception – and its Application in Fuzzy Semantic Similarity Measures

Naeemeh Adel, Keeley Crockett, Alan Crispin
School of Computing, Mathematics and Digital Technology,
Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
N.Adel@mmu.ac.uk

Joao P. Carvalho
INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa, Portugal
David Chandran
Institute of Psychiatry, Psychology & Neuroscience, Kings College London, 16 De Crespigny Park, London,
SE5 8AF, UK

Abstract* - Fuzzy Semantic Similarity Measures are algorithms that are able to compare two or more short texts that contain human perception based words and return a numeric measure of similarity of meaning between them. Such similarity is computed using a weighting, comprised of the semantic and the syntactic composition of the short text. Similarities of individual words are computed through the use of a corpus, and ontological structures based on both WordNet – a well-known lexical database of English, and on category specific fuzzy ontologies created from the derivation of Type-I or Type-II interval fuzzy sets from human perceptions of fuzzy words. Currently, linguistic hedges are not utilized in the similarity calculation within fuzzy semantic similarity measures and are ignored. This paper describes a study, which aims to capture human perceptions for linguistic hedges typically used in natural language. Twelve linguistic hedges used within natural language are selected and an experiment is conducted to capture human perceptions of the impact of hedges on fuzzy category words. A dataset of hedge sentence pairs is created and rated in terms of similarity by human participants. Excellent inter-rater correlations and inter-class correlations are established between the average human ratings and an established fuzzy semantic similarity measure.

Keywords— hedges, lingustic variables, fuzzy semantic similarity measures, interval type-II

I. Introduction

In the field of fuzzy logic, linguistic variables are a well-defined concept where the value of the variables are words that are used in natural human language [1]. In [1], Zadeh defines a term-set for each linguistic variable (i.e. Age) which constitutes all its possible numerical values i.e. [0..130], with linguistic values (i.e. *young*) acting as labels for fuzzy restrictions based upon the meaning and interpretation by a human in a given context. A number of linguistic hedges, designed to modify fuzzy sets through concentration, intensification and dilation, were first defined as mathematical models. In some cases there is not an agreement on the model [2, 3, 4] and more recent work by Le and Tran [5] on dual hedges (i.e. hedges can be used simultaneously to express different levels of emphasis) consider extensions to fuzzy logics through two axiomatizations for multiple hedges [5]. Novak [4] proposes that the branch of fuzzy logic, known as Fuzzy Natural Logic provides a rational model of linguistic semantics, and argues that hedging is more complex than previously known when applied within the field of linguistics. Current work on hedges includes the use of linguistic terms with weakened hedges (LTWH) to enhance natural uncertainty in decision-making [6, 7]. This work uses two frequently used hedges within qualitative decision-making, and argues that the formulation of more complex linguistic expressions improves decision making under uncertainty. Work in [7, 8] acknowledges that more linguistic hedges need to be determined especially for use within modelling natural language. The effect of hedges applied to fuzzy systems has been studied in many application domains such as enhancing a student's academic evaluation [8], the selection of a supplier based on a number of live parameters within a product's supply chain in small and medium businesses [9] and vehicular traffic density estimation [10]. However, their effect within the application of fuzzy semantic similarity measures has not been studied.

Fuzzy Natural Language Processing (FNLP) can be addressed with the formulation of fuzzy computational models of words [4]. We define fuzzy words, within this work, as any word that has a subjective meaning in natural language and is based on a human's perception in a given context. Fuzzy words are often defined from the bottom up – based upon obtaining a representative sample of the human population for a given word and context and then modelling the range of perceptions using either a Type-I or Type-II fuzzy set representation [11, 12]. This process of using humans' subjective opinions is adopted from the field of natural language processing [13] and from work undertaken by Mendel [14, 15, 16], first in his code-book using a Type-I representation and then following the Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets [17].

The motivation for the work in this paper stems from a weakness in the application of fuzzy semantic similarity measures (FSSM) which are used to find a measure of the semantic and syntactic similarity, between short texts, typically of 25 words or less [18]. Currently, linguistic hedges are not utilised in the similarity calculation within FSSMs. Two such FSSM measures are FAST [11] and FUSE [12]. FAST was the first FSSM built on a limited number of categories of words represented by Type-I fuzzy sets used to derive category ontologies similar to WordNet [19]. FUSE (FUzzy Similarity

978-1-5386-1728-1/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

mEasure) determined similarity using expanded categories of perception based words that were modelled using Interval Type-2 fuzzy sets [12]. We hypothesise that the inclusion of the semantic meaning of linguistic hedges will improve the precision of the similarity measurement through obtaining a higher correlation of similarity with human ratings. Hence, linguistic hedges are expected to make a weighted contribution when calculating the overall semantic similarity.

This paper is organized as follows; Section II provides a brief summary of background work on hedges and related work on FSSMs. Section III defines the study that aims to capture human perceptions for linguistic hedges typically used in natural language. In section III, the methodology for natural language hedge selection and obtaining human perceptions of hedges in relation to fuzzy words is described. Following the modelling of the hedges using Type-II interval fuzzy sets, the methodology for creating 16 hedge sentence pairs is presented. Section IV presents the results obtained from capturing perceptions of humans for 12 hedge words and obtaining human similarity ratings between hedge sentence pairs. Section V explores further work in exploring hedge weightings within FSSMs.

II. Background and Related work

A) Hedges

A linguistic variable carries with it the concept of fuzzy set qualifiers, called hedges. A hedge is a marker of uncertainty in language. Hedges are terms that modify the shape of fuzzy sets. They include adverbs such as *very*, *somewhat*, *quite*, *more or less* and *slightly* [20]. Linguistic variables represent crisp information in a form, and precision, appropriate for the problem. Linguistic variables associate a linguistic condition with a crisp variable. A crisp variable is the kind of variable that is used in most computer programs: an absolute value. A linguistic variable, on the other hand, has a proportional nature: in all of the software implementations of linguistic variables, they are represented by fractional values in the range of 0 to 1 [21]. Hedges can modify verbs, adjectives, adverbs or even whole sentences. They are used as [20]:

- All-purpose modifiers, such as *very*, *quite* or *extremely*
- Truth-values, such as *quite true* or *mostly false*
- Probabilities, such as *likely* or *not very likely*
- Quantifiers, such as *most*, *several* or *few*
- Possibilities, such as *almost impossible* or *quite possible*.

Hedges act as operations themselves. For instance, *very* performs concentration and creates a new subset from the fuzzy set it is applied to i.e. applying the hedge *very* to the set of *tall men*, derives the subset of *very tall men*. Hedges are useful as operations, but they can also break down continuums into fuzzy intervals. For example, the following hedges could be used to describe temperature: *very cold, moderately cold, slightly cold, neutral, slightly hot, moderately hot* and *very hot*. Obviously, these fuzzy sets overlap. Hedges help to reflect human thinking, since people usually cannot distinguish between *slightly hot* and *moderately hot* [20]. This makes them important when measuring human perceptions of the similarity of short texts.

According to Zadeh [22], a linguistic variable is a variable, whose values are words or sentences in a natural or artificial language, as opposed to numerical values. Therefore for the category *Age*, it would be considered a linguistic variable if its values were linguistic rather than numerical, this means A*ge* = {*young, not so young, very young… old, not very old, not very young*} is a linguistic variable, as opposed to *Age* = {20, 21, 22 … 60, 61…} which is a numerical variable.

A linguistic variable is characterised by a quintuple (*L*, *T*(*L*), *U*, *G, M*) where [22]:

- *L* is the name of the linguistic variable
- *T*(*L*) is the term set of *L* (collection of linguistic values)
- *U* is the universe of discourse
- *G* is a *syntactic rule* which generates the terms in *T*(*L*)
- *M* is a *semantic rule* which associates with each linguistic value *X* its meaning *M*(*X*)
- Where *M*(*X*) denotes a fuzzy subset of *U*.

Considering the example of *tall men*, application of the concentration hedge, *very* operation, will reduce the degree of memebership of fuzzy elements [20]. The application of hedge *very*, can be calculated using a mathematical square as follows:

$$\mu_A^{very}(x) = [\mu_A(x)]^2 \tag{1}$$

Thus if a person had a 0.84 membership in the set of *tall men*, then they will have a 0.7056 membership in the set of *very tall men*.

B) Fuzzy Semantic Similiarty Measures

Traditionally, Semantic Similarity Measures stemmed from the field of natural language processing and are used for measuring the degree to which a sentence or short-texts are subjectively evaluated by humans to assess whether or not they are semantically similar to each other. Traditional measures did not capture the use of fuzzy words - words that have subjective meanings to different people in different contexts, are typically ambiguous and are characteristically used in everyday human natural language dialogue [12]. The FAST algorithm (Fuzzy Algorithm for Similarity Testing) [1], is an ontology based similarity measure that uses concepts of fuzzy words represented by Type-I fuzzy sets. However, Type-I fuzzy sets were not able to correctly model the subjective options of humans on the meanings of fuzzy words in different contexts. FUSE, attempted to overcome this problem, by using Interval Type-II fuzzy sets to model relationships between categories of human perception based words using fuzzy category ontologies. The FUSE algorithm which can be found in [12], consisted of both syntactic and semantic components which were weighted. FUSE was able to model intra-personal (the uncertainty a person has about the word) and inter-personal (the uncertainty that a group of people have about the word) uncertainties, which are intrinsic to natural language. In [11], FUSE gave better correlations compared to human ratings than FAST over three benchmark datasets [16]. In these results, the modelling of linguistic hedges and the impact on the similarity measurement value was not considered. Hedges were not represented in the fuzzy category

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

ontologies and therefore did not form part of the similarity measurement.

III. Capturing Human Perceptions of Hedges – A Study

A) Overview of study

The aim of this study is to investigate the effect of inclusion of hedge modifiers within the similarity calculation of fuzzy sentence similarity measures. The hypothesis is that their inclusion will improve the precision of the similarity measurement through obtaining a higher collaboration of similarity with human ratings. To investigate the hypothesis, a study consisting of two experiments was undertaken. The first experiment was to obtain human perceptions of the intensity that a hedge had on a fuzzy word. Fuzzy intensity in this research refers to the perceptive numerical measure a word is given, be that measure positive, or negative by a human rater.

For this experiment, let the fuzzy subset *Hedges = {Below, Approximately, Neighbouring, Roughly, About, Around, Quite, Indeed, Definitely, Positively, Very, Above}.* The fuzzy words were selected from the 6 original categories proposed in FUSE [12] as follows: *{Adequate (Level of Membership), Satisfactory (Worth), Middle-Aged (Age), Mild (Temperature), Fair (Frequency), Average (Size & Distance)}.*These fuzzy words were chosen by selecting the word with the value closest to 0 in each category on a scale of [-1, +1]. Once human perceptions were captured they could be used to construct Type-II interval models similar to those used in FUSE [12] and used to derive a hedge ontology. The ontology would be used to determine the path length and depth between words as part of the word component similarity measures in FUSE. The path length and depth of hedge words are relative to their position in the hedge ontology where each hedge category is treated as a concept. Each concept is constructed using a taxonomy (binary tree) where the root node always takes the value 0. Defuzzified hedge words are then placed into tree nodes at intervals of ± 0.2 [12] From the hedge taxonomy, the path length and depth of the Lowest Common Subsumer can be determined for hedge words in a category. This would allow the defuzzified hedge value to influence its associated defuzzified fuzzy word values, in terms of intensity, be this positively in that the sentence similarity value increased or negatively in that the sentence similarity value decreased.

B) Hedge Intensity Experiment

To determine intensity of hedges when applied to fuzzy words, 32 participants consented to take part in a study, all of whom were native English speakers above the age of 18. In total there were 12 hedge words that were not already present in the FUSE Fuzzy Dictionary [12] that had mathematical definitions. When the mathematical value of a hedge word, (such as *Very* as defined in *Eq. (1)*) was applied to a fuzzy word it did not represent the mathematical model that was linguistically represented, therefore a different approach was needed to cater for hedge words. As an example, the hedge word *Very* has a mathematical equation of x^2 [23], where x is the fuzzy value. Therefore taking the word *Hot=0.6193*, and computing the phrase *Very Hot*= $(0.6193)^2$ = *0.3836*, calculated the mathematical value of *Very Hot* to be smaller than the mathematical value of Hot, whereas linguistically *Very Hot* has a more positive intensity then *Hot*. Therefore a different approach to measuring the intensity was required that required the perceptions of humans. To achieve this the subset of 12 hedge words where each added prior to the fuzzy words, one from each of the 6 categories represented in the FUSE FSSM [12]. The middle word in each category with the value closest to zero was selected, and a random hedge word was added to the beginning of each of these six words. Participants were first given a description of the task, which included a simple linguistic definition of a hedge and a fuzzy word. An extract from the experiment description is as follows: *"The aim of this experiment is to help contribute towards computer systems that will understand the English language. This experiment is about HEDGES. Hedges are terms that modify the shape of a sentence. They include adverbs such as very, somewhat, quite, more or less and slightly. In this experiment, I am going to give you 6 words belonging to 6 categories. A category in this instance is just the name given for a group of words that fall under a similar meaning. For instance, for the category TEMPERATURE, it will contain words such as [hot, cold, mild, boiling, scorching, freezing…]. I am going to give you a scale of 0 to 10. Each word sits in the middle of this scale (5). I am going to pair each word with some hedge words and would like you to tell me where these new words would sit on this scale. You can use one decimal place (e.g. 3.2) for finer precision."*

An image of a ruler (Figure 1) was used as a visual aid to make understanding the word placement visually easier. The chosen word from each category was always located at mark 5 on the ruler and was highlighted in red. The participants were then asked to rate the new *hedge word* when applied to the fuzzy *word* on this ruler on a scale of [0-10] with 1 decimal place permitted for accuracy. One example of a word used in this experiment is the hedge word *Below*. Taking the fuzzy word *Fair*, belonging to the category *Frequency*, one participant felt that the word *Below Fair* would be represented by a value of 3.4 as shown in Figure 1. Their opinion was that the hedge, *Below*, negatively reduced the intensity of the category word Fair.

The aim of the hedge intensity experiment was to try and mimic the perceptions of humans using natural language, despite them not actually thinking about words on a scale. On obtaining all human measurements, the average value for each hedge word was calculated and this was scaled on a scale of [-1, +1] to create a hedge ontology. This was done to match the same scale and ontological structure as the words in the fuzzy dictionary used within FUSE [12].

