On the Creation of a Fuzzy Dataset for the Evaluation of Fuzzy Semantic Similarity Measures

David Chandran, Keeley Crockett, David Mclean
The Intelligent Systems Group, School of Computing, Mathematics and Digital Technology,
The Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK
K.Crockett@mmu.ac.uk

***Abstract*—Short text semantic similarity (STSS) measures are algorithms designed to compare short texts and return a level of similarity between them. However, until recently such measures have ignored perception or fuzzy based words (i.e. very hot, cold less cold) in calculations of both word and sentence similarity. Evaluation of such measures is usually achieved through the use of benchmark data sets comprising of a set of rigorously collected sentence pairs which have been evaluated by human participants. A weakness of these datasets is that the sentences pairs include limited, if any, fuzzy based words that makes them impractical for evaluating fuzzy sentence similarity measures. In this paper, a method is presented for the creation of a new benchmark dataset known as SFWD (Single Fuzzy Word Dataset). After creation the data set is then used in the evaluation of FAST, an ontology based fuzzy algorithm for semantic similarity testing that uses concepts of fuzzy and computing with words to allow for the accurate representation of fuzzy based words. The SFWD is then used to undertake a comparative analysis of other established STSS measures.**

I. INTRODUCTION

THE fields of natural language processing and sentence similarity have, since their inception, had a major impact on a wide range of areas of computer science and artificial intelligence. In the field there is a requirement for the comparison of sets of short text to determine the level of similarity between them which is achieved through the use of a short text semantic similarity (STSS) measure. The earliest STSS measures determined similarity based on the comparison of syntax [1]-[3] between sets of text. These measures worked by looking at common words between the two texts that were being compared and determining the distances between them. The distances between these common words can be used to determine a similarity vector giving a representation of the level of similarity for the two compared texts. There was however an issue with these early measures that limits the accuracy of their analysis. While they are capable of representing the level of syntactic similarity, they were incapable of accurately representing the level of semantic similarity between two sets of text. This limits these algorithms to the superficial similarities between texts while not being able to determine the effect of their semantic meanings on the overall level of similarity. In 2006, a new STSS measure called STASIS [4]-[5] was proposed for the specific purpose of accurately representing the level of similarity between short pieces of text. This method determined the level of similarity between two sentences through the use of ontological relations between words using Wordnet [6] - a large lexical database that contains ontological relations between large numbers of entities.

Since the establishment of STASIS a number of other similarity measures have been created [7]-[10]. Islam and Inkpen [8] avoided the use of ontologies by devising a method combining corpus statistics and string matching. The string matching component used a rule-based mechanism to determine semantic similarity based on specific structural similarities and differences between strings within sets of texts. The OMIOTIS [9] measure utilized both corpus statistics and the WordNet based thesaurus approach by considering the relative distances of words in a semantic network. More recent offerings include SEMILAR [10] – a semantic similarity toolkit which incorporates a number of text similarity measures. The toolkit currently only looks at similarity between nouns and verbs. Many of these new similarity measures have adopted the corpus-based approach towards sentence similarity, with varying levels of success. However, none of the STSS measures prior to 2013 have explicitly addressed the challenge of perception based or fuzzy words [11] in the calculation of similarity. In this work we define a fuzzy word as an imprecise word in natural language which may be vague in meaning, ambiguous and has context dependence [12]. Fuzzy words include but are not limited to the linguistic values which a linguistic variable may take [13]. For example, the linguistic variable temperature may have values {very hot, hot, lukewarm, cold} depending on the context.

To address the challenge of incorporating fuzzy words into similarity measures, the solution would be to develop new measures, which incorporated Zadeh's Computing with Words (CWW) framework [14] through the representation of human perceptions using fuzzy sets. Research into fuzzy theory and CWW presents vital concepts that can be used towards the goal of finding representations of natural language or fuzzy words that are used by humans. Through acknowledging that different people have different interpretations of fuzzy words and that they have no singular

qualities, their values can instead be represented with fuzzy sets. Therefore, the work that has been done on CWW allows for the generation of a method to use representations of the values of fuzzy words to determine their similarity and from that create a fuzzy sentence similarity measure. Further expansion on Zadeh's work in CWW came from Mendel who applied fuzzy type-2 methods to CWW [13]-[14]. Mendel noted that perceptions around words differed from individual to individual, which should be represented. The use of type-2 fuzzy sets allowed for the representation of the range of different perceptions about a particular word that allowed for the collection of type-1 fuzzy sets from a range of people to become elements of a type-2 fuzzy set. This could then be defuzzified, to return a single value. Incorporating fuzzy or perception based words has only been recently addressed in the creation of specific fuzzy word [16] and fuzzy semantic sentence similarity measures (FAST) [17]. Such measures will be briefly described in section II.

