THE DEVELOPMENT OF A FUZZY SEMANTIC SENTENCE SIMILARITY MEASURE

By

GAUTAM DAVID CHANDRAN

A Thesis submitted in fulfilment of the requirements of the Manchester Metropolitan University for the Degree of Doctorate of Philosophy

School of Computing, Maths and Digital Technology

Manchester Metropolitan University

2013

Acknowledgements

I'd like to firstly thank my supervisors Dr Keeley Crockett, Dr David McLean and Dr Zuhair Bandar for their tireless assistance and support throughout the project. I would also like to thank my family and my wonderful fiancée who supported and stood by me throughout my research. I would also like to acknowledge my many colleagues at Manchester Metropolitan University who supported and stood by me.

Abstract

A problem in the field of semantic sentence similarity is the inability of sentence similarity measures to accurately represent the effect perception based (fuzzy) words, which are commonly used in natural language, have on sentence similarity. This research project developed a new sentence similarity measure to solve this problem. The new measure, Fuzzy Algorithm for Similarity Testing (FAST) is a novel ontology-based similarity measure that uses concepts of fuzzy and computing with words to allow for the accurate representation of fuzzy based words. Through human experimentation fuzzy sets were created for six categories of words based on their levels of association with particular concepts. These fuzzy sets were then defuzzified and the results used to create new ontological relations between the fuzzy words contained within them and from that a new fuzzy ontology was created. Using these relationships allows for the creation of a new ontology-based fuzzy semantic text similarity algorithm that is able to show the effect of fuzzy words on computing sentence similarity as well as the effect that fuzzy words have on non-fuzzy words within a sentence. In order to evaluate FAST, two new test datasets were created through the use of questionnaire based human experimentation. This involved the generation of a robust methodology for creating usable fuzzy datasets (including an automated method that was used to create one of the two fuzzy datasets). FAST was evaluated through experiments conducted using the new fuzzy datasets. The results of the evaluation showed that there was an improved level of correlation between FAST and human test results over two existing sentence similarity measures demonstrating its success in representing the similarity between pairs of sentences containing fuzzy words.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Background and Motivation

Sentence similarity is the process by which algorithms determine how alike sets of text are to each other through returning a similarity value between them (Li et al. 2006) (Islam and Inkpen 2008). It is a fast developing area that has an increasing number of applications in a number of different fields (Foltz et al. 1999a) (Lemaire and Dessus 2001) (Bigham 2007). The earliest sentence similarity measures utilized the syntactic likeness between pairs of text to determine their level of similarity to each other (Salton and Lesk 1968) (Salton and Buckley 1988). This involved comparing the locations of common words in the texts and determining how close they were to each other (with a greater level of closeness determining a higher level of similarity). A problem with these types of approaches was that they did not deal with sentences that were syntactically similar but semantically different, for example;

“The dog is in a kennel”

and

“The man is in a house”

Therefore, new approaches were required that were able to deal with the semantic component of sentence similarity. Towards this end, new measures were developed. The first and most popular of these was called Latent Semantic Analysis (LSA), which determined the semantic similarity of texts based on statistics derived from the occurrences of the words in corpuses (Landauer et al. 1998). When LSA was created, it was built specifically to deal with large sets of texts (Landauer et al. 1998) (Li et al. 2006) but there was still a need for a short text similarity measure. As a result, the STASIS similarity measure was created (Li et al. 2006), which took an ontology-based approach to determining similarity for short sets of text (with 35 words or fewer). LSA and STASIS are discussed in detail, in

Chapter 2. A fundamental problem with both of these measures was that they were unable to accurately determine the level of similarity between words with subjective meanings that are based on human perception such as;

“big”

or

“good”.

Words with subjective definitions (as opposed to words with crisp, objective definitions) are defined as Fuzzy Words (Zadeh 1999). Other examples of such words are ‘huge’, ‘tiny’, ‘hot’ and ‘giant’. As a result of this problem, existing sentence similarity measures have been unable to accurately determine the level of similarity between sets of text with these words within them. This is a substantial challenge as it prevents sentence similarity measures from accurately representing natural language sentences that are commonly used in human dialogue (that frequently contain perception based words). The main motivation behind this research project has been to address the challenge of creating sentence similarity measures that are able to accurately represent fuzzy words. The aim of this project is to create a new fuzzy sentence similarity measure that can represent the level of similarity between perception based words more accurately than any existing Sentence Similarity Measure. Towards this end, examination of the field of Computing with Words (CWW) was required (Zadeh 1996) (Zadeh 1999). This was a field that was specifically based on the representation of perception based words (or fuzzy words) towards computer systems. This field is an expansion of the field of fuzzy sets (Zadeh 1996) (Zadeh 2008) (Mendel 2007a). Both CWW and fuzzy sets are explored in greater detail in Chapter 2. Through using techniques developed in CWW that dealt with fuzzy words, a methodology could be developed to create the new fuzzy text similarity measure. The measure that was developed was named Fuzzy Algorithm for Similarity Testing (FAST).