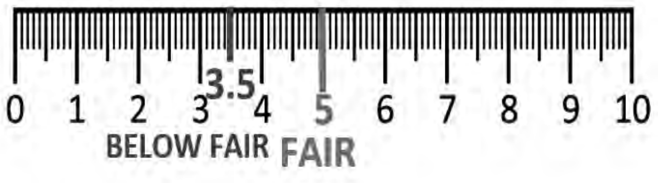


Fig. 1 - Scale for Hedge Intensity Experiment

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

C) *Human Ratings of Hedged Sentence Pairs*

In order to assess the intensity of hedges in the natural language context, it was necessary to compute the sentence similarity between pairs of sentences, which contained hedge words. Following analysis, it was established that the fuzzy sentence benchmark datasets, known as SFWD and MFWD [12], did not contain a sufficient number of hedge words in order to conduct a rigorous evaluation. Therefore, a dataset containing 16 sentence pairs containing hedge words was created. The methodology comprised of randomly extracting 16 sentences pairs from the MFWD [12] ranging from high to low similarity based on human ratings [12]. For each fuzzy word in the hedge sentence pair (HSP), a hedge word was assigned prior to that fuzzy word, i.e. for *HSP1* "The little village of Resina is also situated *approximately* near the spot", the hedge *approximately* was added. The sentence pairs were then checked by an English language expert, to ensure they were grammatically correct. Table I shows the full set of hedge sentence pairs.

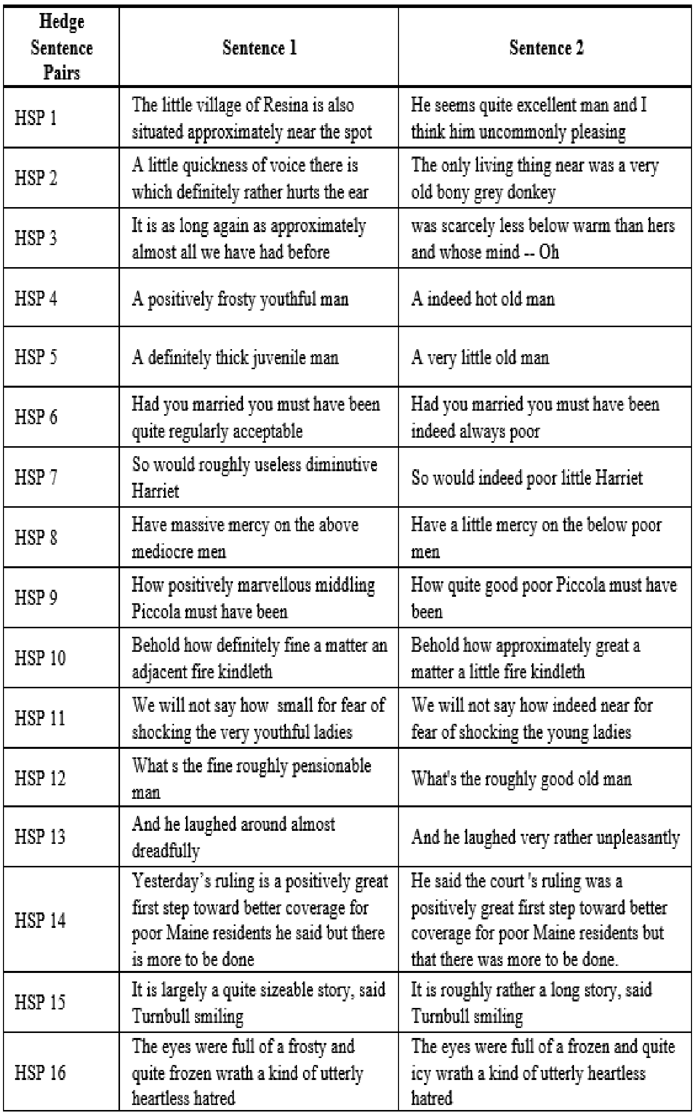
O'Shea et. al. [13] emphasized the importance of establishing rigorous methodology when obtaining human ratings of similarities between words and sentence pairs, especially in relation to sample size, population distribution and the inclusion of calibration pairs providing representation of the highest and lowest sentence similarity pairs within the data set. Adopting this methodology, the second experiment consisted of 16 participants who were all native English speakers above the age of 18 from a diverse range of backgrounds. They were provided with the 16 HSPs and were asked to rate each sentence on a scale of [0-10], with 1 decimal place permitted for accuracy, based on how similar they were to each other. The scale of [0-10] was adopted to be consistent with approaches in [11, 12, 13, 24].

IV. Experimental Results and Discussion

A) *Hedge Intensity Results*

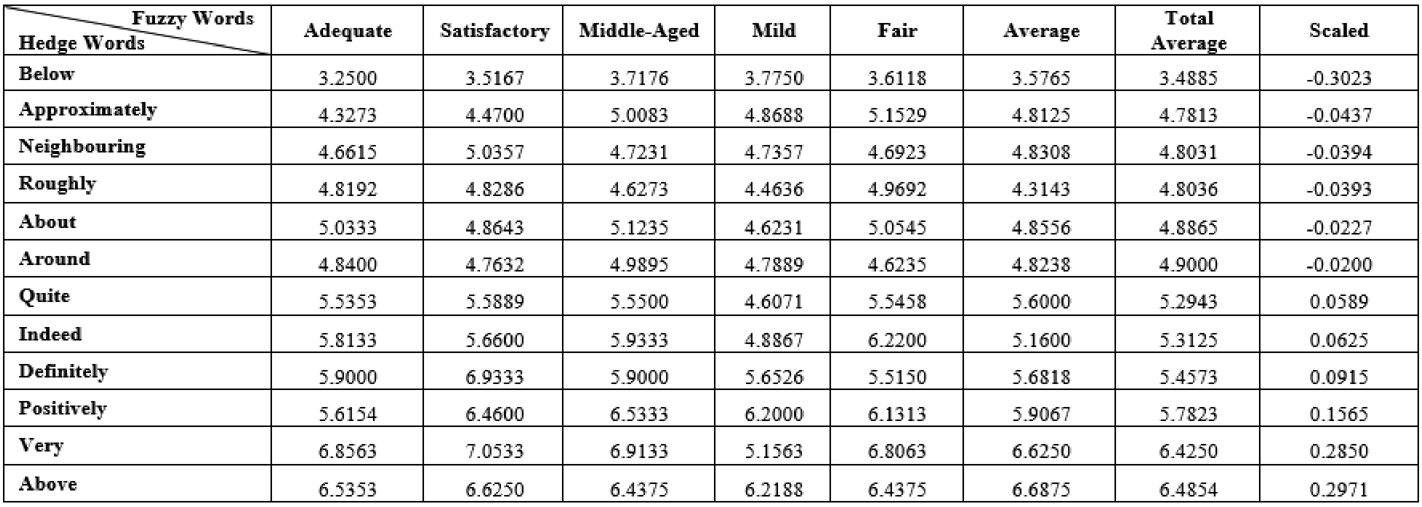
Table II shows the results of the Average Human Ratings (AHR) for the hedge intensities. The table shows the 6 words from the fuzzy dictionary categories (Fuzzy Words), and the 12

TABLE I HEDGE SENTENCE PAIRS CONTAINING FUZZY & HEDGE WORDS



Hedge Sentence Pairs	Sentence 1	Sentence 2
HSP 1	The little village of Resina is also situated approximately near the spot	He seems quite excellent man and I think him uncommonly pleasing
HSP 2	A little quickness of voice there is which definitely rather hurts the ear	The only living thing near was a very old bony grey donkey
HSP 3	It is as long again as approximately almost all we have had before	was scarcely less below warm than hers and whose mind -- Oh
HSP 4	A positively frosty youthful man	A indeed hot old man
HSP 5	A definitely thick juvenile man	A very little old man
HSP 6	Had you married you must have been quite regularly acceptable	Had you married you must have been indeed always poor
HSP 7	So would roughly useless diminutive Harriet	So would indeed poor little Harriet
HSP 8	Have massive mercy on the above mediocre men	Have a little mercy on the below poor men
HSP 9	How positively marvellous middling Piccola must have been	How quite good poor Piccola must have been
HSP 10	Behold how definitely fine a matter an adjacent fire kindleth	Behold how approximately great a matter a little fire kindleth
HSP 11	We will not say how small for fear of shocking the very youthful ladies	We will not say how indeed near for fear of shocking the young ladies
HSP 12	What s the fine roughly pensionable man	What's the roughly good old man
HSP 13	And he laughed around almost dreadfully	And he laughed very rather unpleasantly
HSP 14	Yesterday's ruling is a positively great first step toward better coverage for poor Maine residents he said but there is more to be done	He said the court 's ruling was a positively great first step toward better coverage for poor Maine residents but that there was more to be done.
HSP 15	It is largely a quite sizeable story, said Turnbull smiling	It is roughly rather a long story, said Turnbull smiling
HSP 16	The eyes were full of a frosty and quite frozen wrath a kind of utterly heartless hatred	The eyes were full of a frozen and quite icy wrath a kind of utterly heartless hatred

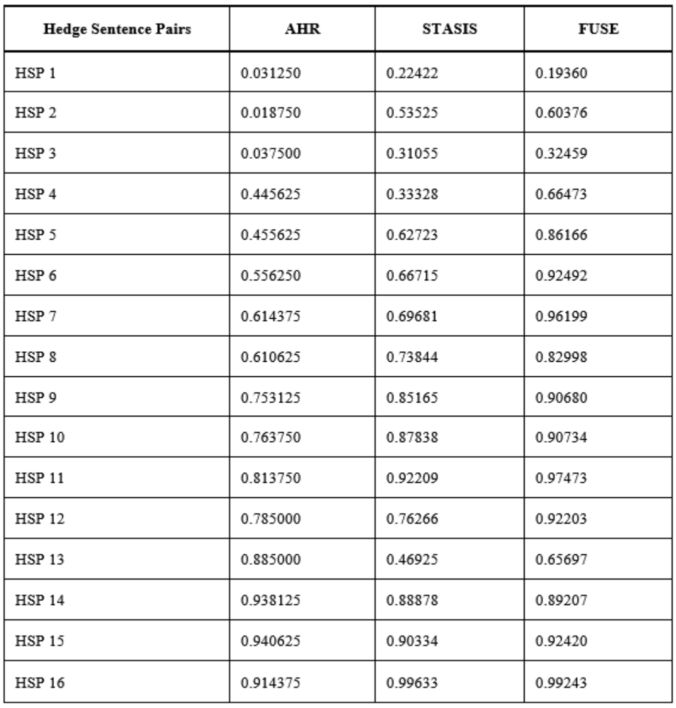
TABLE II. AHR FOR HEDGE INTENSITIES ORDERED HIGH TO LOW



Fuzzy Words Hedge Words	Adequate	Satisfactory	Middle-Aged	Mild	Fair	Average	Total Average	Scaled
Below	3.2500	3.5167	3.7176	3.7750	3.6118	3.5765	3.4885	-0.3023
Approximately	4.3273	4.4700	5.0083	4.8688	5.1529	4.8125	4.7813	-0.0437
Neighbouring	4.6615	5.0357	4.7231	4.7357	4.6923	4.8308	4.8031	-0.0394
Roughly	4.8192	4.8286	4.6273	4.4636	4.9692	4.3143	4.8036	-0.0393
About	5.0333	4.8643	5.1235	4.6231	5.0545	4.8556	4.8865	-0.0227
Around	4.8400	4.7632	4.9895	4.7889	4.6235	4.8238	4.9000	-0.0200
Quite	5.5353	5.5889	5.5500	4.6071	5.5458	5.6000	5.2943	0.0589
Indeed	5.8133	5.6600	5.9333	4.8867	6.2200	5.1600	5.3125	0.0625
Definitely	5.9000	6.9333	5.9000	5.6526	5.5150	5.6818	5.4573	0.0915
Positively	5.6154	6.4600	6.5333	6.2000	6.1313	5.9067	5.7823	0.1565
Very	6.8563	7.0533	6.9133	5.1563	6.8063	6.6250	6.4250	0.2850
Above	6.5353	6.6250	6.4375	6.2188	6.4375	6.6875	6.4854	0.2971

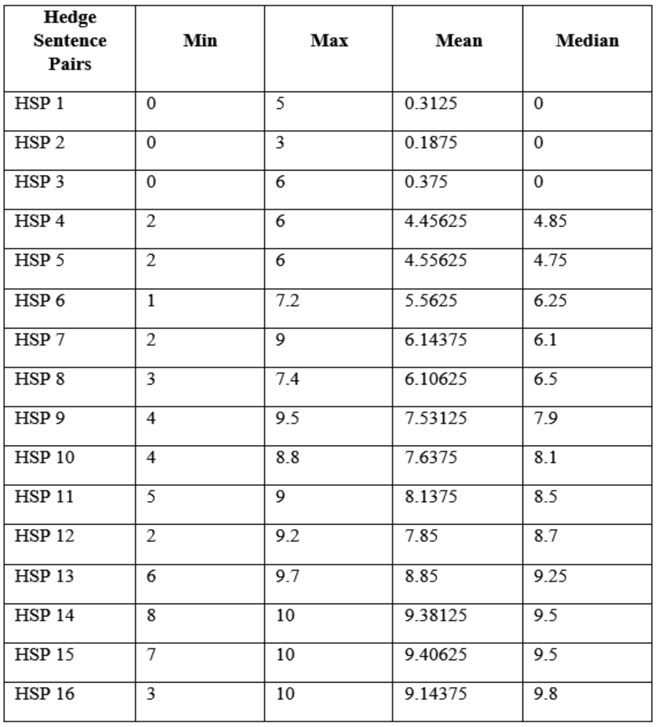
Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

TABLE IIIA. COMPARISON OF HEDGE SENTENCE PAIRS (MEAN AHR), STATIS (CRISP) AND FUSE (FUZZY) SIMILARITY MEASURES



Hedge Sentence Pairs	AHR	STASIS	FUSE
HSP 1	0.031250	0.22422	0.19360
HSP 2	0.018750	0.53525	0.60376
HSP 3	0.037500	0.31055	0.32459
HSP 4	0.445625	0.33328	0.66473
HSP 5	0.455625	0.62723	0.86166
HSP 6	0.556250	0.66715	0.92492
HSP 7	0.614375	0.69681	0.96199
HSP 8	0.610625	0.73844	0.82998
HSP 9	0.753125	0.85165	0.90680
HSP 10	0.763750	0.87838	0.90734
HSP 11	0.813750	0.92209	0.97473
HSP 12	0.785000	0.76266	0.92203
HSP 13	0.885000	0.46925	0.65697
HSP 14	0.938125	0.88878	0.89207
HSP 15	0.940625	0.90334	0.92420
HSP 16	0.914375	0.99633	0.99243

TABLE IIIB. DISTRIBUTION OF THE HUMAN RATINGS OF HEDGE SENTENCE PAIRS



Hedge Sentence Pairs	Min	Max	Mean	Median
HSP 1	0	5	0.3125	0
HSP 2	0	3	0.1875	0
HSP 3	0	6	0.375	0
HSP 4	2	6	4.45625	4.85
HSP 5	2	6	4.55625	4.75
HSP 6	1	7.2	5.5625	6.25
HSP 7	2	9	6.14375	6.1
HSP 8	3	7.4	6.10625	6.5
HSP 9	4	9.5	7.53125	7.9
HSP 10	4	8.8	7.6375	8.1
HSP 11	5	9	8.1375	8.5
HSP 12	2	9.2	7.85	8.7
HSP 13	6	9.7	8.85	9.25
HSP 14	8	10	9.38125	9.5
HSP 15	7	10	9.40625	9.5
HSP 16	3	10	9.14375	9.8

hedge words chosen (Hedge Words). It gives a (Total Average), which is the average of each hedge row, that is then scaled between [-1, +1] (Scaled) to match the rest of the values scaling in the fuzzy dictionary, ordered from low to high. On examimning the results it can be seen that *Very Fair* is more positively intensified than *Fair*, and the results indicate this closely i.e. *Fair*= 0.085 and *Very Fair*= 0.285. The same applies to *Mild*= -0.2387 and *Very Mild*= 0.285; thus the hedge *Very* positively intensifies a fuzzy word between the ranges of [0.0462,..,0.37]. An example of the affect of negative intensity is the hedge word *Below*, with *Below Fair* = -0.2173 and *Below Mild* = -0.5411, thus *Below* negatively intensifies a fuzzy word between the range of [-0.541,..,-0.2173].

