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

João Paulo Carvalho
Instituto Superior Técnico, Technical University of Lisbon, 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

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

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 $\lambda \to \infty$, 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

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

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

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%

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

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):*

$$COG = \frac{\left(\left(\frac{a+b}{2}\right)+\left(\frac{c+d}{2}\right)\right)}{2} \quad (3)$$

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).*

$$y = a + \frac{(x-A)(b-a)}{B-A} \quad (4)$$

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).

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

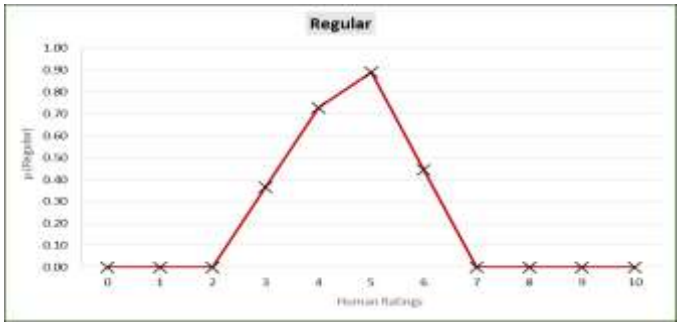

Fig. 1. Defuzzified Figure for 'Regular'

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

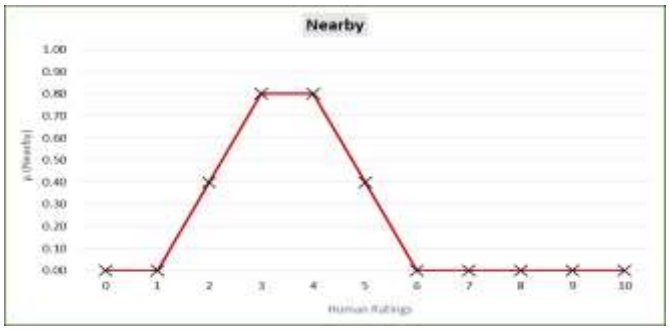

Fig.2. Defuzzified Figure for 'Nearby'

IV. FUSE (FUZZY SIMILARITY MEASURE)

This section first defines how the fuzzy category ontologies are constructed and then defines the proposed FUSE algorithm.

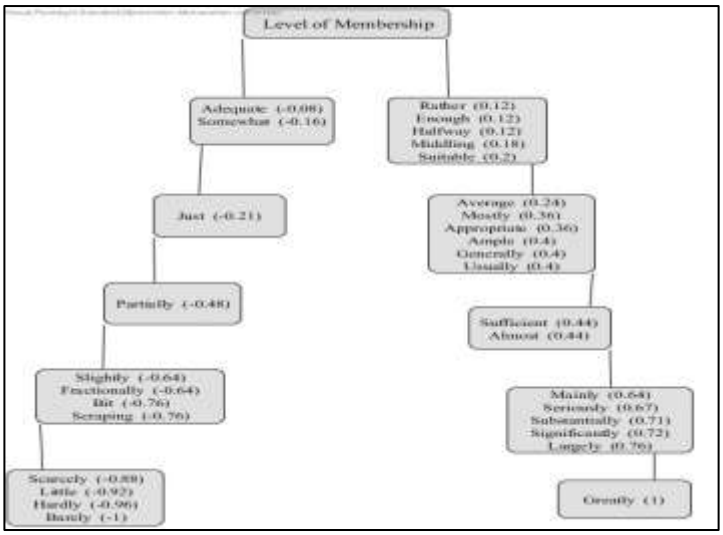

Fig. 3. Ontology for Level of Membership

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 below, 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

1. ***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
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]*
3. *Wordbag* → $U_1[w_1 \ldots w_n] \cup U_2[w_1 \ldots w_n]$
4. *Pair every combination of tagged words* $\{wp_1 \ldots wp_m\}$ *where*

$$m = \frac{n!}{(n-wn)!wn!} \quad (1)$$

5. ***For*** *every word pair* $\{wp_1 \ldots wp_m\}$ *calculate word similarity:*
6. ***If*** $\{wp_m\}$ *are both fuzzy words* ***then***
7. ***If*** $\{wp_m\}$ *are in the same fuzzy category, C where C =* {Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership} ***then***
8. *Calculate* Lowest Common Subsumer *depth, d, from associated fuzzy category ontology.*
9. *Calculate path length, l, and the length of the shortest path between* $\{wp_m\}$ *from the associated fuzzy category ontology*
10. *Calculate word similarity, S between* $\{wp_m\}$

$$S(wpm) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (2)$$

where α and β *were empirically determined as 0.15 and 0.85 respectively*

11. ***Else***
12. *Apply original STASIS word similarity measure (2), calculating* Lowest Common Subsumer *depth, d and path length, l, from the WordNet ontology.*
13. ***End If***
14. ***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

15. ***If*** *Associated Fuzzy Words are Present* ***then***

Calculate new Lowest Common Subsumer, d and length, l modifications

16. *Recalculate Word Similarity using (1)*
17. ***Else***
18. *Return level of word similarity for {wpm}*
19. ***End If***
20. *Return level of word similarity for {wpm}*
21. ***End If***

Calculate word frequency information using Browns Corpus statistics [3]

$$i(w) = 1 - \frac{\log(n+1)}{\log(N-1)} \quad (3)$$

where i(w) is the information weight, N is the total number of words in the Corpus and n is the words frequency.

22. ***End for***
23. *Calculate overall utterance similarity, S:*

$$S(U_1, U_2) = \delta \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} + (1-\delta) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (4)$$

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*

24. ***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

[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]:

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)}.\sqrt{\text{var}(y)}} \quad (5)$$

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

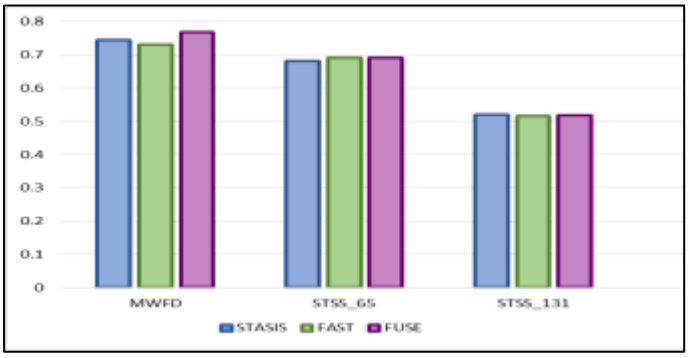

Fig. 4. Correlation Results for Datasets

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

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.