1.2. Research Goals and Objectives

- Determine quantitative relationships between a large set of fuzzy words based on how they are scaled against each other and represent these relationships in a suitable structure.
- Use the relationships between fuzzy words to create a fuzzy word similarity measure that can return a semantic similarity value for pairs of fuzzy words.
- Implement the fuzzy words similarity measure into a sentence similarity measure (FAST), this system should allow for the comparison of pairs of texts with fuzzy components, returning a similarity value between them.
- Evaluate all aspects of the FAST measure and benchmark it against existing sentence similarity measures. This involves the creation of a suitable dataset for the evaluation to be based on.

1.3 Contributions

The main contributions of the project are:

- A Fuzzy word similarity measure and a description of the methodology used to build it. This includes
    - A method for quantifying fuzzy words using human participants
    - A method for determining the relationships between the words based on these quantities.
    - A formula for using the relationships between words to determine the level of similarity between pairs of words.
- The FAST sentence similarity measure, which is a functional fuzzy sentence similarity measure that can be implemented in any system that has a sentence similarity component.

- A new dataset for evaluation of a fuzzy sentence similarity measure and a method for creating this dataset. This includes
    - A Single Fuzzy Word Dataset (SFWD) which contains pairs of sentences with a single fuzzy word in each sentence and the similarity between them and the novel methodology that was used to create the dataset
    - A Multiple Fuzzy Word Dataset (MFWD) which contains pairs of sentences with multiple fuzzy words in each sentence and the similarity between them and the novel methodology that was used to create the dataset

1.4 Thesis Structure

The following subsections detail the structure of the thesis.

1.4.1. Overview of Chapter 2: Literature review

Chapter 2 involved background research into the challenges present in creating the FAST similarity measure and approaches that could provide solutions to these challenges and allow FAST to be built. This was focussed in three areas;

1) The field of sentence similarity
2) The field of CWW and Fuzzy Sets
3) Ontology creation

The first area involved an in-depth analysis of the history and current state of the art in the field of sentence similarity. This involved looking at the different measures that were built, including STASIS and LSA, exploring how they operated and determining how successful they were (Deerwester et al. 1990) (Landauer et al.1998) (Li et al. 2006) (Islam and Inkpen 2008) . This was to provide evidence of whether or not STASIS' ontology-based method would be suitable for FAST.

The second area involved an exploration of the areas of fuzzy and CWW (Zadeh 1996) (Zadeh 1999). This is because these fields areas discussed the representation of fuzzy words in a manner that computer systems could understand. The focus of the literature review was on how fuzzy words could be considered as components of a larger concept and on how fuzzy words could be represented by a range of quantities through the use of fuzzy sets. This provided an important reference on how fuzzy words could be quantified and scaled against each other (which could then be used to determine how similar they are to each other)

The third area explored was the construction of Ontological structures. This is because the FAST measure makes use of ontologies to determine semantic similarity (such as other semantic similarity measures such as STASIS (Li et al. 2006) and OMIOTIS (Tsatsaronis et al. 2009)). The justification for using an ontology-based methodology is discussed in Chapter 2. It was therefore important to explore different methodologies and standards that need to be adhered to in creating ontologies and determining which ones would be most appropriate to use in creating FAST.

1.4.2. Overview of Chapter 3: Collecting and Quantifying sets of Fuzzy Words

Chapter 3 involved the creation of six categories of fuzzy words and the quantification of the words within those categories on a particular scale. Before the FAST measure could be created, sets of fuzzy words would have to be scaled against each other, to allow their similarity to be determined. For this to be done (and to allow the relationships between the words to be determined), the fuzzy words had to be first quantified on given scales. The process required that sets of fuzzy words be collected around a particular concept (or category) that could be used to construct a scale for the sets of words to be quantified on. The whole process was conducted using human testing with two sets of empirical experiments.

1) Populating a set of categories with fuzzy words.

2) Quantifying the sets of fuzzy words.

The first experiment involved a set of participants filling sets of pre-determined fuzzy categories with words that they thought were appropriate for these categories. These categories contained words based around particular subjects (such as size or temperature), with the intention of them holding a large number of commonly used words. The categories that were used were selected based on the large number of fuzzy words they could hold. Through the results of these experiments, six complete sets of fuzzy words were returned. The second experiment involved determining quantities for the fuzzy words. This was done using methods that were developed by Mendel in his work on CWW and Fuzzy Words (Mendel 2007a) (this is discussed in *Chapters 2* and *3*). The method involved giving a group of participants a scale and asking them to quantify the words in each of the categories on that scale. Through methods that were inspired by previous work done in the field of CWW (which are explored in *Chapter 2*), the results from the experiment were used to return representative quantities for the fuzzy words. At the end of this stage of the project was a set of six distinct categories of quantified fuzzy words were generated. The categories were

- Size
- Temperature
- Goodness
- Frequency
- Age
- Level of Membership

The scales contained within the categories could then be used to create ontological structures to determine the overall level of similarity between pairs of fuzzy words within a given category. This is discussed in more detail, in *Chapter 4*.