B) Hedge Sentence Pairs results

Table IIIA shows the average human ratings (AHR) obtained from the 16 participants who rated the HSPs. The 16 participants were different from those who had taken part in the Hedge Intensity Experiments outlined in Section III(B); all of whom were native English speakers above the age of 18. Sentence similarity measurements are shown for FUSE and for comparison the similarity is also shown for the measure STASIS which does not incorporate any fuzzy words. Table IIIB shows the distribution of the human ratings showing the Minimum, Maximum, Mean and Median values for each of the 16 sentence pairs.

Table IVA shows one example of a hedge sentence pair (HSP) with average human rating (AHR= 0.8850) taken from Table IIIA. The hedges used in this example are *around* and *very*. The fuzzy words in the sentence pairs are *almost* and *rather* belonging to the category *Level of Membership,* and *dreadfully* and *unpleasantly* belonging to the category *Worth*. STASIS ignores all fuzzy and hedge words and therefore similarity is low (STASIS=0.46925), FUSE on the other hand caters for both fuzzy words and hedge words, therefore has a higher similarity rating (FUSE=0.65697) which is closer to the AHR. This goes to show that fuzzy words and hedge words play an important role in the similarity rating of a short text. On the other hand, Table IVB which relates to HSP12 shows that STASIS (0.76266) has a closer rating to the AHR (0.785000) then FUSE (0.92203). This is likely to be due to the human sample size being relatively small [13] and/or the variations of WordNet used in STASIS and FUSE, as WordNet is constantly being updated.

Looking at the Inter-Rater Correlation in Table V, FUSE gave a higher correlation to Average Human Ratings, with 0.803, compared to STASIS with Average Human Ratings at 0.796. Although the correlation difference was not significant, it is still an improvement over STASIS, which shows that fuzzy hedge intensity does play an important role in sentence similarity. This small improvement can be attributed to 1) the fact that only twelve hedge words were modelled, 2) the coverage of the hedge words in the HSP dataset was limited and 3) the number of human raters was only 16 – acceptable in the NLP community but on the low end of the scale where 32 participants is typically recommended.

Conduction of an Inter-Rater Correlation produces some positive results as can be seen in Table V, with FUSE=0.886 as opposed to STASIS=0.796.

Cicchetti gives the following guidelines for intra-class correlation coefficient agreement measures [25]:

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

- Less than 0.40 - Poor.
- Between 0.40 and 0.59 - Fair.
- Between 0.60 and 0.74 - Good.
- Between 0. 75 and 1.00 – Excellent

Each of the algorithms STASIS and FUSE is compared against the Average Human Ratings (AHR). Looking at the AHR which is referred to as (a) in this instance, for each of the algorithms it can be seen that in Table VI for STASIS (a= 0.865) and in Table VII for FUSE (a= 0.867) with a confidence interval of 95%. Based on Cicchetti's guidelines, it can be concluded that the intra-class correlation coefficient is deemed as excellent for both datasets.

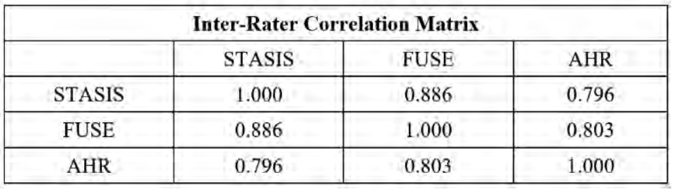
TABLE IVA. A GOOD EXAMPLE OF HSP

Hedge Sentence Pairs	Sentence 1	Sentence 2	AHR	STASIS	FUSE
HSP 13	And he laughed around almost dreadfully	And he laughed very rather unpleasantly	0.885000	0.46925	0.65697

TABLE IVB. A BAD EXAMPLE OF HSP

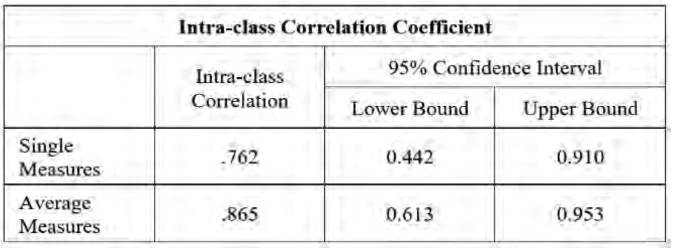
Hedge Sentence Pairs	Sentence 1	Sentence 2	AHR	STASIS	FUSE
HSP 12	What s the fine roughly pensionable man	What's the roughly good old man	0.785000	0.76266	0.92203

TABLE V. INTER-RATER CORRELATION OF FUSE & STASIS



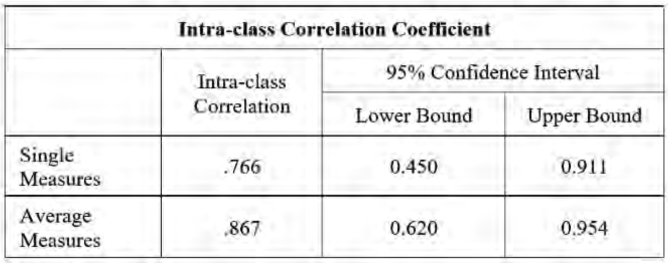
Inter-Rater Correlation Matrix			
	STASIS	FUSE	AHR
STASIS	1.000	0.886	0.796
FUSE	0.886	1.000	0.803
AHR	0.796	0.803	1.000

TABLE VI. INTRA-CLASS CORRELATION COEFFICIENT FOR STASIS



Intra-class Correlation Coefficient			
	Intra-class Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Single Measures	.762	0.442	0.910
Average Measures	.865	0.613	0.953

TABLE VII. INTRA-CLASS CORRELATION COEFFICIENT FOR FUSE



Intra-class Correlation Coefficient			
	Intra-class Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Single Measures	.766	0.450	0.911
Average Measures	.867	0.620	0.954

V. CONCLUSION AND FURTHER WORK

This paper has presented a study on the application of linguistic hedges within fuzzy semantic similarity measures. This has involved first obtaining human intensity ratings of a small selection of hedges to fuzzy words. These hedges were then modelled using Type-II interval fuzzy sets for inclusion in the FUSE fuzzy dictionary. A set of 16 hedge sentence pairs were constructed using the modelled hedges and 16 participants rated their similarity. Although there was minor improvement on the similarity measurement correlation between average human ratings and the fuzzy measure FUSE, it was not significant. This is mainly due to the number of hedges modelled and the number of participants involved in rating the hedge sentence pairs. However even with this small sample, it can be seen that linguistically modelled hedges have a positive effect on sentence similarity. Current work consists of, but is not limited to, expanding the hedge sentence pairs and also expanding the sample size to cater for more human ratings. A future experiment will investigate the impact of hedges on the degree of intensification of a sentence, by determining the fuzzy similarity of pairs of sentences, first with hedges and then without, and comparing both results to the average human rating of each variation. This would allow a greater evaluation of the impact of hedge words applied to individual fuzzy words beyond this paper by looking at how a human interprets the hedge words in the context of a sentence.

Current work is incorporating FUSE into dialogue systems, which will allow a wider range of natural language dialogue to be explored and tested in real-world dialogue utterance exchanges.

ACKNOWLEDGMENT

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under reference UID/CEC/50021/2019, grant SFRH/BSAB/136312/2018 and project LISBOA-01-0145-FEDER-031474

REFERENCES

[1] Zadeh, L.A., "Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges", Journal of Cybernetics, 2:3, 1972, pp. 4-34, DOI: 10.1080/01969727208542910

[2] Kerre E.E., De Cock M., "Linguistic Modifiers: An Overview. In: Chen G., Ying M., Cai KY. (eds) Fuzzy Logic and Soft Computing." The International Series on Asian Studies in Computer and Information Science, vol 6, 1999, Springer, Boston, MA

[3] Shi, H., Ward, R., Kharma, N., "Expanding the definitions of linguistic hedges", Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), 2001

[4] Novak, V., Fuzzy logic in natural language processing, in Proc. IEEE Int. Conf. Fuzzy Systems Conf.erence, 2017.

[5] Le,V., Tran, D., Extending fuzzy logics with many hedges, Fuzzy Sets and Systems, Vol345, 2018, pp.126-138, https://doi.org/10.1016/j.fss.2018.01.014

[6] Wang, H. Xu, Z. Zeng, X., Linguistic terms with weakened hedges: A model for qualitative decision making under uncertainty, Information Sciences, Volumes 433–434, 2018, pp. 37-54.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

[7] Wang, H. Xu, Z. Zeng, X, Huchang, L., Consistency measures of linguistic preference relations with hedges, 2018, IEEE Transactions on Fuzzy Systems, DOI: 10.1109/TFUZZ.2018.2856107.

[8] Hameed, I., Enhanced fuzzy system for student's academic evaluation using linguistic hedges,, 2017, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), DOI: 10.1109/FUZZ-IEEE.2017.8015462

[9] Djatna, T., Luthfiyanti, R., Abbas, A., "Intuitionistic fuzzy hedges modeling for supplier selection of responsive agroindustrial multi products supply chains in small and medium enterprises", 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), DOI: 10.1109/ICACSIS.2017.8355042,

[10] Borkar, P., Sarode, M.V. and Malik, L.G., Int. J. Fuzzy Syst., 2016, 18: 379. https://doi-org.ezproxy.mmu.ac.uk/10.1007/s40815-015-0069-5

[11] Chandran, D., "The development of a fuzzy semantic sentence similarity measure," Doctorate of Philosophy, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2013.

[12] Adel, N., Crockett, K., Crispin, A., Chandran, D. and Carvalho, J.P., FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets. In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8)., 2018, IEEE.

[13] O'Shea, J.D., Bandar, Z.A. and Crockett, K. "A new benchmark dataset with production methodology for short text semantic similarity algorithms," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, no. 4, 2013, p. 19.

[14] Mendel, J.M., "Computing with words and its relationships with fuzzistics," Information Sciences, vol. 177, no. 4, 2007, pp. 988-1006.

[15] Mendel, J.M. and Wu, D., "Determining interval type-2 fuzzy set models for words using data collected from one subject: Person FOUs", 2014 IEEE International Conference on Fuzzy Systems, pp.768 – 775, 2014.

[16] Mendel, J.M., "Type-2 Fuzzy Sets as Well as Computing with Words", Published in: IEEE Computational Intelligence Magazine (Volume: 14 , Issue: 1 , Feb. 2019), DOI: 10.1109/MCI.2018.2881646,

[17] Hao., M. and Mendel, J.M., "Encoding words into normal interval type-2 fuzzy sets: HM approach", IEEE Transactions on Fuzzy Systems, vol. 24, no. 4, 2016, pp. 865-879.

[18] Li,Y., McLean, D., Bandar, Z.A., O'Shea, J.D. and Crockett, K. "Sentence similarity based on semantic nets and corpus statistics", *IEEE transactions on knowledge and data engineering,* vol. 18, no. 8, 2006, pp. 1138-1150.

[19] Princeton University, "About Wordnet." Retrieved 13 June 2014 from http://wordnet.princeton.edu/

[20] Negnevitsky, M., Artificial Intelligence: A Guide to Intelligent Systems. 2 ed., 2005, p. 435. England: Pearson Education Limited.

[21] Banks, W., (2003) "Linguistic Variables: Clear Thinking with Fuzzy Logic." Retrieved 30 May 2016 from http://www.phaedsys.co.uk/principals/bytecraft/bytecraftdata/LinguisticVariables.pdf

[22] Zadeh, L.A., "The Concept of a Linguistic Variable and its Applications to Approximate Reasoning", Information Sciences, Vol. 8,1975, pp 199-249 (Part I), pp 301-357 (Part 11), Vol. 9, 1975, pp 43-80 (Part In)

[23] Cox, E. "The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems" vol. 1,. AP Professional, 1994

[24] Feilong, L. and Mendel, L.M., "Encoding words into interval type-2 fuzzy sets using an interval approach." IEEE Transactions on Fuzzy Systems, vol. 16, no. 6, 2008, pp.1503-1521.

[25] Cicchetti, D.V., "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." Psychological Assessment, vol. 6, no. 4, 1994, p. 284.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:35:07 UTC from IEEE Xplore. Restrictions apply.

FUSE (Fuzzy Similarity Measure)-A Measure for Determining Fuzzy Short Text Similarity using Interval Type-2 Fuzzy Sets

IEEE International Conference on Fuzzy Systems

2018

FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets

Naeemeh Adel, Keeley Crockett, Alan Crispin
School of Computing, Mathematics and Digital Technology,
Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
N.Adel@mmu.ac.uk

David Chandran
Institute of Psychiatry, Psychology & Neuroscience, Kings College London, 16 De Crespigny Park, London,
SE5 8AF, UK
Joao P. Carvalho
INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa, Portugal

***Abstract*—Measurement of the semantic and syntactic similarity of human utterances is essential in developing language that is understandable when machines engage in dialogue with users. However, human language is complex and the semantic meaning of an utterance is usually dependent on context at a given time and also based on learnt experience of the meaning of the perception based words that are used. Limited work in terms of the representation and coverage has been done on the development of fuzzy semantic similarity measures. This paper proposes a new measure known as FUSE (FUzzy Similarity mEasure) which determines similarity using expanded categories of perception based words that have been modelled using Interval Type-2 fuzzy sets. The paper describes the method of obtaining the human ratings of these words based on Mendel's methodology and applies them within the FUSE algorithm. FUSE is then evaluated on three established datasets and is compared with two known semantic similarity algorithms. Results indicate FUSE provides higher correlations to human ratings.**

Keywords*—*fuzzy semantic similarity measures, fuzzy natural language, fuzzy words, interval type-2

I. INTRODUCTION

The dream of humanoid robots with intelligence is becoming more of a reality than science fiction [1]. One area of intensive research is in the communication and understanding of human language between humans and machines. For a machine to truly understand a human language, it must be understood in the context of the conversation in a timely manner and the response provided by the machine must also relate to the context so the human understands. Goal orientated conversational agents (GCA) [2] are one such example where machines support humans in achieving a goal, but to do so each human utterance – in the form of a simple statement or question, must be interpreted, analysed and an appropriate response conducted. In the context of GCA, semantic similarity measures [3] can be used to supplement pattern-matching approaches enabling user utterances to be analysed, both in the syntactic and semantic content, thus improving robustness, etc. There is very limited work on developing these measures for understanding a fuzzy utterance in a timely context. In this work, a fuzzy utterance is defined as a short text or sentence, which comprises of at least one fuzzy word. A fuzzy word is a word that has a subjective meaning, and is characteristically used in everyday human natural language dialogue. Fuzzy words are often ambiguous and in meaning, since they are based on an individual's perception [4].