Evaluation of STSS measures has involved testing the measures against existing published datasets. Specifically recognized data sets published for the purpose of word and sentence similarity measure evaluation include Miller and Charles [18][19], Rubenstein and Goodenough [20] and O'Shea [21]-[23]. The creation of such datasets enabled the development of a methodology for which other datasets could be created [24]. In creating a STSS benchmark dataset, O'Shea [23][24] identified two desirable properties. The first is the precision and accuracy of the judgments by human participants in obtaining similarity ratings of sentence pairs. The second, being the scale on which the similarity measures are made i.e. an absolute zero point (unrelated in meaning) to a maximum (identical in meaning). Expanding on the work done by Miller and Charles and Rubenstein and Goodenough, O'Shea created a dataset of quantified pairs of sentences, SPSS-65 [23] which was followed by the SPSS-131 dataset [24][25]. Unfortunately, none of the existing datasets contained a suitable number of fuzzy words which would allow a fair and unbiased comparison of fuzzy semantic sentence similarity measures.

This paper proposes a methodology to construct a single fuzzy word dataset, which contains a set of sentences containing one fuzzy word per sentence, the Single Fuzzy Word Dataset (SFWD). The creation of the SFWD involved fuzzification of sentences in an existing dataset of sentence pairs [24][25] which had already been used to evaluate the STASIS and LSA sentence similarity measures [4]. The SFWD dataset is then presented along with ratings generated from a set of human participants on each sentence pair based on its level of semantic similarity. The SFWD is then used in a comparative evaluation of three STSS measures: STASIS, LSA and FAST to determine the effect of perception based words when computing semantic sentence similarity.

This paper is organized as follows; Section II provides a brief discussion of related work in word and semantic similarity measures including a description of FAST. Section III describes the methodology for the creation of a new dataset known as SFWD. Section IV presents a comparative evaluation of three STSS measures using the SFWD dataset. Finally, section V presents conclusions and future work.

II. Word and Semantic Sentence similarity measures

The first semantic similarity algorithm was called latent similarity analysis (LSA) and was developed by Landauer et al [3]. This similarity measure worked on the principle of determining semantic similarity through looking at relevant statistics for words in a large corpus. The LSA system calculated the level of similarity between two blocks of texts through the use of a vector system. This semantic approach dealt with the issue prevalent in previous similarity measures, that texts could be syntactically very similar but have very different semantic meanings [3]. Subsequent tests of LSA demonstrated it being able to show a high correlation with human ratings in terms of the level of similarity of sentences within a dataset. A problem with the approach taken by LSA however was, that it was more suited towards comparing large texts as opposed to short texts (texts where fewer than 30 words exist). This left a gap in the field for a measure that was able to accurately represent the level of similarity between short pieces of text.

In [4] a new sentence similarity measure called STASIS was developed. This took the work from a previous word similarity measure developed to take relations between words from the WordNet ontology [6] as well as statistical information about the words from a corpus, to calculate semantic similarity [4][5]. In using WordNet, the system calculated the distance between words in the ontology as well as the distance between them and their lowest common subsumer. This system was tested against the original LSA system in [4] and was demonstrated to give a higher correlation with results from a human dataset.

Little research has been done on word or sentence similarity measures that incorporate perception or "fuzzy" based words. In 2013, Carvalhoet et al [16] proposed a word similarity function known as UWS and its fuzzy counterpart, FUWS (partially implemented), which combined the edit distance and n-gram to automatically detect and correct typographical errors in word lists. Preliminary results were presented mainly for UWS and indicated good discrimination capability, which indicated that when FUWS is completed it could be a good candidate for a general fuzzy word similarity measure. Also in 2013, a Fuzzy Algorithm for Similarity Testing (FAST) was proposed [17]. FAST is a novel ontology based similarity measure that uses concepts of computing with words [13]-[15] to allow for the accurate representation of perception based words. The difference between FAST and existing semantic similarity measures is that FAST is able to show the effect that fuzzy words have on the overall level of similarity between short texts. The main components of FAST include a fuzzy ontology, a fuzzy word similarity measure; an algorithm to determine the association of non-fuzzy words with fuzzy words. Initial work involved the creation of a series of fuzzy sets for six categories of words based on their levels of association with particular concepts. All category words were then quantified using a group of human subjects. These values are used to make a fuzzy set for that category word. The union of human ratings, for each word in each category, created a fuzzy set that could then be defuzzified to create a single value to be used that is

representative of that word. The results were used to create new ontological relations between the perception words contained within them. These relationships formed the basis of a new ontology based fuzzy semantic text similarity algorithm that was able to show the effect of perception based words on computing sentence similarity as well as the effect that fuzzy words have on non-fuzzy words within a sentence. The FAST measure will be used as part of the evaluation of the SFWD and will now be explained in further detail.

