An Automatic Corpus Based Method for a Building Multiple Fuzzy Word Dataset

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

Abstract—**Fuzzy sentence semantic similarity measures are designed to be applied to real world problems where a computer system is required to assess the similarity between human natural language and words or prototype sentences stored within a knowledge base. Such measures are often developed for a specific corpus/domain where a limited set of words and sentences are evaluated. As new "fuzzy" measures are developed the research challenge is on how to evaluate them. Traditional approaches have involved rigorous and complex human involvement in compiling benchmark datasets and obtaining human similarity measures. Existing datasets often contain limited fuzzy words and do allow the fuzzy measures to be exhaustively tested. This paper presents an automatic method for the generation of a Multiple Fuzzy Word Dataset (MFWD) from a corpus. A Fuzzy Sentence Pairing Algorithm is used to extract and augment high, medium and low similarity sentence pairs with multiple fuzzy words. Human ratings are collected through crowdsourcing and the MFWD is evaluated using both fuzzy and traditional sentence similarity measures. The results indicated that fuzzy measures returned a higher correlation with human ratings compared with traditional measures.**

I. Introduction

Sentence similarity is the process by which algorithms determine how alike sets of short texts (typically 10 – 25 words in length) are towards each other [1] through returning a similarity vector between them. On evaluation of early sentence similarity measures (SSM) it was clear that short texts could be syntactically similar but convey different semantic meanings [2]. Today, most SSM measures stem from Latent Semantic Analysis (LSA) developed by Laudener et al. [2] or STASIS developed by Li et al [1]. Whilst both methods were based or a corpus statistics approach, STASIS also combined a word semantic similarity measure, word order and an ontology system (WordNet [3]) to calculate the similarity value.

The common factor of LSA, STASIS and their derivatives, were that they were unable to determine the level of similarity between words with subjective meanings that are based on human perception such as "big" or "good" as such words needed to be measured in the context that they were applied. Words with subjective meanings can be referred to as fuzzy words and are typically used in everyday human natural language dialogue and are often ambiguous and vague in meaning [4]. However, this limitation did not stop traditional algorithms such as STASIS [1] being successfully applied in real world applications such as conversational agents [5], determining literal and intelligent plagiarism [6], to predicting activity attendance in event-based social networks [7].

Early work on incorporating fuzzy set theory to word and similarity measures stems back from 1991, when Ogawa et al. [8] proposed a fuzzy document retrieval system using the keyword connection matrix and a learning method. This work inspired other researchers such as Yerra et al. [9] to propose a fuzzy set information retrieval approach for sentence based copy detection on Web documents. The approach successfully determined the existence of overlapped portions of any two Web documents and was evaluated on a series of randomly sampled sentence pairs. Alzahrani and Salim [10] combined work on constructing a fuzzy similarity model by Yerra and Ng [9] and STASIS to fuzzy semantic-based string similarity for extrinsic plagiarism detection. Due to a variety of reasons reported by the authors the model only achieved a precision of 57% in correctly detecting plagiarism. There is little discussion on the training and testing corpus that was obtained and used for the purpose of evaluation.

Carvalhoet et al. [11] proposed a fuzzy word similarity function (FUWS) to detect and correct typographical errors in word lists. Results on a limited domain indicated that FUWS could be adapted as a general fuzzy word similarity measure however further training and testing would be required. As part of a Twitter Topic Fuzzy Fingerprint algorithm for the detection of topics, Rosa et al. [12] developed a Tweet-Topic Similarity Score which measures how much a tweet fits to a given topic. Whilst the results were promising in that the algorithm outperformed traditional classifiers such as SVM, the authors acknowledged that only a comparatively small data set had been used and recognized in future work that the test sets need to be extended and manually annotated for further result validation.

From this brief review, it is generally acknowledged that in general, many similarity measures have been proposed and evaluated on limited sized datasets, selected for specific problem domains which often involve human participants providing ratings using a variety of different methods. In terms of a recent SSM, Chandran et al. [13] proposed FAST (Fuzzy Algorithm for Similarity Testing) – an novel ontology-based similarity measure which was based on STASIS. This algorithm was evaluated on a Single Fuzzy Word Dataset which was developed using methodologies adapted from the field of traditional evaluation of SSM measures. This allowed FAST to be performance benchmarked and compared with crisp SSMs in terms of correlations with human similarity ratings. The work highlighted measuring the similarity of fuzzy words in relationship to other fuzzy words in a sentence, allowed similarity ratings obtained by FAST to be much closer to human ratings. However, the dataset only looked at the impact on similarity of one fuzzy word in each sentence. The question then arises, how do the inclusion of multiple fuzzy words in a sentence affect the similarity measurement?