Computing with Words (CWW) [5] relates to developing intelligent systems that are able to receive as input, words, perceptions, and propositions drawn from natural language and can then produce a decision or output based on these words. CWW becomes a necessary tool when the available information is perception-based or not precise enough to use numbers, as is the case of most real world applications involving humans. CWW adds to conventional modes of computing the capability to compute with interpreted words and propositions drawn from natural language [6]. Type-1 fuzzy sets were originally used to construct fuzzy sets to model words [6, 7]. Zadeh first introduced Type-1 fuzzy sets, where membership is non-binary and concepts are subjective [8]. According to Mendel [9], words can mean different things to different people and this causes linguistic uncertainty when modelling perception based words. Therefore, Mendel states that using a Type-1 fuzzy set to model a word is scientifically incorrect, because a word is uncertain whereas a Type-1 fuzzy set is certain, therefore, Type-1 cannot cater for linguistic uncertainties [9]. For this reason, Mendel concluded that Type-2 fuzzy sets should be used to model words instead. The 3D nature of Type-2 allows uncertainties to be better modelled. Type-2 fuzzy sets are computationally intensive because Type-reduction is very intensive, and for this reason, Mendel later proposed the use of Interval Type-2 fuzzy. Interval Type-2 is simpler to use because the membership functions are interval sets, and therefore the secondary memberships will either be zero or one [10, 11]. Thus, concepts from CWW provide an ideal platform for handling uncertainties in natural language in the context of semantic similarity measures.

Fuzzy Sentence Similarity Measures (FSSM) are algorithms that are able to compare two or more short texts which contain

978-1-5090-6020-7/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

human perception based words and return a numeric measure of similarity of meaning between them. The Fuzzy Algorithm for Similarity Testing (FAST) [12], is the only current FSSM to date, that uses concepts of CWW to allow for the accurate representation of fuzzy based words. Through human experimentation, fuzzy sets were created for six categories of words using Type-1 fuzzy sets (Size & Distance, Age, Goodness, Frequency, Temperature and Completeness). The application of Type-1 fuzzy sets caused a weakness within FAST; since these words are not a true representation of each category, because the rating of the words is still the subjective opinion of those individuals [9]. This adversely affected the accuracy of the defuzzified values in each category by the potential bias of an individual's views in experiments to quantify fuzzy words.

This research investigates and develops a new algorithm called FUSE (FUzzy Similarity mEasure). FUSE is an ontology based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception based words. The proposed algorithm is more suited to modelling *intra-personal* (the uncertainty a person has about the word) and *inter-personal* (the uncertainty that a group of people have about the word) uncertainties, which are intrinsic to natural language; because the membership grade of an Interval Type-2 fuzzy set is an interval instead of a crisp number as in Type-1 fuzzy sets [10]. In addition, Type-1 fuzzy sets have been shown to not provide the flexibility for simultaneously incorporating both kinds of linguistic uncertainties [13]. Therefore, the key research question addressed in this paper is; can a Type-2 fuzzy set be used to represent an individual's perception within a FSSM?

FUSE identifies fuzzy words in a human utterance and determines their similarity in context of both the semantic and syntactic construct of the sentence. There are a number of key differences between FUSE and FAST. First of all a larger vocabulary of fuzzy words are included in FUSE [12] giving a 57.65% increased coverage of perception based words. Secondly, a new set of fuzzy ontologies has been developed for these categories in FUSE. Thirdly where FAST only modelled words in Type-1, FUSE models words within the category and deduces the fuzzy membership using Interval Type-2 fuzzy sets. The paper also presents the methodology for collecting people's subjective values of fuzzy words using the Hao-Mendel Approach (HMA) [11], for estimating words as Interval Type-2 fuzzy sets which are then defuzzified.

This paper is organised as follows; Section II provides an overview of Type-2 fuzzy sets within CWW, reviews word and short text similarity measures and looks at the challenges associated with using humans to gather similarity ratings. Section III describes how Mendel's HMA method was applied to the task of rating words for the purpose of constructing ontologies of fuzzy words. Section IV introduces the FUSE algorithm and Section V describes the experimental design and results that show that FUSE gives better correlation to human results compared with other known similarity measures. Finally, Section VI presents the conclusions and future work.

II. Related work

A) Type-2 Fuzzy Sets within CWW

Zadeh first introduced Computing with words (CWW) in 1996, where he explained CWW as a methodology for reasoning, computing and decision-making with information described in natural language. In CWW, words are modelled using fuzzy sets [5, 11]. There are three main principles to CWW according to Zadeh [7]. The first, recognized that human knowledge is often described using words and phrases associated in natural language. Secondly, that when using natural languages, words are used when exact amounts or numbers are unknown and therefore allow less precise meaning to be conveyed. Zadeh also stated, "*Precision carries a cost. If there is a tolerance for imprecision, it can be exploited through the use of words in place of numbers*" [7]. The first step in using fuzzy logic for CWW is to construct fuzzy sets to model words. Since words can mean different things to different people according to Mendel [9], this can cause linguistic uncertainty, which is involved in CWW. Therefore using Type-2 fuzzy sets to model words allows for this uncertainty to be catered for. Hence, Mendel concludes that one should use Interval Type-2 fuzzy models in order to model first-order word uncertainties [14].

When people rate words in terms of their similarity, it is still the subjective opinion of those individuals. Groups of people rate words to either belong in a set or not belong in a set; this generally leads to gaps and noise, such as large differences in opinions or missing information. An example of this may be: '*Today is such a hot day, I'm roasting!*'; different people will have different opinions of how hot the day is to them depending on their heat tolerance, the geographical location etc. therefore, will rate the concept of "hot" and hence the word hot differently. This is why Type-1 sets are not able to directly model such uncertainties because their membership functions are totally crisp and two-dimensional. However, Type-2 fuzzy sets are able to model such uncertainties because their membership functions are fuzzy and three-dimensional [15]. By being three dimensional, Type-2 fuzzy sets provide additional degrees of freedom that make it possible to directly model uncertainties.

B) Word and Semantic Similarity Measures

A general issue in linguistic, AI and cognitive science is the measurement of semantic similarity for a given pair of words/sentences. Therefore, the performance of applications can be greatly improved with a proper metric for measurement. Metrics are usually divided into two classes: Path Based Metrics and Information Content (IC) Based Metrics [16]. Semantic similarity has been successfully applied in [17, 18, 19, 20, 21].

Path based metrics proceed from the position of each concept in the taxonomy to obtain semantic similarity and assess semantic similarity by computing geometric distance separating two concepts, such as the number of edges. It is based on the assumption that the similarity of two concepts is related with the path length between two concepts and depth of each concept in the taxonomy respectively. Wu and Palmer presented a scaled metric for measuring the similarity between a pair of concepts [22]. Rada et al. utilized the minimum path length connecting

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

the concepts containing the compared words as a measure for calculating the similarity of words [23]. In 1998, Leacock and Chodorow proposed a similar method for measuring word similarity [24]. They used the WordNet taxonomy to compare words and calculated the shortest path between the words taking into account the maximum depth of the WordNet taxonomy.

The notion of information content of the concept is directly related to the frequency of the term in a given document collection. The frequencies of terms in the taxonomy are estimated using noun frequencies in some large collection of texts. The idea behind semantic similarity information content metrics is that each concept includes information in WordNet. It assumes that the similarity of two concepts is related to information they share in common. The more common information two concepts share, the more similar the concepts are. In 1995, Resnik first proposed an information content (*IC*) based similarity metric [25]. Resnik assumed that for a concept *c*:

$$IC = -\log p\,(c) \tag{1}$$

Where *p(c)* is the probability of encountering an instance of concept *c* [16].

Jiang and Conrath presented an approach for measuring semantic similarity/distance between words and concepts in 1997 [26]. The proposed measure is a combined approach that inherits the edge-based approach of the edge-counting scheme, which is then enhanced by the node-based approach of the information content calculation. If the compared concepts share a lot of information, then the IC will be high and the semantic distance between the compared concepts will be smaller [26].

The edge based approach is a more natural and direct way of evaluating semantic similarity in a taxonomy. It estimates the distance (e.g. edge length) between nodes, which correspond to the concepts/classes being compared. Given the multidimensional concept space, the conceptual distance can conveniently be measured by the geometric distance between the nodes representing the concepts. Obviously, the shorter the path from one node to the other, the more similar they are [26].

Li et al., uses multiple information sources to calculate the semantic similarity of concepts and proposes a metric based on the assumption that information sources are infinite to some extent while humans compare word similarity with a finite interval between completely similar and nothing similar [27]. Intuitively, the transformation between an infinite interval to a finite one is non-linear [16, 27]. Li et al define local semantic density as a monotonically increasing function of *wsim* (w_1, w_2):

$$f_3(wsim) = \frac{e^{\lambda.wsim(w_1,w_2)} - e^{-\lambda.wsim(w_1,w_2)}}{e^{\lambda.wsim(w_1,w_2)} + e^{-\lambda.wsim(w_1,w_2)}} \tag{2}$$

Where $\lambda > 0$. If $\lambda \rightarrow \infty$, then the information content of words in the semantic nets is not considered [16, 27].

The only known FSSM is FAST (Fuzzy Algorithm for Similarity Testing) [12], which is an ontology based similarity measure that uses concepts of fuzzy and computing with words to allow for the accurate representation of fuzzy based words. FAST is designed to be able to represent the effect fuzzy words have in the semantic meaning of a human utterance on the level of semantic similarity. In FAST, levels of similarity between sets of fuzzy words can be calculated by examining the position of the word (based on its Type-1 fuzzy set defuzzified values derived from human ratings) through calculating the similarity between pairs of fuzzy words. FAST has shown an improvement over existing algorithms STASIS and LSA (Latent Semantic Analysis) which do not take into consideration fuzzy words when computing semantic sentence similarity [12]. Furthermore, the improvement that both FAST and STASIS showed over LSA indicates that it is necessary for an ontology to be used in conjunction with a corpus, rather than a corpus alone in terms of determining the level of similarity between sentences with fuzzy words. The results have shown that an increased number of fuzzy words in sentences do have an effect on the performance of SSM. This is demonstrated through the improvement that FAST had over STASIS and LSA [4] but this depends on the domain and coverage of fuzzy words.

C) Challenges in Gathering Human Ratings

There are several challenges that arise when creating a dataset that will be used for measuring semantic similarity which were identified by O'Shea et al. [28] in developing his gold standard dataset known as STSS-131. Firstly, obtaining a valid sample that is representative of the domain - this may either be words or in this research, utterances in the English language. Next is the task of collecting valid human ratings of similarity between the words/utterances. In the case of the research proposed in this paper, native English speakers were used to collect ratings to ensure that words did not have meanings that were too far apart, lessening the risk of distorting the results. It was noted in [28] that regional dialect might also interfere with the ratings given by participants in an experiment, however in this research, these experiments were conducted in the UK and ratings obtained from participants from the Manchester region. The third challenge is in knowing what statistical measures are needed to measure fuzzy similarity. The Pearson correlation coefficient [29] is a long-established measure of agreement used in semantic similarity that assumes a linear relationship between the two variables being compared and will be applied as the statistical measure in this work to evaluate FUSE.

III. Method For Obtaining Human Ratings Of Words

A) Data Collection

FUSE uses six fuzzy categories to hold fuzzy words (Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership). It was recognized that the coverage on words in the first FSSM, FAST, was very limited, with just 196 words over the six categories. In order to expand the categories, the Oxford English Synonyms Dictionary was used. The words that already existed in FAST were taken and, using the dictionary, all the one word synonyms for the existing words were also added to each category. Only one-word synonyms were added, such as '*hot*' or '*cold*', and 2 word synonyms such as '*fairly-hot*', were not added [11]. Once all the categories had been

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

TABLE I. Full List of Participation Breakdown

Category	Before Cleaning	Gender		Age		Education	
Size / Distance	38	M F	26 6	(18-23) (24-29) (30-35) (36-41) (42-47) (54 +)	18 6 5 1 1 1	(A-Levels) (Undergraduate) (Postgraduate) (PhD) (Other)	11 10 8 2 1
Temperature	32	M F	25 7	(18-23) (24-29) (30-35) (36-41) (54 +)	24 4 2 1 1	(GCSE) (A-Levels) (Undergraduate) (Postgraduate) (PhD) (Other)	1 18 5 6 1 1
Age	41	M F	26 6	(18-23) (24-29) (30-35) (42-47) (48-53)	22 7 1 1 1	(Below GCSE) (A-Levels) (Undergraduate) (Postgraduate) (PhD) (Other)	1 13 12 3 1 2
Frequency	35	M F	25 7	(18-23) (24-29) (30-35)	25 4 3	(GCSE) (A-Levels) (Undergraduate) (Postgraduate) (Other)	1 20 7 3 1
Worth	37	M F	26 6	(18-23) (24-29) (30-35) (48-53) (54 +)	22 6 2 1 1	(A-Levels) (Undergraduate) (Postgraduate) (PhD) (Other)	16 9 3 1 3
Level Of Membership	37	M F	26 6	(18-23) (24-29)	26 6	(A-Levels) (Undergraduate) (Postgraduate) (Other)	15 12 2 3

TABLE II. Percentage Increase of Words for FUSE

Categories	Words Per Category	Percentage Increase on FAST
Size/Distance	91	102.22%
Temperature	36	16.13%
Age	42	31.25%
Frequency	48	84.62%
Worth	61	48.78%
Level of Membership	31	47.62%

updated with the additional words, the total increased to 309 words, giving a 57.65% increase over FAST (Table I shows full breakdown below).

B) Methodology

The method for obtaining human ratings of words to be used to construct fuzzy ontologies (similar to those constructed for the lexical database WordNet [30]) for FUSE is based on Mendel's Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets [11].