A. *Creation of a Fuzzy Ontology*

In FAST, it was necessary to create an ontological structure that was able to show the relationships between fuzzy words in a category. The categories of size, temperature, goodness, frequency, age and level of membership were justified in previous work [147 and used to provide distances between words as well as the subsumer depth distances from the lowest common subsumer to the top of the hierarchy. Through the creation of the ontology, a new word similarity measure was built specifically around determining the level of similarity between pairs of fuzzy words. The methodology for the creation of categories, the generation of a set of fuzzy words for each category and the quantification of each of the fuzzy words on scales related to the categories by participants can be found in [17].

To create the fuzzy ontology, each category was first divides into nodes that were related to each other through subsumer relations. This allowed for sets of words from the categories to be stored within these nodes so that relations between these words could be represented by their distances and subsumer depths. Each category was divided into five nodes with the central subsumer being representative of the area around the midpoint of the range. Figure 1 shows the ontology for the size category.

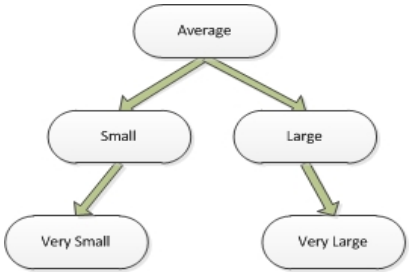

Fig. 1. Size Category Ontology

In order to classify the words in each category, the quantified fuzzy words were re-scaled to reflect them moving away from a central point which represented the top subsumer node. Each word, based on participant ratings [17] were re-scaled on a -1 to 1 scale with the midpoint representing a value of 0. Then through evenly dividing five points along the range, the words were associated with a particular node. An example of the classification of the words in the *size* category is shown in Table I.

TABLE I. CLASSIFICATION OF THE SIZE CATEGORY

Category	**Words in Category**
Very Small	Microscopic, Miniscule, Minute, Tiny, Alongside, Insignificant, Diminutive, Petite, Adjacent
Small	Close, Near, Nearby, Small, Thin, Proximal, Proximate, Little
Average	Regular, Standard, Medium, Normal, Middle, centre, Midpoint, Average
Large	Sizeable, Large, Loads, Thick, Big, Substantial, Distant
Very Large	Massive, Remote, Long, Great, Far, Huge, oversized, Immense, Enormous, mammoth, Giant, Gargantuan, Gigantic

The ontology allows the differences in quantities between the words within a given node category be represented. As each node category covered words that had a range of values, it was essential factor in this range during scaling e.g. “Gargantuan” and “Immense” both belong to the same category (*very large*) but both had different values returned from human ratings. This could show a difference in the level of similarity between words. Therefore, to be able to deal with this issue, each node in itself needed to be re-scaled between {-1..1}, with the word with the middle value, based on participant ratings [17] representing the midpoint. Table II shows an example of rescaling the words in the *very small* category in proportion to the defuzzified participant ratings [17].

TABLE II. SCALE FOR VERY SMALL CATEGORY

Word	Defuzziifed Participant Rating	Re-scaled Value
Microscopic	0.94	-1.00000
Miniscule	1.11	-0.81818
Minute	1.67	-0.27273
Tiny	1.72	-0.27273
Alongside	1.81	-0.18182
Insignificant	1.86	-0.09091
Diminutive	1.94	-0.09091
Petite	2.06	0.090909
Adjacent	2.22	0.181818
Close	2.39	0.363636
Near	2.67	0.636364
Nearby	3.00	0.909091
Small	3.00	0.909091
Proximal	1.00	1.000000
Proximate	1.00	1.000000

B. *Overview Of FAST*

The aim of FAST is to take two sentences containing perception based words as input and return a similarity vector for them. The fundamental building block of FAST is the STASIS measure [4] that in its original form used corpus statistics and syntactic similarity [4] to calculate semantic similarity using nouns within the sentences. Let T_1 and T_2 be two short texts, which the semantic similarity is to be calculated. The FAST algorithm now follows (for a full description see [17]):