This paper proposes a methodology for the automatic creation of a multiple fuzzy word dataset (MFWD) that is automatically generated from a corpus. For the purpose of this study the fuzzy words are selected from 6 fuzzy categories that have been quantified by human participants. The core of this method is the fuzzy sentence pairing algorithm which selects pairs of high, medium and low sentence pairs which are representative of natural language and uses the categories to "fuzzify" these sentence pairs. Quantification of sentence pairs is achieved using selected human participants and through the use of crowd sourcing. The dataset produced is evaluated on three SSM's including FAST and the results show that the increasing the presence of fuzzy words in sentences affects the ability of traditional SSMs to achieve good correlations with humans similarity ratings. In addition, the creation of MFWD validates the proposed methodology which allows for the automatic creation of larger datasets without the need to perform expensive human evaluation.

This paper is organized as follows; Section II provides an overview of related work on methodologies for creating evaluation datasets for SSM's and methods for the quantification of fuzzy/perception based words. Section III describes the formation of fuzzy categories and associated human quantification as used in this work. Section IV describes the method for the automatic generation of a multiple fuzzy word dataset using the fuzzy sentence pairing algorithm and their quantification using crowdsourcing. The new MFDS is also presented. Section V presents the evaluation of MFWD using three SSMs and associated discussion. Finally, section VI presents the overall conclusion and future work.

II. Related Work

A. *Creating Evaluation Datasets for Word and Short Text Measures*

Research into the creation of methodologies to evaluate both word and short text similarity measures is well established and driven by the need to perform comparative evaluations of new and old measures across a benchmark. Creation of such datasets is however challenging in that humans are required at all stages including the collection of words, similarity rating of words, construction of sentences/short texts and the rating of these sentences. O'Shea et al. [14] stated that "Semantic similarity is an artifact of human perception" which means the evaluation of SSM is inherently empirical and ultimately relies on the creation of benchmark datasets comprising of human ratings.

In 1965, Rubenstein and Goodenough [14] presented a methodology and used it to create a dataset comprising of 65 sets of word pairs from which human similarity ratings were collected. This was the first methodology to acquire numeric values for the words from human test subjects. This methodology involved a group of undergraduate students comparing a set of words on a scale of 1 to 4. These experiments showed a sufficiently low level of deviation between the results for them to provide a framework for the numbers of words, participants and the types of scales that were used in this experiment. In 2012, as part of a SEMEVAL 2012 Task 6 [15], a dataset known as (S2012-T6) was produced for training, testing and evaluating semantic similarity algorithms, but the texts were extracted from existing corpora and did not include dialog. An important point to note about all the existing datasets is that the selection of the words used within them is arbitrary. There has been no system of using human respondents to generate the words that were paired. In the work presented in this paper, human participants were used to select fuzzy words within 6 fuzzy categories to ensure that the words were representative of natural language.

More recently, O'Shea proposed two short text benchmark sentence similarity datasets known as STSS-65 [16] and STSS-131 [14]. In [14] O'Shea proposed a robust methodology to create benchmark word and sentence datasets that could be used to evaluate human participant ratings in an unbiased manner. STSS-131 was used to evaluate both STASIS, LSA and more recently FAST [13]. FAST performed similar to STASIS which was expected as STSS-131 did not contain a significant coverage of fuzzy words. Consequently, it was not possible to evaluate aspects of FAST such as its ontological structure or the effects of fuzzy words on the semantic meaning of a sentence.

To address this issue, in [17], the first benchmark data set containing one fuzzy word per short text was proposed. The methodology for the creation of the Single Fuzzy Word Dataset (SWFD) involved the fuzzification of pairs of sentences from STSS-131 by human linguistic experts and a series of quantification experiments using native English speaking human participants. Performing evaluations of SSM measures using SWFD revealed that fuzzy words played a significant part in computing the semantic similarity of sentences, and that considering the similarity of fuzzy words in the semantic context of each sentence gave a higher correlation with human participant ratings than traditional SSMs. However the use of SWFD [17] did not look at the effect of the similarity measurement when the number of number of fuzzy words in a sentence was increased. A further question raised was whether a fuzzy based SSM such as FAST could maintain an improvement over existing SSM with an increased number of fuzzy words. Thus there is a requirement to devise a Multiple Fuzzy Word Dataset (MFWD) which can be used to provide a more extensive evaluation of SSMs.

B. *Quantification of Fuzzy Words*

Methods for the Quantification of fuzzy words stems from Zadeh's work on the concept of granularity [18]. Mendel went on to show explicitly that Zadeh's work on granularity and Computing with Words (CWW) could be used to generate quantities to represent words on a given scale [19]. Pioneering work was conducted to develop a methodology to create a codebook [19] to determine the Footprint of Uncertainty (FOU) of 32 fuzzy Type-2 sets each based on a fuzzy word. The FOU of a type-2 fuzzy set was defined as the union of all primary memberships of the set. The methodology adopted an interval approach to determine these FOUs. All of the 32 words related to the concept of size with the fuzzy sets containing ranges of quantities covered by these words on a scale related to size. These quantities were determined through an experiment where a group of 28 participants were asked what the interval end points on a 0-10 scale were for the words in relation to size. After the FOUs had been determined a series of centroids for the Type-2 fuzzy sets of each word and a mean value for each of them was returned. It was observed that there was a significant amount of overlaps between many of the FOUs. However, each of the different words had a unique mean value. It was noted in [20] that, the word FOU's were generally to fat and wide. Due to limitations, expansion on this work by Wu et al [20, 21] led to the development of an enhanced internal approach to construct word models from intervals collected from human survey participants.