In [11], Mendel used 50 intervals to obtain the person Footprint of Uncertainty (FOU) for the word. He did this by asking one participant to rate words on a scale of *l*-*r* giving the left (x_L,y_L) and right (x_R,y_R) endpoints, this scale can be [1..4], [0..10] etc. Using the one rating Mendel obtained from the one person, he then went on to generate 100 random numbers (L_1, L_2,…,L_{50}; R_1, R_2,…,R_{50}) and used these to generate 50 endpoint interval pairs [(L_1, R_1), (L_2, R_2),…,(L_{50},R_{50})]. In Mendel's approach [11], he used only one participant rating to generate variants as it reduces the time required to collect ratings. In this research, an approach utilized from the field of semantic similarity was adopted and *n* actual participants were used to provide ratings. In obtaining human ratings for words in FUSE categories, each category had a minimum of 32 participants whose ratings per word were obtained; therefore, the person FOU was not used, however the HMA approach was used to collect data from group participants.

Data was collected for the six categories using an online questionnaire and participants were asked to rate the words in each category on a scale of [0-10]. A full list of participant's demographics is shown in Table I.

For example given the word '*Hot*' belonging to the category '*Temperature*' the question would be as follows: *"Rate the word HOT as a measure of Temperature on a scale of 0 to 10. (You can go up to one decimal place). PLEASE ONLY WRITE YOUR ANSWERS IN THE FORMAT "x to y" WHERE x AND y ARE THE NUMBERS YOU HAVE CHOSEN"*. Each category had in excess of 32 participants. This meant that even after removing noise, each category was still left with 32 participants. Each participant was asked to rate a selection of words belonging to a category. Each question asked the user to give a range of where they felt the word would be placed on this scale of [0-10]. Users were permitted to use numbers up to one decimal place for precision (e.g. 3.4). A generic example was provided in each question to ensure users understood what range meant and to ensure they gave a start point and end point [11].

In order to not exhaust the users and potentially affect the quality of the results, each user was asked to fill in one questionnaire relating to only one category at one sitting. The criteria for the candidates was that they had to be native English speakers. Volunteers were emailed a link, which would direct them to the questionnaire. Each questionnaire required a minimum of 32 respondents to make it valid. Once all six categories were complete, cleaning and analysis of the results took place. Due to each category having 32 responses or more, this helped in ensuring that after cleaning and removing any bad or incorrect results, each category was still left with a minimum of 32 responses. Table II shows the percentage increase of words for each category in FUSE compared to that of FAST.

Using Mendel's statistics and probability theory, the following steps below were adapted to remove noise [11].

1. Remove bad data – in this step all nonsensical results were removed; in this case, it was any results that fell outside the [0-10] range requested.
2. Remove outliers - using Box and Whisker tests [31] outliers are removed simultaneously from the results. Only the data intervals that are within an acceptable two-sided tolerance limit were kept. According to Mendel, a

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

tolerance interval is a statistical interval within which, with some confidence level 100 (1- 10)%, a specified proportion (1-0) of a sampled population falls.

3. Remove data intervals that have no overlap or too little overlap with other data intervals. If it overlaps with another data interval, then Mendel and Wu [32] state that it is reasonable.

When all noise has been removed, each category is now left with 32 clean data because of the questionnaires. Once the process of removing noise is complete, the original *n* data intervals have been reduced to a set of *m* data intervals where $m \leq n$. This now results in $m = 32$ for each of the six categories.

Once cleaned data was ready for analysis, each category was analysed word by word. This was achieved by finding the upper FOU and lower FOU for each word; from this, the COG (Centre of Gravity) was calculated as defined in *eq.(3):*

$$COG = \frac{\left(\left(\frac{a+b}{2}\right)+\left(\frac{c+d}{2}\right)\right)}{2} \qquad (3)$$

Where:

a = upper left FOU
b = lower left FOU
c = lower right FOU
d = upper right FOU

Tables III and IV show defuzzified examples for the words '*Regular*' and '*Nearby*' from the category '*Size/Distance*' respectively on a scale of [0-10]. The values are calculated using the triangular membership function. '*x*' is the scale of [0-10], '*lower*' represents the lower boundaries, and '*upper*' represents the upper boundaries. '*t-norm$_{(prod)}$*' is the multiplication of lower and upper, and '*t-norm$_{(min)}$*' is the minimum boundary from the lower or upper. Figures 1 and 2 show the Type-1 defuzzified graphical representation of the word '*Regular*' and the word '*Nearby*' respectively in the category Size/distance that has resulted from the triangular membership calculation. The values of '*t-norm$_{(min)}$*' have been used to plot the graphs.

The results (*y*) were then scaled on a scale of [-1 to +1] using *eq.(4).*

$$y = a + \frac{(x-A)(b-a)}{B-A} \qquad (4)$$

Where

A = smallest number in dataset
B = largest number in dataset
a = minimum normalised value (-1)
b = maximum normalised value (+1)
x = value we want to scale (in this case the COG)

This now meant that every category contained words with values ranging from [-1 to +1]. This scale was selected to allow representation of defuzzified word values in each fuzzy category ontology, required to obtain measurements in FUSE (described in Section IV).

IV. FUSE (FUZZY SIMILARITY MEASURE)

This section first defines how the fuzzy category ontologies are constructed and then defines the proposed FUSE algorithm.

TABLE III. SCALE FOR WORD 'REGULAR'

x	Lower	Upper	T-norm$_{(prod)}$	T-norm$_{(min)}$
0	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00
2	0.00	0.27	0.00	0.00
3	0.36	0.53	0.19	0.36
4	0.73	0.80	0.58	0.73
5	0.89	0.94	0.84	0.89
6	0.44	0.71	0.31	0.44
7	0.00	0.47	0.00	0.00
8	0.00	0.24	0.00	0.00
9	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00

TABLE IV. SCALE FOR WORD 'NEARBY'

x	Lower	Upper	T-norm$_{(prod)}$	T-norm$_{(min)}$
0	0.00	0.00	0.00	0.00
1	0.00	0.29	0.00	0.00
2	0.40	0.57	0.23	0.40
3	0.80	0.86	0.69	0.80
4	0.80	0.86	0.69	0.80
5	0.40	0.57	0.23	0.40
6	0.00	0.29	0.00	0.00
7	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00

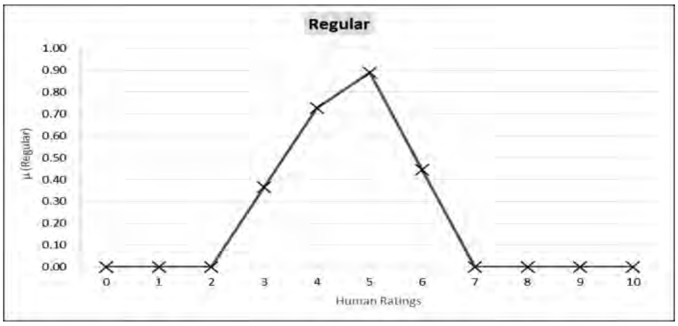


Fig. 1. Defuzzified Figure for 'Regular'

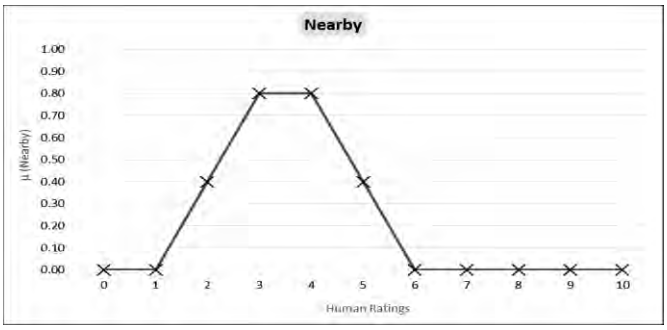


Fig. 2. Defuzzified Figure for 'Nearby'

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

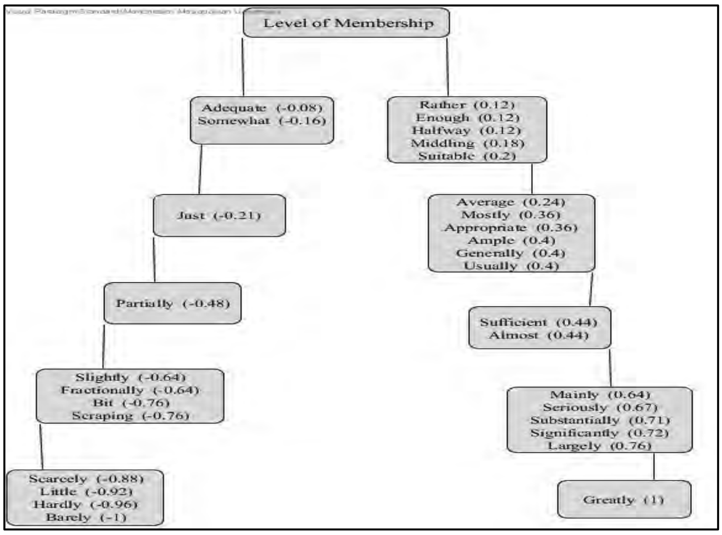


Fig. 3. Ontology for Level of Membership

A) *Fuzzy Ontology Representation*

To show how words in a category are introduced on a scale of [-1, +1] it was necessary to construct an ontology. Each category is treated as a concept. Words within each concept are treated as instances. Each concept has a taxonomy that arranges the words as a binary tree so that the root node always takes the value 0. The defuzzified value of words are equally placed into nodes in intervals of ± 0.2, which was an empirically determined threshold. This approach allows calculation of the path length and depth of the Lowest Common Subsumer (LCS) to be calculated for fuzzy words in a category which could not be done using traditional resources such as WordNet, due to lack of coverage of fuzzy words. Figure 3, shows the words in the category '*Level of Membership*' represented in an ontology structure. The numbers next to each word represent the defuzzified value of that word obtained from the human rating experiment described in Section III. Each partition contains words up to a certain fixed value, with the negative values on one side and the positive values on the other; this allows path length to be calculated.

B) *FUSE Algorithm*

FUSE utilizes a crisp word sentence similarity STASIS, when computing word similarity between nouns and verbs; when it encounters perception based words within an utterance, word similarity is calculated through determining the path length, *l*, and the length of the shortest path from the associated fuzzy category ontology.

```
Input: Let $U_1$ and $U_2$ be two fuzzy utterances, which the semantic similarity is to be calculated.

Output: Similarity measure of $U_1$ and $U_2$
For i = 0 to n in $U_1$ and $U_2$ where n is the total of words ($w_1 \ldots w_n$) in $U_1$ and $U_2$
    Tag every tokenized word ($w_1 \ldots w_n$) in $U_1$ and $U_2$ [ADJ (adjective), ADP (adposition), ADV (adverb), CONJ (conjunction), DET (determiner), NOUN (noun), NUM (numeral), PRT (particle), PRON (pronoun), VERB (verb )] [33]
    Wordbag → $U_1 [w_1 \ldots w_n] \cup U_2[w_1 \ldots w_n]$
    Pair every combination of tagged words {$wp_1 \ldots wp_m$} where
        $m = \frac{n!}{(n-wn)!wn!}$    (1)
    For every word pair {$wp_1 \ldots wp_m$ } calculate word similarity:
        If {$wp_m$ } are both fuzzy words then
            If {$wp_m$} are in the same fuzzy category, C where C = {Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership} then
                Calculate Lowest Common Subsumer depth, d, from associated fuzzy category ontology.
                Calculate path length, l, and the length of the shortest path between {$wp_m$} from the associated fuzzy category ontology
                Calculate word similarity, S between {$wp_m$}
                    $S(wpm) = e^{-\alpha l} \cdot \frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}$    (2)
                where $\alpha$ and $\beta$ were empirically determined as 0.15 and 0.85 respectively
            Else
                Apply original STASIS word similarity measure (2), calculating Lowest Common Subsumer depth, d and path length, l, from the WordNet ontology.
            End If
        Else
            Apply original STASIS word similarity measure(1), calculating Lowest Common Subsumer depth, d and path length, l, from the WordNet ontology.
            Apply fuzzy word association algorithm [12] to determine presence of fuzzy words and associated with the non-fuzzy words
            If Associated Fuzzy Words are Present then
                Calculate new Lowest Common Subsumer, d and length, l modifications
                Recalculate Word Similarity using (1)
            Else
                Return level of word similarity for {wpm}
            End If
            Return level of word similarity for {wpm}
        End If
        Calculate word frequency information using Browns Corpus statistics [3]
            $i(w) = 1 - \frac{\log(n+1)}{\log(N-1)}$    (3)
        where i(w) is the information weight, N is the total number of words in the Corpus and n is the words frequency.
    End for
    Calculate overall utterance similarity, S:
        $S(U_1, U_2) = \delta \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} + (1-\delta) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|}$    (4)
    with S being defined as the total sum of all possible values and $S_1$ and $S_2$ referring to pairs of semantic similarity vectors which were determined in (1) and r is a short joint word vector set vector comprising of word frequency information and word order
End for
```

V. Experimental Design

A) *Dataset Description*

In order to test the FUSE algorithm, three published datasets were used. These consisted of:

- Multi-Word Sentence Pair Fuzzy Dataset [MWFD]
- STSS 65 Sentence Pair [STSS_65]
- STSS 131 Sentence Pair [STSS_131]

MWFD consists of 30 sentence pairs that have two fuzzy words in each sentence. Sentences were taken from the Gutenberg Corpus [33] and random fuzzy words from the same category were substituted in each sentence to create this dataset

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

[12]. STSS_65 contained 65 short text sentence pairs and STSS_131 contained 131 short text sentence pairs. Both datasets are Gold Standard [2, 28].

B) *Experimental Methodology*

FUSE was run against each of the three datasets (MWFD, STSS_65 and STSS_131) and the sentence similarity results for each Sentence Pair [SP] was recorded. In order to be able to test the improvement of FUSE, all three datasets were also run with FAST and STASIS algorithms and the sentence similarity results for each SP was again recorded. Using Pearson's correlation coefficient [29], the correlation for each dataset was compared to the Average Human Ratings [AHR]. Pearson's correlation provides statistical evidence for a linear relationship between two variables *x* and *y* and can be computed as follows [29]:

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\,.\sqrt{\mathrm{var}(y)}} \qquad (5)$$

Where r_{xy} is the correlation coefficient, cov(x, y) is the sample covariance of x and y; var(x) is the sample variance of x; and var(y) is the sample variance of y.

Table V and Figure 4 show the correlation (r) of results recorded for the three datasets versus their AHR tested against STASIS, FAST and FUSE. The r-value should be between [-1 ... +1]. (-1) shows a perfectly negative linear relationship, (0) shows no relationship, and (+1) shows a perfectly positive linear relationship. A negative correlation will mean a decreasing relationship, while a positive correlation will mean an increasing relationship. The magnitude of the value (how close it is to -1 or +1) will indicate the strength of the correlation [29, 34].