For all words ($w_1 \ldots w_n$) in T_1 and *T_2 where n is the total of words in T_1* and *T_2*
Tag every word in T_1 and *T_2*
Pair every combination of tagged words $\{w_1, w_2\}$
For every word pair$\{w_1, w_2\}$:
If $\{w_1, w_2\}$are both fuzzy words:
If $\{w_1, w_2\}$ are in the same category:
Calculate subsumer depth, d from Fuzzy ontology
Calculate path length, l, and the length of the shortest path between $\{w_1, w_2\}$ from the appropriate fuzzy ontology
Calculate word similarity, S between $\{w_1, w_2\}$

$$S\{w_1, w_2\} = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \qquad (1)$$

Where α *and* β, *were empirically determined as 0.2 and 0.6 respectively in [3,4]*

Else:
Apply original STASIS word similarity measure using equation 1, calculating subsumer depth, d and path length, l, from the WordNet ontology [4].
End If
Else
Apply original STASIS word similarity measure using equation 1, calculating subsumer depth, d and path length, l, from the WordNet ontology [4].
Apply fuzzy word association algorithm to determine presence of fuzzy words associated with the non-fuzzy words [14]
If Associated Fuzzy Words are Present:
Calculate new subsumer depth, d and length, l modifications [14].
Recalculate Word Similarity using (1)
Else:
Return level of word similarity for $\{w_1, w_2\}$
End If
Return level of word similarity for $\{w_1, w_2\}$
End If
Calculate Corpus statistics (word frequency information) [4]
Next
Determine Syntactic similarity in terms of word order[4]
Calculate overall semantic similarity $SS(T_1, T_2)$:

$$SS(T_1, T_2) = \delta \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} + (1 - \delta) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \qquad (2)$$

with δ 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 [4].

III. Creating A Single Fuzzy Word Dataset

This section describes the methodology for the creation of a single fuzzy word dataset known as SFWD. The aim was to create a dataset that contained a set of pairs of quantified sentences with a single fuzzy word from the same concept/domain in each of the two sentences. To build this data set there were two different steps that had to be completed to ensure that SFWD was accurate, unbiased and representative of human dialogue.

- A methodology had to be created which generated a set of 30 fuzzy sentence pairs [20]-[26] and then paired them to ensure representation of low, medium and high similarity.
- An experimental methodology was required to return human similarity ratings for the sentence pairs.

It was identified in section I, existing STSS datasets failed to contain a significant number of sentence pairs which contained fuzzy words. However given that recent datasets had been collected through established methods [23][24], using the pairs from an existing benchmark dataset as a basis for the SFWD dataset would ensure that the same level of quality is retained. Using sentences from an existing dataset would require the addition of fuzzy components, which would then need to be re-quantified through human participants. It was also important that these new sentence pairs continued to be representative of natural language while care had to be taken to avoid bias when they were being created. Once the fuzzified sentences had been created, they then had to be paired in such a way to ensure that there was a relatively even distribution of high, medium and low similarity words were returned when the sentence pairs were quantified. Pairing was achieved by a panel of 3 experts in the English language. After pairing the sentences, a methodology to quantify them using human participants was required. It was important that the method to quantify the fuzzy sentence pairs was robust, unbiased and would not lead human participants towards specific answers [23]-[25].

A. Fuzzifying Sentences using Linguistic Experts

The STSS dataset known as STSS-131 [24] was used as the dataset to be fuzzified due to its acceptance as a benchmark dataset [23]. A total of 30 sentence pairs would be required to generate 60 unique sentences which, when paired, gave a complete set. This was achieved using paraphrasing [27] which involved rewriting the sentence whilst changing some of its characteristics. The use of pairs of paraphrased sentences in a sentence similarity dataset can be seen in the large Microsoft Research Paraphrase Corpus [28]. This is a large corpus of pairs of paraphrased sentences with human similarity ratings for each pair. The widely used nature of the corpus [29] evidences the viability of paraphrasing as a method of creating a sentence similarity dataset. The reason that the sentence pairs from the paraphrase corpus could not be used to evaluate FAST is because, as with other datasets, there were very few sentence pairs with fuzzy words in each sentence.

Having established paraphrasing sentences as a means of creating fuzzy sentences, the question then became which method to use to accomplish this task. In papers such as [30], the effect the orientation of fuzzy words could have on a words semantic meaning was discussed. In section II, it was stated

that fuzzy words could be either positively or negatively oriented within the fuzzy ontologies where classes move either positively or negatively from a single central point. For example, the word "Bad" would be considered a negatively oriented word, while the word "Good" would be considered positively oriented. Taking this into consideration, the method that would be applied to the fuzzy sentences was to apply either positively or negatively oriented fuzzy words to either enhance or decrease the impact of a particular aspect of the non-fuzzy sentence. For example, consider the non-fuzzy sentence:

"There is a house".