1.4.3 Overview of Chapter 4: Implementation of FAST

Chapter 4 involved implementing the FAST similarity measure. This was done through determining the similarities between pairs of fuzzy words based on the quantities on given scales that had been calculated in Chapter 3. These similarities were then used to implement a complete similarity measure.

From the conclusions of the literature review chapter, it was decided that the word similarity component of FAST should be created through use of ontologies (Noy and McGuinness 2001) (Gruber 1993). For each category, the words would be placed into an ontology (Gruber 1993), and their relations determined by their relative positions within it. Two candidate ontology structures were created to later (Chapter 6) be evaluated against each other in terms of effectiveness at determining similarity at a later stage.

A formula was developed that allowed the ontological distances between words and their distances to a common subsumer to be used to calculate the semantic similarity value between the two words. This formula was inspired by the work done in by Li et al. (2003).The same formula was applied to both the candidate ontology structures. With this formula, the similarity between any pairs of fuzzy words within a given category could be calculated.

After the ontology-based word similarity component had been created, the component was adapted into wider text similarity measure (that could return a similarity value for pairs of sentences that contained fuzzy words). To do this, inspiration was taken from the work that was done with STASIS, another ontology-based similarity measure (Li et al. 2006). The results of this stage of the project were two implementations of FAST, one for each of the ontology structures. The aim was to test each of them during evaluation and determine which one could more accurately determine sentence similarity.

1.4.4 Overview of Chapter 5: Creating a Fuzzy Evaluation Dataset

Chapter 5 describes the creation of fuzzy datasets. Before FAST could be evaluated there needed to be a suitable evaluation dataset. This dataset needed to contain pairs of sentences (short texts) with human ratings stating how similar they were to each other. Unfortunately, none of the existing datasets contained a suitable number of fuzzy words. Therefore, a new dataset had to be built for this purpose. This new dataset comprised of two smaller datasets, a set of sentences containing one fuzzy word per sentence, the Single Fuzzy Word Dataset (SFWD) and a dataset containing multiple fuzzy words per sentence, the Multiple Fuzzy Word Dataset (MFWD). Methodologies inspired by the work done in.(O'Shea et al. 2008a) were adapted to create these datasets.

The creation of the SFWD involved fuzzifying (adding a fuzzy component) to words in an existing dataset of sentence pairs. The dataset that was chosen was created by James O'Shea (2010), the STSS-131 dataset. The process of fuzzifying the sentences was undertaken by a panel of experts. Once a set of fuzzy sentence pairs had been created, they then had to have the levels of similarity between their constituent sentences determined. This was done through a set of human participants rating each sentence pair based on its level of similarity. The methodology was inspired by the work on dataset development by James O'Shea (2010).

The creation of the MFWD involved extracting fuzzy sentences from a large corpus and creating new sentences for them to be paired with through replacing their fuzzy words with other random fuzzy words. An algorithm was designed and implemented to accomplish this goal. Once the fuzzy sentence pairs had been created, they were then rated against each other in terms of their levels of similarity through use of human participants as had been done with the SFWD.

1.4.5 Overview of Chapter 6 : Experimental Results

Chapter 6 described the procedure through which all aspects of the FAST measure were fully evaluated. This evaluation was now possible based on the datasets that were created in Chapter 5. Firstly it involved determining the level of effect that the presence of fuzzy words had on sentence similarity. The next stage of Chapter 6 involved determining the effect that fuzzy sentence pairs had on existing sentence similarity measures. This was done through selecting two measures (STASIS and LSA) and comparing their accuracy with the STSS-131 dataset to their accuracy with the SFWD. Through examining the results of this, it was possible to determine if the accuracy of the measures decreased when dealing with fuzzy words.

The next stage involved determining which ontology structure to use to form a general implementation of FAST. The implementations were tested against both the SFWD and the MFWD. Through looking at these results, a clear picture was returned about the effectiveness of the structures. Through using that information, a general implementation was selected.

The final stage of the evaluation involved testing FAST, LSA and STASIS
The final stage of the evaluation involved testing FAST, LSA and STASIS
against SFWD and MFWD. Through this evaluation, a clear picture is returned about the effectiveness of FAST as a measure and how useful it would be for any future work. From the testing, FAST was shown to have returned a statistically significant improvement over STASIS and LSA. This showed that FAST and its ontology-based methodology was able to address the challenge or representing fuzzy words.

1.5 Conclusion

In conclusion, this project chronicles the work that was required to create the FAST sentence similarity measure; this involves the research, the different challenges that had to be addressed in its implementation and the results of its evaluation. Through the successful completion of the project, the issue of creating a fuzzy semantic similarity measure is addressed.