TABLE V. CORRELATION RESULTS FOR DATASET

Algorithms / Datasets	STASIS	FAST	FUSE
MWFD	0.74525	0.73050	0.76820
STSS_65	0.68130	0.69080	0.69097
STSS_131	0.52078	0.51630	0.51799

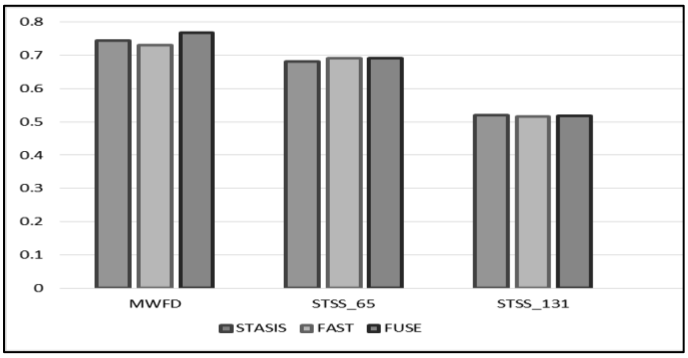


Fig. 4. Correlation Results for Datasets

C) *Results and Discussion*

Table V shows for MWFD, that FUSE gave a higher correlation (r = 0.76820) with human ratings compared to STASIS (r = 0.74525) and FAST (r = 0.73050). For STSS_65, FUSE gave a higher correlation coefficient (r = 0.69097) than both STASIS (r = 0.68130) and FAST (r = 0.68130), and for STSS_131, FUSE gave a higher correlation (r = 0.51799) than FAST (r = 0.51630). These can also be viewed in Figure 4. It was found that FUSE gave a higher correlation against both STASIS and FAST for the datasets MWFD and STSS_65.

Consider the following examples of SPs. The first is an example from the MWFD dataset.

[SP_{a1}] *So would useless diminutive Harriet*

[SP_{b1}] *So would poor little Harriet*

For MWFD, the r-value was STASIS r = 0.7141, FAST r = 0.9089, and FUSE r = 0.9647.

The second example SP is from the STSS_131 dataset.

[SP_{a2}] *If you continuously use these products, I guarantee you will look very young.*

[SP_{b2}] *I assure you that, by using these products consistently over a long period of time, you will appear really young.*

For STSS_131, the r-value was STASIS r = 0.8573, FAST r = 0.8021, and FUSE r = 0.8772.

From the two sentence pair examples it can be seen that FUSE provided better correlation (as evidenced by the r-value) compared to both STASIS and FAST. In addition, FUSE had better human ratings compared to FAST, which also helped with the improvement of the r-value. This can be shown using the two examples given. In MWFD, the words '*useless*' and '*poor*' had defuzzified values of (-0.695 and -0.65) respectively in FAST; however, in FUSE, those values were (-0.95862 and -0.89655) respectively. For STSS_131, the same also applies; the words '*young*' and '*consistently*' have values of (-0.45 and 0.4) respectively in FAST, and values of (-0.58969 and 0.4) respectively in FUSE; also the word '*continuously*' did not exist in FAST, but this word exists in FUSE with the value of (0.425). This goes to show that not only does the increased coverage of words in FUSE, with an almost 60% increase in words in total over the six categories compared to FAST, play an important part in giving a higher correlation; but the improved defuzzified values for the fuzzy words using Interval Type-2 allows better representation of the uncertainty of words in the context of FSSM and aligns to the findings that Interval Type-2 is the scientifically correct way to model linguistic uncertainties [35].

VI. CONCLUSION AND FURTHER WORK

In conclusion, the FUSE algorithms showed better correlation compared to human ratings than other similar algorithms on human utterances. The improvement FUSE had over STASIS and FAST for the three datasets of MWFD, STSS_65 and STSS_131 is down to several factors. Firstly, the coverage of words is far greater, with an increase of 57.65%. Secondly, a new set of fuzzy ontologies has been developed for these categories in FUSE. Finally, the ability to represent uncertainty using Interval Type-2, as opposed to Type-1 has

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

been shown to contribute towards a higher correlation between FUSE and human ratings. However, it is noted that in this kind of work, there is a degree of subjectivity in gathering human ratings. The results from FUSE are promising and will allow a deeper understanding of the semantic meaning, in context of human utterances by a machine, especially within Conversational Agents.

Future work will involve the incorporation of linguistic hedges, such as {very, mostly, slightly} etc. [8] into FUSE. Currently, hedges are not utilized in FSSMs. This will help further with precision of utterance similarity measurement, in that such words will make a weighted contribution when calculating the overall semantic similarity.

ACKNOWLEDGMENT

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

REFERENCES

[1] L. Nocks, "500 years of humanoid robots automata have been around longer than you think [Resources_Review]," *IEEE Spectrum*, vol. 54, no. 10, pp. 18-19, 2017.

[2] J. D. O'Shea, "A framework for applying short text semantic similarity in goal-oriented conversational agents," Doctorate of Philosophy, Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, 2010.

[3] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138-1150, 2006.

[4] D. Chandran, K. Crockett, D. Mclean, and Z. Bandar, "FAST: A fuzzy semantic sentence similarity measure," in Fuzzy Systems (FUZZ), 2013 *IEEE International Conference on*, 2013, pp. 1-8: IEEE.

[5] L. A. Zadeh, "Fuzzy logic= computing with words," *Fuzzy Systems, IEEE Transactions on*, vol. 4, no. 2, pp. 103-111, 1996.

[6] J. M. Mendel et al., "What computing with words means to me," *IEEE Computational Intelligence Magazine*, vol. 5, no. 1, pp. 20-26, 2010.

[7] J. M. Mendel, "Computing with words and its relationships with fuzzistics," *Information Sciences*, vol. 177, no. 4, pp. 988-1006, 2007.

[8] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Information sciences*, vol. 8, no. 3, pp. 199-249, 1975.

[9] J. M. Mendel and R. B. John, "Type-2 fuzzy sets made simple," *IEEE Transactions on fuzzy systems*, vol. 10, no. 2, pp. 117-127, 2002.

[10] J. M. Mendel, R. I. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *Fuzzy Systems, IEEE Transactions on*, vol. 14, no. 6, pp. 808-821, 2006.

[11] M. Hao and J. M. Mendel, "Encoding words into normal interval type-2 fuzzy sets: HM approach," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp. 865-879, 2016.

[12] D. Chandran, "The development of a fuzzy semantic sentence similarity measure," Doctorate of Philosophy, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2013.

[13] J. M. Mendel, "A comparison of three approaches for estimating (synthesizing) an interval type-2 fuzzy set model of a linguistic term for computing with words," *Granular Computing*, vol. 1, no. 1, pp. 59-69, 2016.

[14] A. Bilgin, H. Hagras, A. Malibari, M. J. Alhaddad, and D. Alghazzawi, "Towards a general type-2 fuzzy logic approach for computing with words using linear adjectives," in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, 2012, pp. 1-8: IEEE.

[15] J. M. Mendel and R. I. B. John, "Type-2 fuzzy sets made simple," *Fuzzy Systems, IEEE Transactions on*, vol. 10, no. 2, pp. 117-127, 2002.

[16] L. Meng, R. Huang, and J. Gu, "Measuring semantic similarity of word pairs using path and information content," *Int. J. Futur. Gener. Commun. & Netw*, vol. 7, pp. 183-194, 2014.

[17] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2003, pp. 241-257: Springer.

[18] D. Sánchez, D. Isern, and M. Millan, "Content annotation for the semantic web: an automatic web-based approach," *Knowledge and Information Systems*, vol. 27, no. 3, pp. 393-418, 2011.

[19] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 71-78: Association for Computational Linguistics.

[20] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for B2C eCommerce," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 946-950, 2008.

[21] J. Atkinson, A. Ferreira, and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts," *Knowledge-Based Systems*, vol. 22, no. 7, pp. 502-508, 2009.

[22] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133-138: Association for Computational Linguistics.

[23] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE transactions on systems, man, and cybernetics*, vol. 19, no. 1, pp. 17-30, 1989.

[24] C. Leacock, G. A. Miller, and M. Chodorow, "Using corpus statistics and WordNet relations for sense identification," *Computational Linguistics*, vol. 24, no. 1, pp. 147-165, 1998.

[25] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[26] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[27] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on knowledge and data engineering*, vol. 15, no. 4, pp. 871-882, 2003.

[28] J. O'Shea, Z. Bandar, and K. Crockett, "A new benchmark dataset with production methodology for short text semantic similarity algorithms," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, no. 4, p. 19, 2013.

[29] K. S. University. (2012, 09/12/2017). *SPSS Tutorials: Pearson Correlation*. Available: https://libguides.library.kent.edu/SPSS/PearsonCorr

[30] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[31] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*. Macmillan New York, 1993.

[32] J. M. Mendel and D. Wu, Perceptual Computing: Aiding People in Making Subjective Judgments. John Wiley & Son, 2010.

[33] P. Gomes et al., "The importance of retrieval in creative design analogies," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 480-488, 2006.

[34] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[35] J. M. Mendel, "Computing with words: Zadeh, Turing, Popper and Occam," *IEEE computational intelligence magazine,* vol. 2, no. 4, pp. 10-17, 2007.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:36:09 UTC from IEEE Xplore. Restrictions apply.

Application of Fuzzy Semantic Similarity Measures to Event Detection within Tweets

IEEE International Conference on Fuzzy Systems

2017

Application of Fuzzy Semantic Similarity Measures to Event Detection Within Tweets

Keeley Crockett, Naeemeh Adel, James O'Shea, Alan Crispin,
Intelligent Systems Group, School of Computing, Mathematics and Digital Technology,
Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK
K.Crockett@mmu.ac.uk

David Chandran
Institute of Psychiatry, Psychology & Neuroscience, Kings College London, 16 De Crespigny Park, London, SE5 8AF, UK
João Paulo Carvalho
INESC-ID / Instituto Superior Técnico, Universidade de Lisboa, Portugal

***Abstract*— This paper examines the suitability of applying fuzzy semantic similarity measures (FSSM) to the task of detecting potential future events through the use of a group of prototypical event tweets. FSSM are ideal measures to be used to analyse the semantic textual content of tweets due to the ability to deal equally with not only nouns, verbs, adjectives and adverbs, but also perception based fuzzy words. The proposed methodology first creates a set of prototypical event related tweets and a control group of tweets from a data source, then calculates the semantic similarity against an event dataset compiled from tweets issued during the 2011 London riots. The dataset of tweets contained a proportion of tweets that the Guardian Newspaper publically released that were attributed to 200 influential Twitter users during the actual riot. The effects of changing the semantic similarity threshold are investigated in order to evaluate if Twitter tweets can be used in conjunction with fuzzy short text similarity measures and prototypical event related tweets to determine if an event is more likely to occur. By looking at the increase in frequency of tweets in the dataset, over a certain similarity threshold when matched with prototypical event tweets about riots, the results have shown that a potential future event can be detected.**

Keywords- fuzzy semantic similarity measures, Twitter, semantic analysis

I. INTRODUCTION

Twitter continues to be key player in the social media market with reporting on average 317 million monthly active users in late 2016 [1] and has the opportunity to be a rich source of information for organisations. However, the most difficult challenge lies in extracting meaning from the unstructured and inherently noisy textual tweets. Typically, to extract useful information from tweets i.e. sentiment analysis [2,3], tweets undergo significant pre-processing that can include removal of all URLs, correction of spelling errors, tagging of named entities, removal of stop words, punctuation etc., acronym look up and even the removal of hashtags. The cleaned tweet can, in cases, project a different semantic meaning than what was intended in the context of the analysis being conducted.

Tweets are known as a mechanism for spreading news information fast. For example, tweets were influential during the Arab Spring uprising in 2010-2011 [3], allowing large groups of people to communicate quickly and organise protest rallies against regimes. Twitter also, perhaps more importantly, allowed information in the form of photographs and videos to be broadcast by members of the public. Munroe [4] suggested that people's communications over the Internet, especially Twitter could overtake an initial earthquake. This is because seismic waves travel a lot more slowly than data traveling along fiber optical cables. This effect has occurred a couple of times, more recently on 23rd August 2011 where a 5.9 magnitude earthquake struck close to Richmond, Virginia. The effects were initially felt in Washington D.C. where the initial tweets were posted, and various people reported having read them in New York City before the earthquake reached them [5,6].

Event detection and user profiling using Twitter is a predominant research area [2..13]. An event can be defined simply as an occasion of importance that happens at a given time. Sakaki et al [8] states that events have three key properties: they are often large scale, have an influence on an individual person's daily lifestyle and can have both spatial and temporal locations. Early research in [8] used support vector machines to classify a tweet into positive and negative classes in relation to the event being predicted. This classification was used to semantically analyse tweets and in conjunction with spatial estimation was incorporated into an earthquake reporting system. The authors [8] identified that the search query or term(s) used for classification are of vital importance and improvement in recall heavily relied on this factor. Pavlyshenko [9], used frequent sets, association rules and formal concept analysis to build semantic concepts between individuals, had more success detecting events that had fewer random factors. Arias et al [10] used summary decision trees and support vector machines to improve the power of Twitter forecasting models in predicting the stock market and box office revenue trends. Rui et al [11] proposed a Twitter based event detection and analysis system for crime and disaster related events. Ribeiro et al [12] proposed a method to identify traffic events and conditions using Twitter and report them in real time (with 50% to 90% accuracy). More recently, Polhl and Bouchachia [13] conducted a review of how social media networks were used to disseminate information during a crisis i.e. a police emergency. To the authors' knowledge, the incorporation of

978-1-5090-6034-4/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.

semantic similarity as an additional dimension to the models produced was not considered.

Semantics are concerned with the literal meaning of morphemes, words, phrases and sentences and the way that they are combined. Semantic similarity is therefore, a complex concept with a long history in cognitive psychology and linguistics [14.15], which can analyse the deep semantic structure of a short text to convey meaning. An operational definition often used in studies of semantic similarity is "How close do these two sentences come to meaning the same thing?" [16] and getting a machine to be able to answer the question in a similar way to a human being is particularly challenging.

Due to the inherent natural language of tweets, fuzzy sentence similarity measures (FSSM) are of particular interest in this work. FSSM are algorithms that are able to compare two or more short texts which contain human perception based words and return a numeric measure of similarity of meaning between them. FAST [15] (Fuzzy Algorithm for Similarity Testing) is an ontology based similarity measure that uses concepts of type 1 fuzzy sets to model relationships between categories of human perception based words (fuzzy words). Previous work has shown that FAST gives higher correlations with human ratings of similarity than leading other measures [17], which tend to ignore fuzzy properties of words when measuring similarity. To the knowledge of the authors, none of the work on event detection to date has measured the semantic similarity of the tweets using fuzzy short text similarity measures. A brief review of FAST can be found in Section II.