When asked to add a word to either increase or decrease the size of the house, a positive or negatively oriented word from the size category could accomplish this task. Consider adding the word "huge" (positively oriented to make the house bigger);

"There is a huge house".

The sentence has, through the task of changing the impact of "house", been converted to a fuzzy sentence. Converting a full set of non-fuzzy sentences in that manner generates a set of fuzzy sentences.

With the concept behind fuzzifying sentences having been decided the next issue to decide would be who would actually be responsible for fuzzifying the sentences to prevent bias. Fuzzification of sentences was achieved through the use of human test subjects. Each newly fuzzified sentence had to be semantically and syntactically accurate and representative of natural language. This is because the ability to handle natural language sentences was a critical attribute of FAST and other STSS measures [3][4][8]. As a result of this, some selectivity was required regarding which group of subjects would actually fuzzify the sentences.

In his work, O' Shea [23]-[25] discussed both the importance and usefulness of the use of linguistic experts in the generation of a natural language sentence dataset. He stated that experts, through their in depth knowledge of the English language and sentence construction, could be relied upon to construct natural language sentences. As they are also impartial to the project, the risk of biases within their responses is also reduced. To further reduce the risk of bias, precautions had to be taken to ensure that the instructions that were to be followed were to be constructed in such a manner so as not to unnecessarily lead respondents towards particular answers. Furthermore the instructions also had to clearly illustrate the task at hand. Extensive discussion of how this could be achieved can be found in [24].

For the purpose of creating the SFWD, three English language experts were chosen. They were selected based on them working in professions that involved advanced and extensive knowledge of all aspects of English and its regular practical application. Following the selection of the experts, they were given a set of 30 randomly selected sentence pairs from STSS-131. 20 sentence pairs were selected for high levels of similarity, 5 for medium and 5 for low. This was to ensure a distribution of results across the range of possible similarity levels. Each expert was asked to fuzzify using the method of amplifying or diminishing a particular aspect. For example when given the instruction;

Increase or diminish, if possible, the level of delay for the sentences, T_1 and T_2:

T_1: When I was going out to meet my friends there was a delay at the train station

T_2: The train operator announced to the passengers on the train that there would be a delay.

The returned fuzzified sentences were; -

T_{1f}: When I was going out to meet my friends there was a significant delay at the train station

T_{2f}: The train operator announced to the passengers on the train that there would be a brief delay

Through this method a total of 90 pairs of sentences (180 unique sentences in total) were created. To further reduce the problem of bias, no full sentence pair from a single expert could be added to the dataset. Therefore, for each of the sentence pairs to be generated, two random sentences, each one from a different expert were taken. The final result of this was a set of 30 fuzzy sentence pairs that covered a broad spectrum of levels of similarity. Table III contains the acquired sentence pairs (SP) which formulate the SFWD dataset.

B. Quantification of Sentences in the SFWD

Quantification of sentence pairs in the SFWD required further human experimentation. There had been a number of different methodologies already established for quantifying both word [5][17] and sentence similarity [21]-[24] . As was the case in the construction of all previous sentence similarity datasets, the collection of the similarity data is questionnaire based. 20 participants were selected. A suitable questionnaire was designed which would not lead or bias the respondents' answers. The questionnaire asked participants to rate pairs of sentences based on their level of similarity on a scale of 0 to 10. This scale had been previously used in the Mendel's Codebook [26] that was specifically geared towards fuzzy quantification. There were some common parameters to all previous sentence similarity experiments that aided in addressing this problem [23][24]. They illustrated that examples could be used (just as was the case in the initial collection of sentences), to clearly give participants knowledge of what to do, while at the same time avoiding leading them towards particular answers. This did however mean that careful selection was needed to determine the sentences used. Furthermore, [23] also noted the importance of the positioning of the sentence pairs (i.e. avoiding grouping high similarity sentence pairs together) to further decrease the potential level of bias. Table IV shows the similarity results collected for the SFWD dataset in terms of the average human rating (AHR) and standard deviations (SD-AHR) for each sentence pair (SP).