A paper was presented at The **2013 IEEE** International Conference on Fuzzy Systems that provided a brief description of the creation of FAST (Chandran et al. 2013). (Appendix 5).

A paper has been accepted at the **2014 IEEE** International Conference on Fuzzy Systems that provides a description of the creation of the Evaluation dataset (Chandran et al. 2014) (Appendix 6).

2. Literature review

2.1. Overview

This chapter contains a detailed exploration of all the background material that was reviewed in preparation for the research project. The background research focussed on three particular areas.

- The evolution and state of the art of the areas of Fuzzy and Computing with Words.
- The history and development of Sentence Similarity Measures (SSM).
- The general development of ontological structures and how they relate to sentence similarity.

The goal of the project, as stated in the introduction (Chapter 1), was to create a sentence similarity measure that can accurately represent human perceptions. Therefore, the first objective (Section 2.2) was to review the field of fuzzy logic focusing on Computing with Words (CWW). This gives an illustration of the different methodologies that are in place to deal with subjective (or fuzzy words), in terms of representing them to computer systems.

The second objective (Section 2.3) was an exploration into the background of sentence similarity measures. As was discussed in chapter 1, these are systems that are able to take in two sets of text as input and return a single similarity value to denote how alike they are. This objective provides important context for the project, particularly in terms of what existing measures and methodologies may be implemented towards the project being successful.

The third objective (Section 2.4) deals with the concept of ontologies. As is discussed in Section 2.3, one of the most successful semantic text similarity measures was the STASIS similarity measure (Li et al. 2006). STASIS dealt with similarity through use of an ontological structure (differing from the methods used in other similarity measures). This was a novel approach that allowed the similarities between every word pair combination in two

sentences to be calculated and then have those similarities used to calculate the semantic similarity value. This method used ontological structures to determine the inter-relatedness of particular words. Therefore, the third objective of the literature review is to explore the area of ontologies with a focus on its applicability to the concept of creating a fuzzy sentence similarity measure.

2.2. Fuzzy Sets and Computing with Words

2.2.1. Introduction

In his seminal 1999 paper and subsequent work, Lofti Zadeh identified a significant issue in human-computer communication (Zadeh 1999) (Zadeh 2008). This was the introduction of CWW as a field. He noted that while computers tend to communicate with each other using crisp quantities, humans tend to communicate information to each other using perception based words. These are words whose meanings are dependent on an individual's previous experience with those words. For example, a human being when describing (to another human being) the location of a nearby area to another area might use an expression such as;

"It is a short walk from here".

Whereas if a computer system was trying to communicate that information (to either a human or another computer system) it may use an expression such as

"It is 20 metres away" – a more precise answer.

Zadeh noted that the perception based approach that humans used enabled many accomplishments (an example Zadeh provided was the moon landing with a discussion on how our ability to communicate issues in that endeavour in terms of perceptions enabled it). However, given that it allowed people to communicate quantities in an abstract manner, it had also been problematic in that it had limited the scope of human-computer communication. It created a situation where a computer could not perfectly understand what a human's intent was when the human worded a statement in terms of their perceptions. To deal with this issue, Zadeh created a new

framework called CWW (Zadeh 1999) (Zadeh 1996) through which perception-based words could be communicated to computer systems. CWW expanded on the pioneering work that was done by Zadeh in the field of fuzzy sets (Zadeh 1965) (Zadeh 1999). Therefore, before any further discussion of CWW can take place, the area of fuzzy sets must be briefly examined.

2.2.2. Fuzzy Sets

The concept of fuzzy sets was first introduced by Zadeh in (Zadeh 1965). This paper moved away from the existing model of set theory which stated that an element was either completely a member of a set or it was completely outside of it

$$e \in \{0,1\}$$

Under this model, a set of items would either have to completely describe the items with it or have no relationship with them. This was adequate to describe sets where membership could easily be stated as true or false. For example, given the set 'Fruit'; membership and nonmembership can easily be determined. Membership in the set is crisp, an item can either be a fruit (and a member of the set, having a value of 1), or not a fruit and have a value of 0.

$$Fruit = \{Apple, Orange, Banana \dots\}$$

$$Fruit \neq \{Car, Bread, Cat \dots\}$$

The Zadeh model (called fuzzy sets) instead described partial membership. Zadeh stated that elements do not have to either completely belong to a set or be completely outside of one. Instead, they could be partial members of a set, containing a level of membership while not being a full member. This level of membership was termed an element's membership function. Zadeh defined fuzzy sets as

"*A fuzzy set A in X is characterised by a membership function f(x) which associations each point in X with a real number in the interval [0,1] with a value of f(x) at x representing a" grade of membership" of x in A"*

The higher an elements' grade of membership or membership function, the greater its level of membership in a set with a membership function of 1 denoting full membership of a set.