In [22] Mendel et al. proposed a methodology for determining a words interval type-2 fuzzy set model using one participant. Data collection involved asking each participant two questions relating to the perceived intervals of endpoints for a given word on a scale. The results showed that for 10 probability words, that a single participant could generate a robust FOU. Further work in terms of expanding the categories of words could consider this method to quantify words to expand the categories allowing for greater coverage of fuzzy words in the sentence datasets and also to enhance the ontology of FAST [13].

III. Formation of Fuzzy word categories

This section overviews the methodology used to create a set of fuzzy word categories, then populating those categories with fuzzy words and then quantifying the fuzzy words against each other based on their level of association within a particular category. This results in a set of fuzzy words with quantities on a given scale, thus demonstrating the differences between them. At each stage an independent set of human participants were used. This provides a framework from which fuzzy words can be integrated into a SSM such as FAST [13].

A. *Creating Categories of Fuzzy Words*

The requirement for category creation was to hold a large range of fuzzy words that cover a series of different concepts. Furthermore it was important that the category permitted related fuzzy words to each be scaled in terms of their level of association with the category. The creation method was inspired by [19]. For the purpose of this work, the set of fuzzy categories, *C* is defined as *C* = {Size, Temperature, Goodness, Frequency, Age, Level of Membership}. When Zadeh first described CWW in [23], he described the three categories (size, distance and age) as granules and so it was decided that these categories would be used. Size and distance were then merged into a single one due to the large level of overlap between them in terms of the potential fuzzy words that could be included in either category. This was established using a scoping experiment where a set of 20 humans were asked to list words they thought belonged in each category. The four other categories were selected due to the large number of frequently used fuzzy words contained within them.

B. *Populating categories*

Once the categories had been determined, the next phase was the population of the categories with fuzzy words. The words that were collected were representative of natural language dialogue and commonly used by English speakers. If words were arbitrarily chosen there exists a risk of selective bias in terms of the person who determines the words which then in turn increases the risk of corrupting the value of quantities returned for the words. Furthermore an individual might have particular words that they use that are not widely used or have very commonly used words that they do not consider. The problem in CWW of differing perceptions between individuals was explored in detail by Mendel in a number of papers [20, 21]. Therefore to populate categories, the opinions of a wide range of people are needed to be taken into consideration.

The method proposed by O'Shea et al. [14] to generate benchmark short text datasets for evaluating SSM's was adapted to generate a list of words for each category. To ensure that there was a wide range of words with different values across the categories, a series of guide words (words that could act as stimuli) were used across each category. For example, with the size/distance category, the guide words were 'very small', 'small', 'average', 'large' and 'very large'. When considering which guide words would be used, it was important that the guide words were not selected in such a way so as to bias the results and they would serve their intended purpose and not mislead participants. Twenty English speakers were asked to complete a questionnaire which for each category, asked them to take each guide word and state all the words that they felt had similar meanings. I.e. for the guide word *Cold*, the participant may have written *Cool* underneath. The participants were asked only to include only single words and dual words with a hyphen (such as middle-aged) but not sets of words (such as "As good as it gets").

Through taking the words that had been collected and then collecting a set of synonyms for them, statistics could be collected from the Brown Corpus [24] to determine the usage of these words in natural language. The Brown Corpus was selected due to being a large corpus that contains numerous English language texts from a very wide variety of sources.

This includes a large number of sources where the text is representative of human conversation. Looking at the presence of the collected category words in the Brown Corpus it was determined that they represented 1.6% of all words within the corpus. Then looking at the presence of the words within sentences within the corpus it was determined that at least one of the words was present in 24% of all the sentences in the corpus. This shows the influence even a very limited number of words has and is a strong indication of the significance of fuzzy words in terms of sentence similarity.

C. Fuzzy Word Quantification

The concept of defuzzifying a fuzzy set formulated from a set of different people’s perceptions around a fuzzy word forms the basis of the experiment to quantify them. To perform quantification of the fuzzy words in each category, a further set of twenty human participants were asked to provide a single value that is representative of the point where the membership function of that fuzzy word would be highest. The standard deviation of these points reflects the level of uncertainty. To acquire values for the fuzzy words in all categories, a questionnaire was created that asked participants to rate each word in each category on a scale of 0 to 10 based on which value they felt best represented a numerical value for the word. The union of these results, per fuzzy word, was used to create a fuzzy set. To defuzzify the results, the mean average of each of the sets was used. It was observed that in a vast majority of fuzzy words, the standard deviation was less than 2.00. Table I shows the defuzzified values for the size category and the associated standard deviation.