The aim of the research presented in this paper is to see if groupings of prototypical tweets about a potential event, i.e. a riot, can be used in conjunction with fuzzy short text similarity measures to detect where an event is more likely to occur. For the purpose of this work the chosen event is the London riots [18..19] which took place between the 6th and 10th August 2011, where the UK experienced riots at a level not seen since the eighties. Following the riots, not only did the UK Government announce a public enquiry, the Guardian newspaper began its own analysis which included an examination of the role of social media to try and establish whether Facebook and/or Twitter actually incited the riots [20..22]. This analysis involved examination of 2.57m tweets and concluded that tweets during the period were mainly used as a reaction mechanism. The research in our paper addresses the question that if the tweets of the perpetrators were known days before the actual riots and provided evidence that the riots were incited in some way, could these individuals have been brought to justice sooner? Also, could we potentially avoid or scale down the riots themselves? This would have positive effects for society, including reduced insurance claims. If we can detect that a potentially dangerous or criminal event is about to occur and identify who is initiating it and where it is likely to occur, we can then put measures in place to either stop it happening or reduce the consequences.

This paper is organised as follows; Section II provides an overview of fuzzy short text semantic similarity measures. Section III describes how the London Riot Data set was created and how the data was sampled for this work. Section IV presents a methodology for application of fuzzy semantic similarity measures in detecting potential events using semantic analysis experimental results and accompanying discussion are covered in Section V and finally Section VI presents the conclusions and future directions.

II. Overview Fuzzy Short Text Semantic Similarity Measures

FAST (Fuzzy Algorithm for Similarity Testing) [17] was developed to enable new human perception properties to be taken into consideration when short texts were analysed by a machine to determine their syntactic and semantic similarity. FAST was inspired by STASIS, a short text semantic similarity measure developed by Li et al [23], from which FAST adopted the path length and depth of words relative to their position in a set of fuzzy ontologies with the information content of individual words being derived from a corpus. These were used to form semantic vectors and were then combined with word order vectors (from the word order in each short text) to determine the semantic similarity [17]. FAST identified fuzzy words within a short text and calculated the effect such words would have on the overall similarity. Experimental results showed that FAST gave an improved correlation between the similarity measure and human ratings [18, 24] compared with traditional measures.

Essentially, FAST works through first applying a word similarity measure to every possible pair of words in a short text and using corpus statistics to determine the overall semantic similarity between two short texts. The key stages of the FAST measure are:

1) Tokenize every word in the two short texts. For example, a cleaned tweet (method outlined in section IV) ("UKRiots Those convicted include a primary school teacher a lifeguard a man who works for homeless charity and an 11 year old from Essex") is sorted into a list ["UKRiots", "Those", "convicted", "include", "a", "primary", "school", "teacher", "a", "lifeguard", "a", "man", "who", "works", "for", "homeless", "charity", "and", "an", "11", "year", "old", "from", "Essex"]
2) Pair every combination of tokenized words. A Bag of Words was [23] created as the union between all words within the two short texts which similarity is being measured.
3) Determine the similarity of each word pair. If the word pair comprises of only fuzzy words i.e. [young, old] then use fuzzy category based ontologies to determine path length else use WordNet [25] as the semantic knowledge base to calculate path length.
4) If the word pair contained non-fuzzy words i.e. [teacher, man], determine effect of associated fuzzy words i.e. "small man".

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.

5) Apply sentence similarity measure using word similarities from different word pair combinations from 3) and 4).

Developed from an established traditional sentence similarity measure known as STASIS [24], FAST also incorporates an empirically determined semantic threshold, α, which was used to filter out word pairs with very low similarity scores Li et al [24] justified the use of a semantic threshold, particularly when short texts were very short in length. The work also determined that function words (words that express grammatical relationships with other words within a short text [24] i.e. 'do'), also carried syntactic information and were to be included in the semantic similarity measurement. Typically, tweets are short in length (i.e. 140 characters per tweet or less excluding multimedia). In this work, the application of FAST to determine the semantic similarity to tweets will require the empirical evaluation of a suitable semantic threshold. FAST is fully automatic without requiring the users' intervention and readily adaptable across the range of potential application domains. For the purpose of this research FAST will be used to measure the similarity of Twitter tweets to a set of prototypical tweets in an attempt to detect a potential future event.

III. Creating the London Riot Data set

In order to evaluate the use of FSSM in its suitability to detect possible future events from tweets, it was necessary to construct a dataset to investigate whether or not an event could be predicted. Given the number of tweets generated on a daily basis, it was essential that the dataset contained a balanced proportion of tweets that concerned a particular event. The event selected for this study was the London Riots which occurred between the 6th and 10th August 2011. The riots were seen to be triggered by the shooting of 29-year-old father of four Mark Duggan by the police. The Guardian Newspaper publically released some Twitter data that included a list of 200 influential Twitter users based on re-tweets during the riot period [21]. It also included a list of the most popular Hashtags - relevant keywords, acronyms or phases in order to allow the tweet to be categorised. The dataset, known as the London-Riot dataset was initially populated with tweets from users identified using the Guardian data. The dataset was then expanded by selecting users, which appeared using the Twitter REST API public feed. For each user, tweets were recorded which were created up to and after 1st August 2011 at midnight, or up to the 3,200 tweet limit from the REST API statuses/user_timeline limitation (if the user had posted more than 3,200 tweets since 1st August 2011). A total of 9,913,397 tweets were collected from 8,819 Twitter users.

Due to the time taken to process this quantity of data, using available equipment this dataset was further reduced in size. A total of 1,132,938 individual tweets were extracted between 1st August 2011 00:00:00 and 31st August 2011 23:59:59 to create a new dataset which will be referred to in this work as the Twitter Riot dataset. The quantity of Guardian riot tweets which appeared in the results was 17,795 tweets – a total of 4.6% of the tweets collected were sourced from users listed in the Guardian data. Samples of tweets from an 11 day period were then extracted to test groupings of prototypical tweets. Details are provided in Section IV.

IV. Detecting Events using Fuzzy Short Text Semantic Similarity Measures

This section describes a study that was conducted to test the following hypothesis:

H1: Can Twitter tweets be used in conjunction with fuzzy short text similarity measures and prototypical event related tweets to determine if an event is more likely to occur.

The section first outlines the overall methodlogy and then describes how the prototypical tweets for an event were sampled and a control group formulated.

A. Methodology

Let Event dataset be a generic name used to define a set of tweets that is to be used to investigate the likelihood of an event occurring that is time stamped. For this work, 2200 tweets were randomly sampled without replacement from the Twitter Riot Dataset; 200 tweets were randomly sampled from the period 1st to 11th August to investigate before, during and after the riots started on the 6th April 2011. The methodology for using fuzzy short text semantic similarity measures to detect potential events is defined as follows:

1. Select a series $\{1..k\}$ where $1 \leq k \geq m$ of prototypical tweets, T concerning an event, where m is the maximum number of associated prototypical tweets and is empirically defined. In this work k = 7. Each tweet, t is between 1 and 25 words in length.
2. For all tweets in the Event dataset $\{1..n\}$ where n is the number of tweets, calculate the fuzzy semantic similarity, Si between every tweet, tn, per day stored from the *start-date* to the *end-date* of the month of the event and prototypical tweet km using the pre-selected FSSM.
3. Using the short text semantic similarity measure, Si, plot a graph showing the following:
   a. The total number of tweets stored for each day between the start-date and end-date.
   b. The number of tweets per day where the similarity Si of ti is greater than a given semantic threshold, α, *where* $0.5 \leq \alpha \geq 0.7$ in 0.05 increments. A semantic threshold, α, of 0.5 was chosen as an initial starting point, as this will result in matches which have a moderate to high similarity with the comparative short text.
4. Identify if there is sufficient increase in the frequency of tweets in the dataset, over a specific similarity threshold during the event. This sufficient increase will be

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.

identified if the comparison result, from one day to another, is greater than 0.5%.

No cleaning of the tweets took place prior to running through FAST. FAST removes symbols, such as ($%^|&* etc.) from the short texts, leaving only letters and numbers. For example, the anonymized tweet "\@XXX77 Tottenham has a notorious past for riots, Im sure it aint the last time, :0(" becomes "XXX77 Tottenham has a notorious past for riots, Im sure it aint the last time 0". This ensures word order, path length and the inclusion of function words is maintained – all required to determine the similarity. Hashtag words are also left in place with "#Riots:" becoming "Riots".

B. Selection of Prototypical Tweets

In order to evaluate if FAST could be used to detect if a potential event was more likely to occur, two groups of tweets were selected. The first group contained 7 tweets that were related to the type of event that was to be potentially detected – in this case riots. Tweets ID's 1 to 7 (Table I) were randomly selected from a study on *"Twitter, Information Sharing and the London Riots"* [22] which analyzed 600,000 tweets and retweets about the London riots to investigate whether Twitter was used as a tool to promote illegal group actions. The second group, known as the control group (Tweet ID's 8 to 14) contained a further 7 tweets which were randomly sampled from top tweets of 2011 [26]. The 14 prototypical and control group tweets can be seen in Table I.

C. Experimental Methodology

All 2200 tweets were ran against the 7 prototypical event tweets and the 7 control tweets shown in Table I, for each of the semantic thresholds and the similarity of each was recorded. In order to identify if the semantic similarity of the tweets indicated if an event was likely to occur, each day's tweets which matched the cumulative prototypical tweets over a specific similarity threshold, α had to be scaled as a % of that day's tweets. The relative number of tweets on that given day defined as

$$\%relative\ tweets = \frac{tweets\ measuring\ above\ \alpha}{tweets\ processed\ for\ given\ date} * 100 \quad (1)$$

Where α is in {0.5. 0.55, 0.6, 0.65, 0.7}.

V. Results and Discussion

A. Results

The results were grouped into the date that each tweet was posted, in order for a day-by-day comparison. Tables II and III show the % relative tweets for both the riot prototypical tweets and the control tweets over the eleven-day period, along with the semantic threshold used in each experiment. Higher sematic similarity thresholds did not yield an increase, emphasising the need that prototypical event tweets need to be more generalised to an event type and not a specific occurrence of an event. From Table II, it can be seen that during the days preceding the first riot on the 6th August and in the days afterwards when the rioting spread to further cities in the north (6th to 11th) there was a higher number of tweets which matched the prototypical tweets with high semantic similarity. The higher the threshold, the more semantically similar is the tweet to one of the prototypical tweets from Table I. In comparison, in Table III, it can be observed that the seven control group tweets, when α = 0.50, the % relative control group tweets remains low.

TABLE I. Prototypical Event Tweets and Control Group Tweets

Tweet ID	**Prototypical Event Tweets**
1	There are young people rioting, smashing cars and vandalizing buildings.
2	The Bullring Shopping Centre has been closed amidst fears of looting and rioting. Large police presence
3	Don't understand how people think they can just hear of rioting and go down and loot! Man on news says it was older people!
4	I'm glad I'm in a peaceful country where people respect each other while the UK burns! Philippines
5	Sending in army may clear streets but it would be a sign of major political weakness for Cameron London riots
6	Don't call them anarchists. Anarchy is a political philosophy. This is just shopping with no rules. Call them capitalist
7	Rioting & looting has spread across UK – London
	Control Event Tweets
8	Welcome back Egypt Jan 25
9	Helicopter hovering above Abbottabad at 1AM (is a rare event)
10	my daughter her name is sarah m. rivera
11	This lockout is really boring..anybody playing flag football in Ok..I need to run around or something!
12	Brooms up London!
13	Here's another Photo of the shuttle from my plane.
14	Earthquake

TABLE II. %Relative Prototypical Tweets (1 to 7) using FAST

Day	**α = 0.50**	**α = 0.55**	**α = 0.60**	**α = 0.65**	**α = 0.70**
1	13.21	5.21	1.21	0.21	0
2	12.21	4.57	1.35	0.14	0
3	14.78	6.57	2.42	0.35	0
4	14.85	7.21	2.21	0.78	0.07
5	11.64	4.35	1.07	0.14	0
6	18.64	9	2.71	0.57	0
7	17.71	8.42	3.71	0.85	0.14
8	18.28	10.41	4.85	1.85	0.35
9	17.28	7.92	3.07	0.57	0.21
10	23.21	11.92	4.71	1.14	0
11	32.92	18.92	9.42	3.35	0.71

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.

TABLE III. %RELATIVE CONTROL GROUP TWEETS (8 TO 14) USING FAST

Day	α = 0.50	α = 0.55	α = 0.60	α = 0.65	α = 0.70
1	3.92	0.92	0.28	0	0
2	2.35	0.28	0	0	0
3	5.14	1.85	0.21	0.07	0
4	4.71	2.07	0.42	0.07	0
5	3.07	1.07	0.21	0	0
6	4	1.07	0.42	0.07	0.07
7	3.28	0.92	0.28	0.14	0.07
8	3.35	0.92	0.14	0	0
9	2	0.5	0.14	0	0
10	4.92	1.35	0.28	0.07	0
11	6.57	2.07	0.5	0	0

Figures 1 and 2 visually show the effect of matching the prototypical event tweets (Table I) with varying semantic threshold over the 11 day period. The y-axis shows the % of tweets relative to each result range (i.e. similarity threshold α >0.5 compared to α >0.7) for each day which was calculated in equation 1. The x-axis shows the days 1 to 11. The graph in Figure 1 shows the event clearly peaking on the 6th August and then again showing growth up to the 11th August although this is predominantly with semantic thresholds 0.65.

The graphs in figures 1 and 2 show an increase in matched tweets for higher similarity thresholds for the days of the London riots on the 6th August, which remains relatively stable to the 9th and then sharply increases which corresponds to the triggering of further riots occurring across the rest of the UK. On examining tweets in the sample from day 11, it was observed that the tweets that matched with a higher similarity threshold were focused on clean-up operations undertaken by the general public [25]. These consisted of tweets not only from influential riot Twitter users [3], but also from those who were not. Interestingly where the similarity threshold is >0.6 (Table II) there is an increase in similarity of tweets prior to the events; from 2.7% to 3.7% on the day before the riot and the first day of rioting. The lower α, and hence the more general the semantic match, the more evident the rise between days. The obvious dip in the %relative tweets on day 5 can only be due to the sample selection.