The advantage of these sets was that they could represent sets where the levels of membership of elements are partial and where different elements have different levels of membership (Zadeh 1965) (Zadeh 1997). For example, consider the set of entities that could be represented by the statement "This is an animal that moves fast". The definition of the word fast is completely subjective. As a result of this, a number of different animals would have different levels of membership in the set depending on their speed. As illustrated by Zadeh, various entities would have different values of membership in a set based on how "true" the statement that they move fast was. For example, a sloth, a human, a horse, a cheetah and a race-car could have the values 0.01, 0.3, 0.6 and 0.9 respectively. This could be presented as a fuzzy set (with isv being defined as a veristic constraint to illustrate a level of truth (Zadeh 1997) (Zadeh 1999)):

$$Fast(animal)isv(0.01|Sloth + 0.3|Human + 0.6|Horse + 0.9|Cheetah)$$

In the given example, the statement "Is an animal that moves fast" had varying degrees of truth for each of the entities that were presented. This illustrates how Fuzzy Sets could be used to deal with sets where membership is non-binary and concepts where values are subjective. As work in the field of fuzzy progressed new types of fuzzy sets (that are discussed in Section 2.2.3) were created. Therefore, classic fuzzy sets (which are described in this section) are referred to as type-1 fuzzy sets.

2.2.3. Type 2 Fuzzy Sets

The work that was done by Zadeh and others (Zadeh 1973) (Zadeh 1975) (Yager 1980) expanded on the initial work on fuzzy sets by creating a new fuzzy set. While the initial fuzzy sets created by Zadeh could be used to show how elements might have partial membership of a set, they still had an unresolved issue. The issue of uncertainty about the fuzzy membership functions within a fuzzy set needed to be addressed. For example, consider

a fuzzy set y with a member x. There is uncertainty about whether the membership function of x is 0.2, 0.3 or 0.4 (as is the case for all other members of y). This creates a problem in terms of accurate representation due to the uncertainty regarding which of the values should represent the membership function.

Acknowledging that there could be fuzziness between elements in a fuzzy set, Zadeh created the concept of type 2 fuzzy sets. These were created to represent the issue of fuzziness about the membership functions of elements within existing fuzzy sets (referred to henceforth as Type-1 fuzzy sets) and as such deal with the issue. A type-2 fuzzy set is defined as a set of type 1 fuzzy sets. For example, consider three type-1 fuzzy sets, f1,f2 and f3 related to the same concept but with each containing different ranges of values (e.g. f1 = {0.6,0.7,0.8}, f2 = {0.5,0.6,0.7} and f3 = {0.7,0.8,0.9}). A type-2 fuzzy set F is a set that could hold each of the type-1 sets,

F = {(0.6, 0.7, 0.8), (0.5, 0.6, 0.7), (0.7, 0.8, 0.9)}

Or

F = {f1, f2, f3}

Each of the three type 1 fuzzy sets is now an element of a type two fuzzy set and their differences can be easily represented. Type-2 Fuzzy Sets are therefore an effective method of illustrating the fuzziness between the boundaries of fuzzy sets. The work that Zadeh did on Type 2 fuzzy was further developed by other researchers in the fuzzy community (Zadeh 1965) (Mizumoto and Tanaka 1976) (Lee and Wang 2011). Work of particular importance was done by Jerry Mendel (Mendel and John 2002) (Mendel et al. 2006).

A type-2 fuzzy set could be transformed into a type-1 fuzzy set with representative values through a process called type reduction. There are a number of different methods of type reduction that have been developed (Zadeh 1975) (Mizumoto and Tanaka 1976) (Tanaka et al. 2000). This is important as it allows the fuzzy membership functions to be represented as crisp membership functions. This further allows them to be used in

processes where crisp membership functions for fuzzy words are required. A method that was explored in detail by Mendel and was demonstrated to have a high level of accuracy is to take a centroid based approach (Karnik and Mendel 2001) (Mendel et al. 2006) (Mendel and Wu 2007). That is, for each of the fuzzy sets in a type-2 fuzzy set, their centroids were taken as a representative value for the element's membership function. For example, consider the earlier type-2 fuzzy set that had been defined in this section. If the centroids of the values f1, f2 and f3 were 0.5, 0.7 and 0.9 respectively, a new type-reduced type-1 fuzzy set F1 could be defined as:

$$F1 = \{0.5, 0.7, 0.9\}$$

This solution created a robust method through which uncertainty in fuzzy sets could be dealt with. For example, if in a type-2 fuzzy set of “hot things”, there were three distinct membership functions for the entity “warm”, through type reduction these three membership functions could be projected onto a single membership function. This process is not limited to type-1 fuzzy sets as, if required, uncertainty in type-2 fuzzy sets could be handled using a similar process (defining a new set that is composed of a set of type-2 fuzzy sets).