IV. AUTOMATIC GENERATION OF A MULTIPLE FUZZY WORD DATASET

A. Overview

This section describes a methodology for the generation of a Multiple Fuzzy Word Dataset (MFWD). For the purposes of expediency developing an automated method was considered. The aim of this dataset is to determine if the number of fuzzy words in a sentence affects the ability for a SSM measure to correlate more closely to human ratings for a given set of sentence pairs. To overcome the challenges of creating a dataset using human participants, the methodology involves the use of a new sentence pairing algorithm which is used to select high, medium and low similarity sentence pairs automatically from a corpus. The constraint on this selection is that each set of sentence pairs should contain at least two fuzzy words per sentence per pair.

B. A Corpus Based Method of Building a Fuzzy Dataset

There has been substantial work that has been done in terms of extracting semantic information from corpuses [1] [25]. A problem with automatic generation of sentences is that they may not be as representative of natural language as sentences that were created using a human expert method. However, an automated method would be more efficient and could offer much more control over the number of results that are returned. Furthermore, given that many of the texts from within a corpus are based on natural language [24], using them even after further fuzzification is not likely to significantly reduce their naturalness.

TABLE I QUANTIFICATION OF FUZZY WORDS IN THE SIZE CATEGORY

Word	**Defuzzified Value**	**Standard Deviation**
Adjacent	2.22	1.52
Alongside	1.78	1.31
Average	4.89	1.08
Big	7.22	0.94
Close	2.39	1.85
Diminutive	1.94	2.22
Distant	7.89	1.53
Enormous	8.78	1.63
Far	8.28	1.07
Gargantuan	9.00	2.41
Giant	8.94	1.95
Gigantic	9.11	1.97
Great	8.22	1.56
Huge	8.39	1.65
Insignificant	1.86	1.66
Large	7.17	1.86
Little	3.17	1.86
Massive	8.11	1.32
Medium	4.67	1.37
Microscopic	0.94	1.21
Middle	4.72	1.02
Miniscule	1.11	0.90
Minute	1.67	1.19
Near	2.67	1.53
Nearby	3.00	1.08
Normal	4.67	0.69
Petite	2.06	0.94
Proximal	3.11	1.53
Proximate	3.11	1.45
Regular	4.44	0.92
Remote	8.11	1.75
Sizeable	7.11	1.97
Small	3.00	1.03
Standard	4.56	0.86
Substantial	7.33	1.57
Tiny	1.72	0.89

C. Selecting a Corpus

The Gutenberg Project corpus was selected [26] for sentence extraction as it contained a wide variety of texts from a number of different sources. It has been used extensively in a number of different Natural Language Processing projects [27] and as a result it has had its effectiveness in the field proven. The multitude of texts that are found within it allow for sentences from it to be a fairer representation of the English language than using a corpus that is more focused on a single source would be. This is because the range of sources would cover variations in language that occur when it is used in different circumstances.

D. The Sentence Pairing Algorithm

The Sentence Pairing Algorithm takes as input the maximum length of a sentence in words (*Ln*) the total number of sentence pairs to be generated (*SP*), the total number of fuzzy words per sentence (*Fz*) the number of sentence pairs of high similarity that need to be returned (*H*), the number of sentence pairs of medium similarity to be returned (*M*) and the number of sentences of low similarity to be returned (*L*). Though the initial steps remain constant three different sub-algorithms are used to generate the high, medium and low similarity sets of sentence pairs. For the purpose of using the algorithm to build the required set, the following parameters were set. The maximum length of a sentence (*Ln*) = 30, the number of fuzzy words per sentence (*Fz*) = 2, the number of sentence pairs (*SP*) = 30, the number of high similarity pairs (*H*) = 20, the number of medium similarity pairs (*M*) = 5, the number of low similarity pairs (*L*) = 5. For the purpose of this work, the categories, C are selected from *C* = {Size, Temperature, Goodness, Frequency, Age, Level of Membership} as discussed is section III. The sentence length of 30 was selected as that was considered to be the maximum length a set of text for a sentence [1].