TABLE III. SENTENCE PAIRS IN SFWD

SP	Sentence 1	Sentence 2
SP1	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.
SP2	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.
SP3	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.
SP4	I will make you laugh so very hard that your sides ache and split.	When I tell you this you will split your sides laughing.
SP5	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.
SP6	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.
SP7	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.
SP8	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.
SP9	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.
SP10	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.
SP11	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.
SP12	Does this soggy sponge look dry to you?	Does pleasant music help you to relax or does it distract you too much?
SP13	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.
SP14	Global warming is what everyone is really worrying about greatly today.	Global warming is what everyone is mildly worrying about today.
SP15	Midday is 12 o'clock in the midpoint of the day.	-Midday is 12 o'clock in the centre of the day.
SP16	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.
SP17	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.
SP18	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.
SP19	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.
SP20	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
SP21	There is no point in trying hard to cover up what you said, we all know.	You shouldn't be burying what you feel.
SP22	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?
SP23	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.
SP24	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.
SP25	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.
SP26	Get that wet dog off my latest sofa.	Get that wet dog off my barely new sofa.
SP27	Will you drink a glass of excellent wine while you eat?	Would you like to drink this wonderful wine with your meal?
SP28	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?
SP29	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.
SP30	I am so hungry I could eat a whole big horse plus desert.	I could have eaten another massive meal, I'm still starving.

TABLE IV. HUMAN SIMILIARITY RATINGS FOR SFWD

SP	AHR	SD-AHR	O'Shea et al [93]	Difference
SP 1	3.83	2.02	7.83	3.85
SP 2	0.00	0.00	0.40	0.40
SP 3	7.30	1.99	7.10	0.34
SP 4	7.95	1.85	9.15	1.15
SP 5	1.28	2.43	0.23	1.19
SP 6	8.72	1.00	9.78	0.98
SP 7	7.10	1.74	8.95	1.90
SP 8	6.72	1.76	9.53	2.57
SP 9	0.95	1.80	1.80	0.75
SP 10	8.25	1.01	7.65	0.52
SP 11	4.96	1.49	8.05	2.99
SP 12	0.53	0.98	0.25	0.28
SP 13	3.29	2.57	3.63	0.47
SP 14	6.37	1.83	7.85	1.28
SP 15	9.14	0.89	9.90	0.85
SP 16	6.78	1.81	9.63	2.60
SP 17	3.23	2.39	8.98	5.56
SP 18	2.11	1.99	2.63	0.35
SP 19	6.76	2.21	9.83	2.83
SP 20	8.99	0.78	9.70	0.72
SP 21	3.55	3.24	5.53	1.60
SP 22	8.85	1.45	9.60	0.76
SP 23	7.04	1.62	8.40	1.35
SP 24	3.83	2.30	5.45	1.37
SP 25	8.86	0.96	9.00	0.11
SP 26	7.58	1.83	8.98	1.33
SP 27	8.92	1.08	8.90	0.06
SP 28	6.91	2.02	9.58	2.51
SP 29	1.30	2.21	1.25	0.18
SP 30	6.62	2.40	9.00	2.36

Following collection of the ratings, it was essential to conduct further experimentation to determine if the inclusion of fuzzy words had an effect on semantic sentence similarity ratings. The aim of the experiment was to see if the use of fuzzy words in a sentence significantly changed its semantic meaning (and therefore changed the level of similarity between the candidate sentence and another sentence). This could be achieved through comparing the sentence pairs from the SFWD with the corresponding sentences from the STSS-131 dataset [24] from which the SFWD sentences were derived. Specifically, the difference could be determined through looking at the levels of variance between the quantities from human ratings of the two sets of data. Given the low level of variance among results when the O'Shea results were collected in [22] and the STSS-131 results were collected in [24] if fuzzy words had no effect on similarity, then there should be a low variance between the SFWD results and their corresponding O'Shea results. The experiment showed that there were a number of cases where a large difference exists between the human participants ratings that were collected for the SFWD dataset and those that had been collected for STSS-131 and reported in [24]. Between the two datasets, there was an average difference of 11.4%, which shows that the fuzzy words do exert an effect on sentence similarity and change the meanings of sentences. Table IV shows for each sentence pair, the ratings obtain in [24] and the difference in those human ratings when collected for SFWD.

IV. Comparison of STSS measures using SFWD

In order to evaluate the SFWD, a series of experiments were conducted against a number of STSS measures. These included the traditional measures LSA and STASIS which were selected due to their wide usage and that they had both been previously benchmarked against a human sentence similarity dataset. FAST was selected (as described in section II) as the fuzzy STSS measure. The aim of the experiment was to test the ability of the measures to represent the similarity between pairs of sentences where each sentence contained a single fuzzy word from the same category.

Each sentence pair in the SFWD was executed in turn to LSA, STASIS and FAST. Each of the sets of results for each measure would have a level of correlation with the human similarity ratings from SFWD. These correlations can be compared against each other to determine the representativeness of the data in terms of human similarity ratings. A higher correlation implies that the measure was more successful in representing human sentence similarity.