2.2.4. Computing with Words

Zadeh (1999) expanded upon the work that he had previously done in the field of fuzzy set theory with the creation of CWW. This was to address the issue of representing human perceptions to machines. Zadeh put forward the idea that that perception based (fuzzy) words would cover a range of values effectively being represented by a fuzzy set (Zadeh 1996) (Zadeh 1999). For example if a person was to state where on a scale of temperature the word “warm” would be. A person might consider it to cover an area on the scale rather than a single point. To this end Zadeh talked about the concept of granularity that explored the association of multiple concepts around a single concept (or granule). For example, a second, an hour and a day would all be related to the concept or granule of time, which encompassed all of them. Zadeh discussed how different entities could have different levels of association with a particular concept (for example, if

we were to consider the concept of hotness, “hot” would have a higher level of association with the concept than “lukewarm”) (Zadeh 1996) (Zadeh 1999). This was an expansion on the work that had been done on fuzzy type 1. Each fuzzy element associated with the granule would be members of a fuzzy set with different membership functions. Through this, the different values covered by perceptions of a group of fuzzy words could theoretically be represented in terms of a concept they are associated with. Further expansion on Zadeh’s work in CWW came from Jerry Mendel who applied fuzzy type-2 methods to CWW (Mendel 2007a) (Mendel 2007c) (Wu and Mendel 2007a). Mendel noted that perceptions around words differed from individual to individual. The fact that different people have varying views on fuzzy words meant that the differences between individuals needed to be represented, as well. For an illustration of this, consider the illustration presented by Mendel (Mendel 2007a) regarding the word “some”, showing a set of fuzzy sets, from 3 individuals, containing the range of membership functions of the word “some” on a given scale.

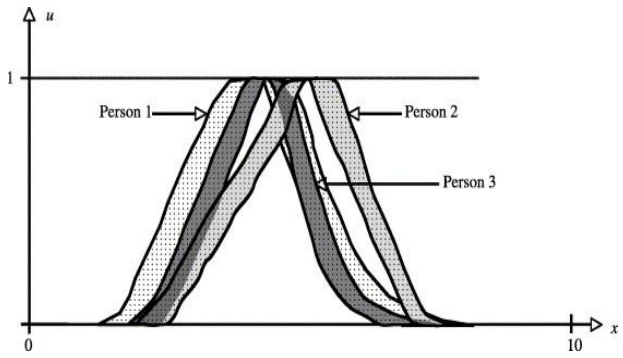

Figure 1 A type-2 fuzzy set (Mendel 2007a)

The use of a type-2 fuzzy set allowed for the representation the range of different perceptions about a particular word (this allowed for the collection of type 1 fuzzy sets from a range of people that were then made elements of a type-2 fuzzy set). Therefore in Mendel’s approach the fuzziness in the boundaries between elements of the type-2 fuzzy set had to be able to be represented to a computer system. In order to reduce a fuzzy type-2 set into a fuzzy type-1. Mendel had earlier proposed determining the centroid of a

type-2 set (Karnik and Mendel 2001) (Mendel et al. 2006), towards type reduction (projecting the type-2 set to a type-1 set). This approach also allowed for the representation of the level of uncertainty that was present in the fuzzy type-2 set (Karnik and Mendel 2001). Mendel described the centroid of a type-2 fuzzy set as “The union of the centroids of its T1 FSs”. This could then be defuzzified, to return a single value. This single value could then be taken as a representative value for level of association a particular element had with a granule based on the perceptions of a multitude of people rather than just one.

Liu and Mendel (2008) developed and implemented a methodology to create a “codebook” to determine the Footprint of Uncertainty (FOU) of 32 fuzzy Type-2 sets each based on a fuzzy word (with the FOU of a type-2 fuzzy set being 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 a particular area (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 T2 fuzzy sets of each word and a mean value for each of them was returned (using the method described by Liu and Mendel (2008)). 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. While Mendel cautioned against using the specific values that were found in the codebook, he suggested that the methodologies he proposed could be used for further experimentation. While this does mean that for any work that is done in this area; new words will need to be collected, it does provide a good reference point and framework.

2.2.5. Applications of CWW

There is a wide range of practical applications of CWW in a wide range of fields. Many of these were identified and explored by Zadeh in his seminal paper (Zadeh 1999) where he first described the concept. Of particular importance is the use of CWW in systems where human-computer communication is required and the human needs to converse with a computer in a natural manner (Zadeh 1996) (Zadeh 1999) (O'Shea 2012a) (Latham et al. 2012). The applicability of CWW and the concept of granularity to the field of data-mining and the applicability of CWW in the area of intelligent database querying as was described by Lin et al. (2002), Martinez et al. (2010) and Herrera et al. (2006). Furthermore, Herrera et al. (2009) described the use of CWW in the area of decision making and stated its use in terms of decision making systems such as expert systems.

2.2.6. Relevance to Project

The 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 through the use of 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 as a method to determine their similarity and from that create a sentence similarity measure.