Sentence Pairing Algorithm

```
1)  Let $T$ = set of sentences $\{ S_1, S_2 \ldots\ldots..Si\}$ in the
    Gutenberg Corpus where $Si \in \{ w_1 w_2 \ldots\ldots..wk\}$ where
    $k$ is the number of words in sentence $Si$.
2)  Let $F$ = list of all fuzzy words $\{ fw_1, fw_2 \ldots\ldots..fwn\}$ in
    fuzzy category $Cx$ where x = 6, defined in section III
    and where $n$ is the total number of fuzzy words.
3)  Let $Fp$ = List of all positively oriented fuzzy words
    $\{ fp_1, fp_2 \ldots\ldots..fpn\}$ in all fuzzy categories, $C$.
4)  Let $Fn$ = List of all negatively oriented fuzzy words
    $\{ fn_1, fn_2 \ldots\ldots..fnn\}$ in all fuzzy categories, $C$.
5)  Tag each Sentence ($Si$) in $T$
6)  For each $Si$ in $T$ :
            If Length of $Si$ < $\boldsymbol{Ln}$ then
              If $\{ fw_1, fw_2 \ldots\ldots..fwn\} \in Si$ and count of
              $\{ fw_1, fw_2 \ldots\ldots..fwn\} \in Si$ = $Fz$ then
                  Add $Si$ to list $Sf$
7)  Apply High Similarity Algorithm or
8)  Apply Medium Similarity Algorithm or
9)  Apply Low Similarity Algorithm
```

High Similarity Algorithm

```
7.1  Select $SP$ random sentences in $Sf$ as $Sr$
7.2  For each $Si$ in $SP$
     7.2.1   Clone $Si$ as $S1$
     7.2.2   For each word $wk$ in $S1$
     7.2.3   If $wk$ in $Fp$ then replace $wk$ with Random word
             $fpn$ in $Fp$ where $wk$ and $fpn$ in $Cx$ Else If $wk$ in
             $Fn$ then replace $wk$ with Random word $fnn$
             in $Fn$ where $wk$ and $fnn$ in $Cx$
7.3  Add $Si$ and $S1$ as pair to $TSet$
7.4  Return $TSet$
```

Medium Similarity Algorithm

```
8.1  Select $M$ random sentences in $Sf$ as $Sr$
8.2  For each $Si$ in $SP$ where $Si \nexists$ TSet
     8.2.1   Clone $Si$ as $S1$
     8.2.2   For each word $wk$ in $S1$
     8.2.3   If $wk$ in $Fp$ then replace $wk$ with Random word
             $fnn$ in $Fn$ where $wk$ and $fnn$ in $Cx$
             Else If $wk$ in $Fn$ then replace W with Random
             word $fpn$ in $Fp$ where $wk$ and $fpn$ in $Cx$
8.3  Add $Si$ and $S1$ as pair to $TSet$
8.4  Return $TSet$
```

Low Similarity Algorithm

```
9.1  Select $L * 2$ Random Sentences, $Si$ and $Sj$ in $Sf$ as $Sr$
     where $Si \nexists TSet$
9.2  Randomly pair all sentences $Si$ and $Sj$ in $Sr$
9.3  Add, $Si$ and $Sj$ as a pair to $TSet$
```

Overview of the Sentence Pairing algorithm

The algorithm first specifies all the sentences in the Gutenberg corpus as a single set, *T*. Once the list has been collected the sentences can be dealt with and parsed as individual entities. The list of all the fuzzy words, *F* in all fuzzy categories, *C* is collated and referenced to determine the presence of fuzzy words in any of the sentences. Within each category, words can either be positively or negatively orientated from the central subsumer of that categories fuzzy ontological structure [13]. Positively or negatively oriented fuzzy words are used to either enhance or decrease the impact of a particular aspect of a sentence. Step 5) of the algorithm, tokenizes each of the sentences so each sentence is represented as a list of words where each word can now be referenced and used as individual entities. This also allows for words in sentences to be easily replaced with other words. Step 6) involves generating a list of all fuzzy sentences where there are two fuzzy words in each of the sentences from *T*. For all sentences in *T*, the length of the sentence in words is computed and the algorithm determines if it can be classified as a sentence. If this is the case the algorithm then looks at all the tagged words in the sentence. Through comparing each of the words in the sentence with the list of fuzzy words contained in *F*, the algorithm determines the presence of fuzzy words in the sentences. The algorithm is specifically looking for sentences that contain a number of fuzzy words equal to *Fz*. If the sentence does have the correct number of fuzzy words, it is then added to another list of sentences *Sf*. *Sr* is defined as a set of random sentences from the corpus, that fit the required criteria for length and number of fuzzy words and have not already been added to the dataset from the high and medium similarity algorithms. It has a size of twice the number of low similarity pairs required (e.g. if L=5, *Sr* would contain 10 sentences). The sentences within *Sf* and *Sr* are used for the purpose of generating sentence pairs. Steps 5) to 7) apply either the high, medium or low similarity algorithms to obtain the correct portion of sentence types in the MFDS.