Table V shows the comparison of FAST, STASIS and LSA in terms of the SFWD. It contains the average human ratings for each sentence pair, and the similarity ratings for each pair returned by LSA, STASIS and FAST. From the results it can be observed that FAST has an overall Pearson's correlation level of 0.77 with human similarity ratings in the SFWD. STASIS and LSA correlation levels were calculated at 0.71 and 0.64 respectively. This shows that FAST was able to return an improvement of 8.1% over STASIS and an even larger improvement of 20% over LSA. These results demonstrate the success of FAST in terms of its ability to represent sentence similarity in the case of sentence pairs with a single fuzzy component in each sentence. It also demonstrates the strength of ontology based similarity measures in this area over non ontology based ones. This is demonstrated by the fact that STASIS and FAST both showed a large improvement over the performance of LSA.

TABLE V. RESULTS FOR SENTENCE PAIRS WITH 1 FUZZY WORD

SP	Scaled AHR	LSA	STASIS	FAST
SP 1	3.83	0.48	0.75	0.72
SP 2	0.00	0.01	0.47	0.47
SP 3	7.30	0.26	0.67	0.67
SP 4	7.95	0.84	0.75	0.74
SP 5	1.28	0.02	0.56	0.56
SP 6	8.72	0.95	0.63	0.63
SP 7	7.10	0.63	0.85	0.85
SP 8	6.72	0.81	0.78	0.77
SP 9	0.95	0.49	0.62	0.68
SP 10	8.25	0.46	0.71	0.82
SP 11	4.96	0.49	0.41	0.41
SP 12	0.53	0.32	0.49	0.48
SP 13	3.29	0.05	0.57	0.60
SP 14	6.37	0.93	0.92	0.89
SP 15	9.14	1.00	1.00	1.00
SP 16	6.78	0.70	0.84	0.84
SP 17	3.23	0.59	0.32	0.32
SP 18	2.11	0.61	0.50	0.50
SP 19	6.76	0.79	0.78	0.77
SP 20	8.99	0.36	0.82	0.84
SP 21	3.55	0.28	0.54	0.54
SP 22	8.85	0.42	0.88	0.90
SP 23	7.04	0.80	0.86	0.87
SP 24	3.83	0.39	0.71	0.71
SP 25	8.86	0.72	0.74	0.77
SP 26	7.58	0.96	0.87	0.92
SP 27	8.92	0.71	0.71	0.79
SP 28	6.91	0.88	0.86	0.86
SP 29	1.30	0.16	0.38	0.38
SP 30	6.62	0.48	0.53	0.57

V. Conclusion and further Work

This paper has described the methodology for the creation of a SFWD, which can be used to evaluate traditional and fuzzy semantic similarity measures. The method comprised of firstly, the fuzzification of pairs of sentences extracted from the STSS-131 dataset by linguistic experts. Secondly, a methodology was proposed for the quantification of the fuzzified sentences using human participants. Experiments conducted on three STSS measures, showed that fuzzy words play a significant part in computing the semantic meaning between sentences, which was illustrated by FAST giving a higher correlation with human participant ratings. The main conclusions that can be drawn from these experiments is that FAST shows a high level of accuracy in terms of dealing with fuzzy words and a notable improvement over both STSS

measures STASIS and LSA which do not take into consideration perception based words. Current work involves validating a second data set that contains multiple fuzzy words. Given the complexity of such sentences that would be required, a new automated approach has been developed which involves extraction of sentences with fuzzy components from a corpus, fuzzifying them and then pairing them to formulate a multiple fuzzy word dataset. Once validated, the dataset will form a richer set of natural language sentences containing perception-based words that could be used to evaluate both traditional and fuzzy semantic similarity measures.

REFERENCES

[1] Joachims, T. Text categorization with support vector Machines: Learning with many relevant features. Springer Berlin Heidelberg, 1998.

[2] Salton, G, Buckle, C. Term-weighting approaches in automatic text retrieval", Information processing & management vol.24, no. 5, 1988, pp.513-523.

[3] Landauer, T., Foltz,P. Laham,D. An introduction to latent semantic analysis" Discourse processes vol. 25, no 3, 1998, pp.259-284.

[4] 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, 2006, pp.1138-1150.

[5] 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, 2003, pp.871-882.

[6] Fellbaum, C. WordNet. Springer Netherlands, 2010.

[7] Agirre, E, A study on similarity and relatedness using distributional and WordNet-based approaches. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, 2009, pp.19–27.

[8] Islam, A, Inkpen, D. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data Vol.2.2:10, 2008.