The main strength of the CWW approach is that it provides a framework through which fuzzy words can be quantified, scaled against each other and then represented to a computer system. The scaling of fuzzy words against each other is a critical step in creating a fuzzy sentence similarity measure. There are some weaknesses in CWW, however, which must be considered. Firstly, given the size of the language and the fact that it is continually evolving, it is unlikely that all the fuzzy words in a given language would be able to be covered with CWW. To create a comprehensive representation of a given language with CWW would require a vast amount of data collection.

2.3. Semantic Sentence Similarity Measures

2.3.1. 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. The premise of Sentence Similarity is the comparison of sets of text to determine the level of similarity between them. This is done through the use of various semantic similarity algorithms. One area where it has a particular level of importance is that of interactive intelligent systems, crucially, conversational agents (Li et al. 2004) (K.O'Shea et al. 2008). These are computer systems with which humans are able to converse through the use of natural language and receive intelligent responses (Li et al. 2004). Systems that use sentence similarity for human-computer interaction do so through comparing human input with a knowledge base that the system holds. By using an algorithm to determine which members of the knowledge base the input has the highest level of similarity with, the system can intelligently determine which rules to fire and thus what the response of the input will be. As these systems continue to develop and technology advances they are capable of performing progressively more complex tasks and fulfilling increasingly complex demands.

The earliest of these sentence similarity measures (SSM) were based on the premise of determining similarity based on a comparison of syntax (Lewis and Croft 1989) (Salton and Lesk 1968) 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 value 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 the similarity of their meanings has on the overall level of similarity (Li et al. 2006)..

2.3.2. LSA

The first sentence similarity measure that was able to factor in the level of semantic similarity was the seminal Latent Semantic Analysis, the creation of which is described by Deerwester et al. (1990). Using a corpus based approach, this system was able to specifically determine the level of semantic similarity between two sets of text. This system worked through the analysis of corpus statistics. The system took words in two blocks of text and referenced them from within a large corpus. Generating statistics based on the occurrences of these words in the corpus allowed the creation of semantic values to determine the level of similarity between the compared sets of text. Tests of the LSA system against a human dataset (O'Shea et al. 2008b) showed that the system returned a high correlation with human results and was able to deliver results with a high level of accuracy (with a positive Pearson's Correlation of 0.884). A weakness that was identified with LSA, however was that it was designed to deal primarily with large sets of text, as opposed to short texts (sentences with a length less than thirty five words) (Landauer et al. 1998) (Li et al. 2006).

2.3.3. STASIS

A new sentence similarity measure called STASIS was proposed for the specific purpose of accurately representing the level of similarity between short pieces of text (Li et al. 2006). This method proposed determining the level of similarity between two sentences through the use of ontological relations between words. The basis of this measure was a word similarity measure that was created earlier (Li et al. 2003). This method expanded on the taxonomy based approach that was taken by Rada et al. (1989) to determine relationships between concepts and entities. The word similarity measure worked through taking through looking at the distances between two words in an ontology and well as the distance between those words and their closest subsumer. By doing this, the measure was able to return a level of semantic similarity for those two words. The ontology that this system

used was the WordNet ontology, a large lexical database that contains ontological relations between large numbers of entities (Miller et al.1990) (Miller 1995) (Pedersen et al. 2004). Having been tested against the highly regarded Rubenstein and Goodenough word-pairs dataset (Rubenstein and Goodenough 1965) the word similarity measure was shown to have a high correlation with human results.

The STASIS system uses the word similarity measure (Li et al. 2003) between all two possible word combinations within two sets of texts. The results from this are used in conjunction with corpus statistics from the Brown corpus (Francis 1965) to create a semantic value. This semantic value is used with a level of similarity derived from comparing the word order between the texts (representing how similar syntactically they are), which is created through the positions of words in the texts, to return an overall level of similarity for the two sentences. O'Shea et al. (2008a), created a dataset showing human similarity ratings between pairs of sentences based on the definitions of words in Rubenstein and Goodenough's dataset. Both LSA and STASIS were compared against these sentences pairs, and while both measures had a high correlation, STASIS was shown to be closer to the human ratings than LSA.

There are a number of ways that the ontology method alongside corpus statistics to determine semantic similarity is advantageous over using solely corpus statistics as similarity measures such as LSA do. This is because certain ontological structures such as WordNet are demonstrated to successfully represent the inter-relatedness between a wide variety of words (Miller 1995) (Pedersen et al. 2004). As such, they can be used to show how related pairs words are to each other and from that it can be derived how similar they are to each other. The success of the ontological approach specifically to word similarity was demonstrated by Li et al. (2003) when an ontology-based word similarity measure was shown to be able to accurately represent the levels of similarity between pairs of words using pre-established datasets (Rubenstein and Goodenough 1965). Further evidence of the success of using the ontological approach to word similarity as a component of a sentence similarity measure was described by O'Shea et al.