High similarity Algorithm

First, all the positively oriented fuzzy words (words that, on the scale that they were quantified on, have a value greater than 0) are stored in a list (*Fp*). Within this list they are furthermore classified into sub-lists based on their domain (e.g. size words are classified into a sub-list, temperature based words are classified into a sub-list, etc.).The classification of the words into sub-lists is to allow them to easily be replaced by other words within the list. A similar procedure is then applied to all the negatively oriented words. Generation of sentence pairs is achieved through replacing fuzzy words in the sentences with other fuzzy words from within the same domain thus creating two different sentences that can be compared. The first step of this procedure is the selection of a random sentence from the set *Sf*. The reason for random selection is to ensure that all the different texts from within the corpus are given a chance to be represented, preventing the risk of bias. Following the selection of the sentence, the fuzzy words within are then identified. They are then replaced with random fuzzy words from the same orientation. At this point the two sentences are added as a pair to the list *TSet*. This process is repeated to generate a number of sentence pairs equal to the *H* value.

Medium similarity Algorithm

Firstly, before any sentences are selected, the algorithm checks to ensure that instances of the sentence do not already exist in the *TSet* list. This is to prevent repetition. For each selected sentence, as with the high similarity algorithm, it is cloned and its fuzzy words are replaced. The difference however is that while in the high similarity algorithm the fuzzy words were replaced with others from the same orientation, in this case they are replaced by words from the opposite orientation. This is done until a number of sentence pairs equal to the *M* value are generated. The sentence pairs that are generated this way are added to the *TSet* list.

Low Similarity Algorithm

A set of random sentences that are not already in *TSet* is selected from *Sf*. The number of sentences is equal to the *L* value multiplied by two. All the sentences in *Sr* are now randomly paired with each other. Given the vast range of different sentences that are present in the corpus, this makes it highly improbable that the sentences will be related to each other. These unrelated sentence pairs are therefore likely to have very low similarity ratings, ensuring that the low range of the spectrum is covered. The sentence pairs that have been generated using this method are added to the *TSet* list.

Table IV shows the complete list of sentence similarity pairs with two fuzzy words that was generated.

TABLE I. MFWD SENTENCE PAIRS

SP	Sentence 1	Sentence 2
SP1	How marvelous middling Piccola must have been	How good poor Piccola must have been
SP2	A frosty youthful man	A hot old man
SP3	Had you married you must have been regularly acceptable	Had you married you must have been always poor
SP4	The little village of Resina is also situated near the spot	He seems an excellent man and I think him uncommonly pleasing
SP5	They hint that all whales on-occasion smell amazing	They hint that all whales always smell bad
SP6	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
SP7	Mr Brown broke into a mostly antiquated giggle	Mr Brown broke into a rather childish giggle
SP8	An unacceptable watcher and very dietetically pathetic is Dr Bunger	A great watcher and very dietetically severe is Dr Bunger
SP9	Have massive mercy on the mediocre men	Have a little mercy on the poor men
SP10	Behold how fine a matter an adjacent fire kindleth	Behold how great a matter a little fire kindleth
SP11	A little quickness of voice there is which rather hurts the ear	The only living thing near was an old bony grey donkey
SP12	And he laughed almost dreadfully	And he laughed rather unpleasantly
SP13	That is somewhat the acceptable complication	That is just the awful complication
SP14	But why the fantastic youthful playthings	But why the nice new playthings
SP15	The advantages of Bath to the child are pretty sufficiently understood	The advantages of Bath to the young are pretty generally understood
SP16	A thick Juvenile man	A little old man
SP17	He seems a great decrepit party, " I remarked	He seems a pleasant old party," I remarked
SP18	It is as long again as almost all we have had before	was scarcely less warm than hers and whose mind -- Oh
SP19	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
SP20	It is largely a sizeable story, said Turnbull smiling	It is rather a long story," said Turnbull smiling
SP21	Do not treat the little Stars so," said the good Moon	Mrs Price s last baking failed for want of good barm
SP22	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
SP23	She constantly travels with her own sheets an excellent precaution	She always travels with her own sheets an excellent precaution
SP24	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
SP25	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.

SP26	Some people were habitually cross when they were temperate	Some people were always cross when they were hot
SP27	But Mr Weston is just a recent man	But Mr Weston is almost an old man
SP28	If indeed it could be restored to our poor little boy --"	Almost sobbed the young man who was in the highest spirits
SP29	So would useless diminutive Harriet	So would poor little Harriet
SP30	What's the fine pensionable man	What's the good old man

E. Quantifying the MFWD of Sentence Pairs through Crowdsourcing

Given the increased number of fuzzy words per sentence, there was a risk that the variance would increase in terms of human similarity ratings. Therefore a larger number of human responses would be required. It was recognized that the traditional method of quantification using questionnaires to acquire ratings was time consuming and therefore an alternative approach was required. A method that had been used in a number of areas for collecting data from human participants was crowdsourcing [28]. Crowdsourcing refers to, in this particular instance, collecting information from a group of people who volunteer to participate through a common interface for a small monetary reward.