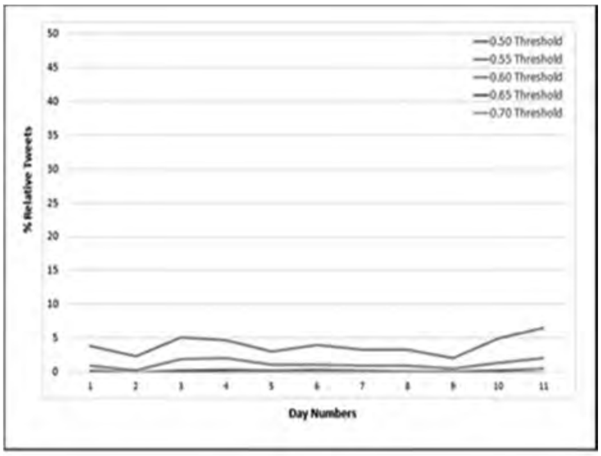


Fig.1. %Relative Prototypical Tweets (1 To 7) Using FAST

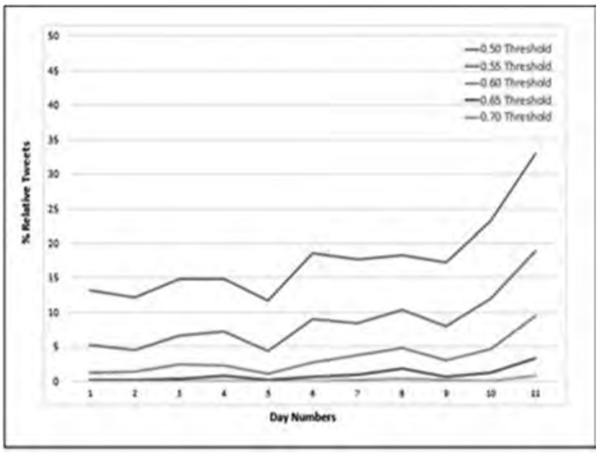


Fig.2. %Relative Control Tweets (8 To 14) Using FAST

In order to identify the day-to-day trends, the scaled results from day p were subtracted from the scaled results from day *p-1*, where *p* is the day of the month. As a result, the similarity threshold trends are shown in table IV for prototypical event tweets and Table V for control group tweets.

TABLE IV. COMPARISON OF PROTOTYPICAL EVENT TWEET SIMILIARITY BETWEEN DAYS

Day	α = 0.50	α = 0.55	α = 0.60	α = 0.65	α = 0.70
1 to 2	-1	-0.64	0.14	-0.07	0
2 to 3	2.57	2	1.07	0.21	0
3 to 4	0.07	0.64	-0.21	0.43	0.07
4 to 5	-3.21	-2.86	-1.14	-0.64	-0.07
5 to 6	7	4.65	1.64	0.43	0
6 to 7	-0.93	-0.58	1	0.28	0.14
7 to 8	0.57	1.99	1.14	1	0.21
8 to 9	-1	-2.49	-1.78	-1.28	-0.14
9 to 10	5.93	4	1.64	0.57	-0.21
10 to 11	9.71	7	4.71	2.21	0.71

TABLE V. COMPARISON OF CONTROL EVENT TWEET SIMILARITY BETWEEN DAYS

Day	α = 0.50	α = 0.55	α = 0.60	α = 0.65	α = 0.70
1 to 2	-1.57	-0.64	-0.28	0	0
2 to 3	2.79	1.57	0.21	0.07	0
3 to 4	-0.43	0.22	0.21	0	0
4 to 5	-1.64	-1	-0.21	-0.07	0
5 to 6	0.93	0	0.21	0.07	0.07
6 to 7	-0.72	-0.15	-0.14	0.07	0
7 to 8	0.07	0	-0.14	-0.14	-0.07
8 to 9	-1.35	-0.42	0	0	0
9 to 10	2.92	0.85	0.14	0.07	0
10 to 11	1.65	0.72	0.22	-0.07	0

Plotting the results of Table IV as shown in Figure 3 shows that we can identify any day-to-day rises from previous day-to-day tweeting i.e. from the 5th to the 6th of August there was a rise of 0.21% for the similarity threshold of α >0.6. In the days leading up to the riot it can be seen that changing the similarity threshold does not yield any significant changes in

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.

tweets matching the prototypical event tweets. Figure 3 shows a clear spike between the 5th and 7th August around the date of the riot and also a further increase between the 9th and 11th again corresponding to further riots that were triggered and the public clean-up operation. Figure 4 visually shows the day-to-day trends of the control group tweets where it was observed that there is no significant difference over the 11 day period.

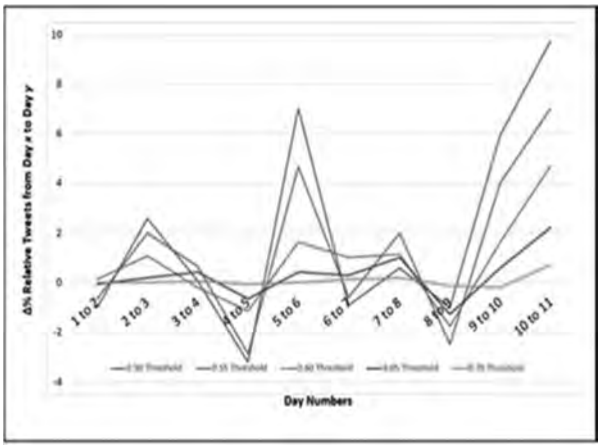


Fig 3. Identifying day-to-day trends surrounding an event (Prototypical Tweets)

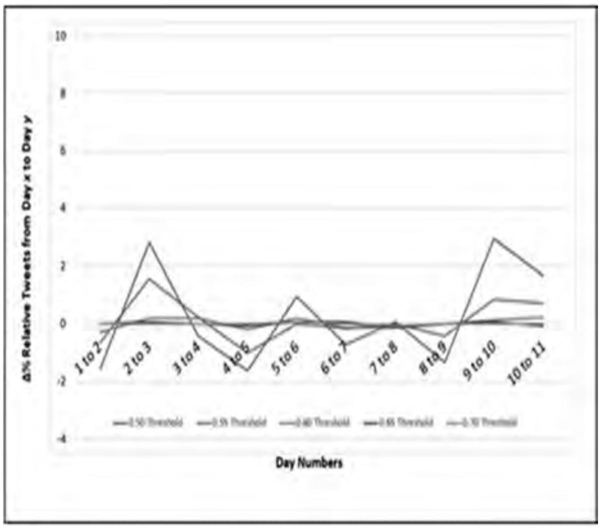


Fig 4. Identifying day-to-day trends surrounding event (control tweets)

B. Discussion

The experiments have shown that an event can be potentially detected, through using a set of prototypical tweets about the event type, by measuring the rise in similarity between the prototypical event tweets. The significance is dependent on the set of prototypical tweets used and the similarity threshold of the FAST measure. It was found that the prototypical tweets used which were extracted from the original sample of 17,795 tweets contained words that were specific to UK riots i.e. London, UK which are within Wordnet [25] which suggests that more general prototypical event tweets would need to be used to produce a more general set of event specific prototypical tweets. The results in figure 3 show that a similarity threshold of $\alpha > 0.55$ with FAST is sufficient to indicate an event, when comparing the change of percentage between two periods of time. However, a more generalised set of prototypical tweets may have yielded a high semantic similarity threshold. In comparison, there were no significant day to day trends using the control group tweets. A clear advantage of using a FSSM (identified from previous work on human correlations[17]) is that words such as ‘young’, ‘older’, ‘rare’ and ‘major’ highlighted only from the prototypical event tweets would allow their semantic meaning in the context of the tweet syntax to contribute towards the overall similarity of two comparison tweets. Therefore, produce a measurement more in line with human interpretation.

VI. Conclusions

The overall conclusion from the experiments conducted is that using a fuzzy semantic similarity measure such as FAST, makes it possible to detect potential events using fuzzy short text semantic similarity measures and prototypical event tweets. This confirms the hypothesis H1 is true. The changes in semantic similarity measurements between dates are able to indicate a potential event. This is based on the assumption that trends in rises of matches with a set of prototypical event tweets is indicative of a potential future event. However, the semantic threshold required to show a potential event was lower than expected, typically = 0.55. FAST was designed to be used on short texts such as sentences which have an established structure and when calculating the total similarity, inherited the weightings between the semantic part and the syntactic component from STASIS [23]. Given that these weights were designed through empirical experiments on structured sentences and not unstructured texts such as tweets, further work will example the weightings between the semantic, syntactic and fuzzy components within FAST. In order to validate the methodology for detecting possible events using FSSM, further experiments would need to be carried out to see if similar patterns occur with other historical events.

Many police departments around the world currently monitor "social media risk” with different degrees of success [27]. A recent report suggested that using such tools, they had the … “potential to remove any bias from the picture presented…” [28]. the report also highlighted that “… police force representatives thought it may not be as flexible as their more qualitative approach…” and they could be improved by adding further information. Hence, further work should seek to integrate analysis of the semantic meaning of tweets into larger social network analysis systems.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] The statistics Portal. [online], Available at https://www.statista.com/statistics/282087/number-of-monthly-active-Twitter-users/ [Accessed 12/1/2017], 2017.

[2] Ifrim, G., Shi, B. & Brigadir, I. (2014), Event detection in Twitter using aggressive filtering and hierarchical tweet clustering, in `Second ACM Workshop on Social News on the Web (SNOW), Seoul, Korea, 2014.

[3] Morozov, E., Guardian.co.uk Facebook and Twitter are just places revolutionaries go. [Online] Available at: http://www.guardian.co.uk/commentisfree/2011/mar/07/facebook-Twitter-revolutionaries-cyber-utopians [Accessed 12/1/2017], 2011.

[4] Munroe, R. [Online] Available at http://blog.xkcd.com/2011/08/24/earthquakes/ [Accessed11/1/2017]

[5] Ford, R. Hollywood Reporter. [Online] Available at: http://www.hollywoodreporter.com/news/earthquake-Twitter-users-learned-tremors-226481 [Accessed 12/1/2017]

[6] Gupta, M. Li, R. Chang, K. Towards a Social Media Analytics Platform: Event Detection and Description for Twitter – a Tutorial, 23rd International WWW Conference, [Online] Available at: http://www2014.kr/asset/slide/Towards%20a%20Social%20Media%20Analytics%20Platform.pdf, [Accessed 11/ 1/ 2017], 2014

[7] Katragadda, S, Virani, S, Benton, R. Raghavan, V. Detection of event onset using, IEEE IJCNN, DOI: : 10.1109/IJCNN.2016.7727381, 2016.

[8] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." In Proceedings of the 19th international conference on World Wide Web, pp. 851-860. ACM, 2010.

[9] Pavlyshenko, B. Forecasting of Events by Tweet Data Mining." arXiv preprint arXiv: 1310.3499 [Online] Available at: http://arxiv.org/abs/1310.3499, [Accessed 12/1/2017], 2013.

[10] Arias , M. Arratia, A. Xuriguera , R, Forecasting with Twitter data, ACM Transactions on Intelligent Systems and Technology (TIST) - Special Section on Intelligent Mobile Knowledge Discovery and Management Systems and Special Issue on Social Web Mining archive , Vol5:1, DOI: 10.1145/2542182.2542190. 2013

[11] Rui, L. Lei, K. Khadiwala, R. Chang, K. TEDAS: A Twitter-based Event Detection and Analysis System, IEEE 28th International Conference on Data Engineering, pp,1273 – 1276, 2012.

[12] Sílvio S. Ribeiro, Jr., Clodoveu A. Davis, Jr., Diogo Rennó R. Oliveira, Wagner Meira, Jr., Tatiana S. Gonçalves, and Gisele L. Pappa, Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '12). ACM, pp.5-11. DOI=10.1145/2442796.2442800, 2012.

[13] Pohl, D. Bouchachia, A. Information Propagation in Social Networks During Crises: A Structural Framework, Propagation Phenomena in Real World Networks, Vol:85, Intelligent Systems Reference Library, pp 293-309, 2015

[14] Rubenstein, H. Goodenough, j. Contextual Correlates of Synonymy. Communications of the ACM, 8, pp.627-633, 1965.

[15] Chandran, D. Crockett, K. Bandar, Z. Mclean, D. FAST: A Fuzzy Semantic Sentence Similarity Measure, accepted for the IEEE International Conference on Fuzzy Systems, India, Digital Object Identifier :10.1109/FUZZ-IEEE.2013.6622344, 2013

[16] O'shea, James, Zuhair Bandar, and Keeley Crockett. "A new benchmark dataset with production methodology for short text semantic similarity algorithms." ACM Transactions on Speech and Language Processing (TSLP) 10.4 (2013): 19.

[17] Crockett, K. Chandran, D, Mclean, D. On the Creation of a Fuzzy Dataset for the Evaluation of Fuzzy Semantic Similarity Measures, IEEE WCCI – FUZZY Systems, China, pp. 752..759, DOI: 10.1109/FUZZ-IEEE.2014.6891571, 2014.

[18] BBC, As it happened: England riots day five. [Online] Available at: http://www.bbc.co.uk/news/uk-14449675 [Accessed 13/6/2014, 2011

[19] BBC, 2011. Riots in Tottenham after Mark Duggan shooting protest. [Online] Available at: http://www.bbc.co.uk/news/uk-england-london-14434318 [Accessed 13 Jun 2014], 2011.

[20] Evans, L., 2011. 200 most influential Twitter users during the riots: are you on the list. [Online] Available at: http://www.guardian.co.uk/news/datablog/2011/dec/08/riot-Twitter-top-200 [Accessed 13 Jun 2014], 2011.

[21] The Guardian and LSE. Reading The Riots: Investigating England's Summer Of Disorder [Online] Available at: http://www.guardian.co.uk/uk/series/reading-the-riots. [Accessed 13th June 2014], 2011.

[22] Tonkin, E. Pfeiffer, H. Tourte, G. Twitter, Information Sharing and the London Riots? by [Online] Available: https://www.asis.org/Bulletin/Dec-11/DecJan12_Tonkin_Pfeiffer_Tourte.pdf. [Accessed 12/1/2017], 2014.

[23] Li, Y. Mclean, D. Bandar, Z. O'Shea, J. Crockett, K. “Sentence similarity based on semantic nets and corpus statistics”, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 8, pp.1138-1150, 2006.

[24] Chandran, D, Crockett, K. Mclean, D, An Automatic Corpus Based Method for a Building Multiple Fuzzy Word Dataset, IEEE-FUZZ 2015, DOI: 10.1109/FUZZ-IEEE.2015.7337877, 2015.

[25] Princeton University, About Wordnet. [Online] Available at: http://wordnet.princeton.edu/ [Accessed 13 June 2014], 2017.

[26] ABC news. The Year According to Twitter: 2011's Top Tweets Available [Online] Available: http://abcnews.go.com/Technology/year-Twitter-2011s-top-tweets/story?id=15065335 [Accessed 12/1/2017], 2012.

[27] As Police Monitor Social Media, Legal Lines Become Blurred, by Martin Kaste, February 28, 2014 8:39 PM ET, [Online] Available: http://www.npr.org/blogs/alltechconsidered/2014/02/28/284131881/as-police-monitor-social-media-legal-lines-become-blurred [Accessed 22 January 2017], 2014.

[28] Gunnell, D. Hillier, J. Blakeborough, L. Social Network Analysis of an Urban Street Gang Using Police Intelligence Data, Research Report 89, The Home Office. [Online]. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/491578/horr89.pdf. [Accessed 22/01/2016], 2016.

Authorized licensed use limited to: Manchester Metropolitan University. Downloaded on April 05,2022 at 17:37:10 UTC from IEEE Xplore. Restrictions apply.