[9] Tsatsaronis, G. Varlamis, I. Vazirgiannis, M. Nørvåg, L. Omiotis: A thesaurus-based measure of text relatedness. Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, Vol.5782, 2009, pp.742-745.

[10] Rus, V., Lintean, M., Banjade, R., Niraula, N., and Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, August 4-9, 2013, Sofia, Bulgaria

[11] Zadeh, L. From Computing with Numbers to Computing with Words—from Manipulation of Measurements to Manipulation of Perceptions. Logic, Thought and Action, International. Journal. Applied Math. Comput. Sci., Vol.12:3, 2002, pp. 307–324.

[12] Glockner, I. Fundamentals of Fuzzy Quantification:Plausible Models, Constructive Principles,and Efficient Implementation, Report TR2002-07, University at Bielefeld, Available: http://pi7.fernuni-hagen.de/gloeckner/tr0207.pdf, Date Accessed: 14/3/14.

[13] Zadeh, L. The concept of a linguistic variable and its application to approximate reasoning - I, Information Sciences, vol. 8, no. 3, pp. 199-249, July 1975.

[14] Mendel, J. Computing with words and its relationships with fuzzistics, Information Sciences vol. 177:4,2007, pp.988-1006.

[15] Wu, D., Mendel, J., Coupland, S. Enhanced Interval Approach for encoding words into interval type-2 fuzzy sets and its convergence analysis, IEEE Transactions on Fuzzy Systems, vol. 20:3 2012, pp.499-513.

[16] Carvalho, J.P.; Coheur, L., Introducing UWS - A fuzzy based word similarity function with good discrimination capability: Preliminary results, 2013 IEEE International Conference on Fuzzy Systems, 2013 doi: 10.1109/FUZZ-IEEE.2013.6622494.

[17] Chandran, D.; Crockett, K.; Mclean, D.; Bandar, Z., FAST: A fuzzy semantic sentence similarity measure, 2013 IEEE International Conference on Fuzzy Systems , 2013, doi: 10.1109/FUZZ-IEEE.2013.6622344

[18] Miller, G, Walter, G., Contextual correlates of semantic similarity, Language and cognitive processes, Vol.6:1, 1991, pp.1-28.

[19] Miller, G, Word Net: a lexical database for English, Communication", ACM Vol. 38:11, 1995, pp.39-41.

[20] Rubenstein, H, Goodenough, J. Contextual correlates of synonymy, Communications of the ACM Vol. 8:10, 1965, pp.627-633.

[21] O'Shea, J, Bandar, Z. Crockett, K, Mclean, D., Benchmarking short text semantic similarity, International Journal of Intelligent Information and Database Systems, Vol. 4:2, 2010, pp.103-120.

[22] O'Shea, J, Bandar, Z. Crockett, K, Mclean, D., A comparative study of two short text semantic similarity measures. Agent and Multi-Agent Systems: Technologies and Applications, 2008, pp.172-181.

[23] O'Shea, J, Bandar, Z. Crockett, K, Mclean, D., Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description, Technical Report Available: http://www2.docm.mmu.ac.uk/STAFF/J.Oshea/TRMMUCCA20081_5.pdf. Date accessed: 12/12/13.

[24] O'Shea, J., Bandar, Z., and Crockett, K. A new benchmark dataset with production methodology for short text semantic similarity algorithms. ACM Trans. Speech Lang. Process. 10, 4, Article 19, December 2013, 57 pages, DOI: http://dx.doi.org/10.1145/2537046

[25] O'Shea, J., A Framework for Applying Short Text Semantic Similarity in Goal-Oriented Conversational Agents, PhD Thesis. School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University: Manchester, 2010: Available: http://semanticsimilarity.net/?attachment_id=138

[26] Liu, F, Mendel, J. Encoding words into interval type-2 fuzzy sets using an interval approach. IEEE Transactions on Fuzzy Systems Vol. 16.6, pp. 1503-1521.

[27] Dolan, B., Quirk, C. Brockett, C. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources, 20th Int. Conf. on Computational Linguistics, 2004, pp. 350–356.

[28] Dolan, B., Brockett. C. Automatically constructing a corpus of sentential paraphrases. Dolan, B., & Dagan, I. (Eds.). Proc. of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment. Ann Arbor, MI, 2005.

[29] Das, D., Smith. N, Paraphrase identification as probabilistic quasi-synchronous recognition. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Vol.1:1, Association for Computational Linguistics, 2009, pp. 468-476.

[30] Pang, B., Lee. L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2 Vol.1:2, 2008, pp.1-135.