One major tool for crowdsourcing was the Crowdflower system [28]. This allows for users to complete a survey for a monetary reward that is specified by the survey's creator. It also allows a designer to set criteria to determine the people who are surveyed. Furthermore, it allows for the creation of "Gold Standard" questions. These are questions where there are expected answers by the users, allowing for easy determination of whether the participant was following the survey's instructions. It was decided that to create a dataset of human similarity for the MFWD, two sources would be used. The collection of results would be divided between a small number of direct surveys to human participants and collecting a larger amount of data through a crowdsourcing system. This would also allow for the testing of whether or not there was any noticeable difference between results from direct surveys and crowdsourced ones. The survey was created using the same methodology that was used to create the SFWD [17] with the use of a 0 to 10 scale and examples to clarify instructions to the users. A total of 36 responses were collected from all participants (22 were from crowdsourced participants). The average ratings (*AHR*) for each sentence pair in the MFWD are shown in Table II, along with the Human Standard deviation (*Human SD*). A t-test on the results returned a *P* value of 0.96, very strongly suggesting that there is no significant difference between Non-Crowdsourced and Crowdsourced result. What this illustrates is the similarity of the two sets of standard deviations from the crowdsourced and non-crowdsourced results are not significantly different. This therefore opens a new avenue in terms of data collection for any future work.

V. Evaluation of MFWD of Sentence Pairs

A series of experiments were devised to evaluate the MFWD through the application of a series of SSMs. The aim of the experiments was to test the ability of the SSMs to represent the similarity between sentences pairs of high, medium and low similarity where each sentence contained two fuzzy words. The experimental methodology consisted of each sentence pair being run through traditional SSM's LSA and STASIS and the fuzzy SSM FAST. Each measure would give a level of correlation with the human similarity ratings from MFWD. A higher correlation with human similarity ratings implies that the measure was more successful in representing human sentence similarity.

TABLE II. RESULTS FOR MFWD SENTENCE PAIRS

SP	AHR	Human SD	LSA	STASIS	FAST
SP 1	5.62	2.94	0.66	0.87	0.90
SP 2	1.72	2.06	0.72	0.40	0.59
SP 3	3.78	2.27	0.82	0.73	0.94
SP 4	0.75	1.62	-0.01	0.24	0.21
SP 5	3.71	2.75	0.84	0.89	0.90
SP 6	8.35	1.91	0.99	0.99	1.00
SP 7	5.68	2.62	0.98	0.90	0.94
SP 8	3.84	2.82	0.9	0.95	0.98
SP 9	4.87	2.59	0.73	0.79	0.82
SP 10	6.87	2.16	0.92	0.90	0.97
SP 11	1.22	2.37	0.08	0.55	0.58
SP 12	7.13	2.37	0.72	0.50	1.00
SP 13	5.29	2.62	0.16	0.86	0.99
SP 14	5.94	2.14	0.59	0.84	0.97
SP 15	7.38	1.95	0.18	0.92	0.94
SP 16	3.24	2.84	0.71	0.67	0.76
SP 17	4.31	2.88	0.86	0.82	0.96
SP 18	1.45	2.39	0.06	0.34	0.36
SP 19	7.79	2.61	1	0.97	0.95
SP 20	7.82	1.97	0.93	0.73	0.79
SP 21	2.112	3.37	0.06	0.63	0.63
SP 22	6.25	2.72	0.78	0.95	0.99
SP 23	8.16	1.91	0.97	0.99	1.00
SP 24	7.22	2.43	0.93	0.84	0.84
SP 25	7.49	1.92	0.92	0.85	0.85
SP 26	6.33	2.48	0.68	0.74	0.86
SP 27	3.84	2.56	0.92	0.95	0.97
SP 28	1.23	1.87	0.07	0.44	0.43
SP 29	6.07	2.66	0.47	0.71	0.91
SP 30	6.49	2.62	0.79	0.75	0.97

From the results shown in Table II, the pearson's correlation between FAST and the human ratings for the MFWD is 0.77. However, the correlation between STASIS and the MFWD drops down to 0.71 while the level of correlation between LSA and the MFWD drops to 0.63. The decreases in the levels of accuracy from both STASIS and LSA were not however significant; with both losing no more that 1% in accuracy, implying that the increase in the number of fuzzy words in the sentence pairs did not substantially diminish their performance. The fact that the results remained so similar between the three measures is an indication that increasing the number of fuzzy words in pair of fuzzy sentences does not substantially change the performance of any of the three SSM. If the slight decrease in accuracy from both STASIS and LSA continued at a consistent rate for both measures as more fuzzy words were added, then the number of fuzzy words that would be required to make this significant are more than could reasonably be expected to be found in a natural language sentence. The results have overall shown that the presence of fuzzy words changed the semantic meanings of sentences enough to change human perceptions of the levels of similarity between them.

VI. CONCLUSION AND FURTHER WORK

This paper has described the methodology for the creation of a corpus based method of building a fuzzy dataset known as MFWD. The methodology incorporates the use of a fuzzy sentence pairing algorithm which is used to automatically generate a set of low, medium and high sentence pairs that contain two fuzzy words. The algorithm uses predefined categories of fuzzy words that have been quantified by human participants. Fuzzy words were selected from a set of pre-defined categories of fuzzy words that have been quantified by human participants. Crowdsourcing and traditional questionnaires were used to obtain human sentence ratings for MFWD. The results have shown that the FAST measure returned a high level of correlation with human ratings while this was not that case with traditional SSM's STASIS or LSA. While the accuracy of FAST remained high, the accuracy of STASIS declined and the accuracy of LSA remained comparatively low. This therefore showed that FAST was a highly suited replacement to existing non fuzzy semantic similarity measures in the area of fuzzy sentences. Further work includes the expansion of fuzzy categories using a less human intensive method such as [22]. This will allow creation of automatic fuzzy datasets which have much more coverage of natural language. The question is - is it possible to create a generic 'codebook' of quantified words that is not domain or context dependent?

VII. REFERENCES

[1] L 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.

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

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

[4] 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: 02/1/15

[5] O.Shea, K. Crockett, K. Bandar, Z. O'Shea, J. An approach to conversational agent design using semantic sentence similarity, Journal of Applied Intelligence, Vol (40): 1, pp. 199-199, 2014.

[6] Alzahrani, S.M. Salim, N. Abraham, A. Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, V:42:2, pp. 133 – 149, 2012.

[7] Du, R. Yu, Z. Mei, T. Wang, Z. Wang, Z. Guo. B. Predicting activity attendance in event-based social networks: content, context and social influence. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, USA, pp. 425-434, 2014

[8] Ogawa, Y., Morita, T., Kobayashi, K.: A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method. Fuzzy Sets and Systems, 39, pp.163–179, 1991.

[9] Yerra, R., Ng, Y.-K. A Sentence-Based Copy Detection Approach for Web Documents. In Fuzzy Systems and Knowledge Discovery, pp. 557-570, 2005.

[10] Alzahrani, S. M., & Salim, N. On the Use of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents. Paper presented at the Second International Conference on the Applications of Digital Information and Web Technologies, UK, 2009

[11] 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, doi: 10.1109/FUZZ-IEEE.2013.6622494, 2013

[12] Rosa, H.; Inst. Super. Tecnico, Univ. de Lisboa, Lisbon, Portugal ; Batista, F. ; Carvalho, J.P., Twitter Topic Fuzzy Fingerprints, 2014 IEEE International Conference on Fuzzy Systems, pp, 776 – 783, DOI:10.1109/FUZZ-IEEE.2014.6891781, 2014

[13] Chandran, D.; Crockett, K.; Mclean, D.; Bandar, Z., FAST: A fuzzy semantic sentence similarity measure, 2013 IEEE International Conference on Fuzzy Systems , 2013

[14] O'shea, J. Bandar, Z. Crockett, K. A new benchmark dataset with production methodology for short text semantic similarity algorithms. ACM Trans. Speech Lang. Process. 10, 4, Article 19, 2014

[15] Agirre, E., Cer, D., Diab, M., and Gonzalez-agirre, A. 2012a. Semeval-2012 task 6: A pilot on semantic textual similarity. Joint Conference on Lexical and Computational Semantics (SEM'12). Association for Computational Linguistics, pp. 385–393, 2012.

[16] 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, pp.103-120, 2010.

[17] Chandran, D. Crockett, K. McLean D, On the creation of a fuzzy dataset for the evaluation of fuzzy semantic similarity measures, FUZZ-IEEE 2014, pp.752-759, 2014.

[18] Zadeh, Lotfi A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy sets and systems 90.2, pp.111-127, 2007

[19] Mendal, J.M. Donggrui, Wu. Perceptual Computing: Aiding People in Making Subjective Judgments, Wiley-IEEE, 2010.

[20] Dongrui Wu ; Mendel, J.M. ; Coupland, S. Enhanced Interval approach for Encoding Words Into Interval Type-2 Fuzzy Sets and Its Convergence Analysis, IEEE Transactions on Fuzzy Systems, V: 20:3, pp.499 – 513, 2012.

[21] Minshen H. Mendel, J.M. Modeling words by normal interval type-2 fuzzy sets. 2014 IEEE Conference on Norbert Wiener in the 21st Century, pp. 1 – 8, 2014.

[22] Mendel, J.M. ; Dongrui Wu, 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.

[23] Zadeh, L. Computing with Words: Principal Concepts and Ideas, Springer Publishing Company, 2012

[24] Nelson, F. A standard corpus of edited present-day American English. College English 26.4, pp.267-273, 1965

[25] 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.

[26] Hart, Michael. Project Gutenberg. Project Gutenberg, 1971.

[27] Schmidt, D. Colomb, R. A data structure for representing multi-version texts online. International Journal of Human-Computer Studies67.6, pp.497-514, 2009

[28] Carvalho, V, Lease, M. Yilmaz, E. Crowdsourcing for search evaluation. ACM Sigir forum. Vol. 44:2, ACM, 2011