(2008b) when STASIS was evaluated against LSA with a sentence pair dataset. The results showed that STASIS was able to more accurately represent sentence similarity than LSA.

There was, however, a problem that existed with the STASIS measure. The WordNet ontology contained insufficient depth between relations between the vast numbers of fuzzy words in the English language. As a result, when comparing two words, even ones as similar as “tiny” and “miniscule” the word similarity component would likely return a similarity of 0. To deal with this, a new measure had to be created. The goal of this measure would be to accurately represent the effect that perception based words had on overall sentence similarity. The measure’s objective was to be able to succeed in doing this in an ontology-based short text sentence similarity measure such as STASIS. This could greatly improve the abilities of a measure to represent the naturalness of human dialogue, which very often contains a significant degree of fuzzy words. For such an ontology-based system to be constructed the relationships between fuzzy words would need to be established. To do this, methods created within the field of CWW can be utilized.

2.3.4. Operation of STASIS

The STASIS algorithm is adapted to deal with words, corpus statistics and syntactic similarity (Li et al. 2006). STASIS takes two sets of text as input. Every pair of words in the texts is referenced in the WordNet ontology (Miller 1995). Their path length, *l*, (the length of the shortest path between them) and their depth *h*, (the subsumer depth) are then retrieved (an illustration of the WordNet ontology is provided in *Section 2.5* (*Figure 2*)).

The level of similarity between the words (w_1 and w_2) is determined with the following formula:

$$sw_1, w_2 = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (1)$$

The parameters α and β , based on calculations done by Li et al. (2003) (2006) take on the values of 0.2 and 0.6 respectively (these values were chosen based on testing done by Li et al. (2003)).

These similarity values are taken along with word frequency information and information on word positions from a short joint word set value(represented as *r* in the following equation) to determine the total level of similarity between the two sentences (T_1 and T_2). Overall similarity (S) is calculated with the following formula.

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

2.3.5. Other Semantic Similarity Measures

Since the establishment of STASIS, a number of other similarity measures have been created (Mitchell and Lapata 2008) (Islam and Inkpen 2008) (Agirre et al. 2009) (Tsatsaronis et al. 2010). Many of these new similarity measures have adopted the corpus based approach towards sentence similarity, with varying levels of success. None of the measures, however, have explicitly addressed the challenge of fuzzy words or have been tested using a dataset that contained large numbers of fuzzy words. It is important to consider these measures to decide if an ontology-based approach would be the most effective approach to take in creating FAST.

A prominent recent method was put forward by Islam and Inkpen (Islam and Inkpen 2008) (Islam and Inkpen 2009), which used 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. Like STASIS, this method aimed to address the problem of similarity between short texts. However it moved away from the ontological approach

that was taken by STASIS and showed an improvement over STASIS using the same Rubenstein and Goodenough dataset that STASIS was initially evaluated with. It needs to be noted, however, that STASIS was subsequently assessed against a more advanced dataset (O'Shea et al. 2008a) (O'Shea et al. 2008b) and was shown to have a high level of accuracy while the Islam and Inkpen measure was never assessed with the dataset (due to how recent the dataset is).

The Islam and Inkpen method was not benchmarked against LSA. However given that STASIS showed an improvement over LSA with that dataset and the Inkpen, and Islam method showed an improvement over STASIS it could be assumed to be able to show an improvement over LSA with that dataset, as well. Neither this nor STASIS were assessed with datasets that contained fuzzy sentences. This method, however, given its use of string matching has no techniques to deal with individual word similarity. This diminishes its usefulness as part of this overall project which requires determining the level of similarity between pairs of fuzzy words in addition to determining the effect of fuzzy words in terms of sentence similarity.

Another important text similarity measure was OMIOTIS (Tsatsaronis et al. 2009) (Tsatsaronis et al. 2010). As with STASIS and the Islam and Inkpen methods it looked at corpus statistics but also took a WordNet based thesaurus approach towards semantic similarity (distinct from the method used by STASIS). The nature of the measure's approach considered the relative distances of words in a semantic network (that was defined by WordNet) and as such was ontological in nature. The results showed that this method was able to represent similarity more accurately than STASIS, the Islam and Inkpen method and LSA when used with them, in terms of the existing (non-fuzzy) datasets. This therefore provides further evidence that a method of determining fuzzy word similarity through considering the ontological inter-relatedness of fuzzy words is a viable strategy.

2.3.6. Disambiguation in Sentence Similarity

A problem that all text similarity measures need to address is that of disambiguation (uncertainty about a word's definition). It is possible that syntactically similar sentences can be semantically very different (or vice versa). For example the sentences;

The male cat is cold;

and

He is a cool cat;

Are syntactically quite similar to each other. There is, however, a clear semantic difference between the words cool and cold and between the two instances of the word cat. Various measures deal with disambiguation in different ways. Corpus based methods (Landauer et al. 1998), do this through examination of the occurrences of words in a corpus to determine the most likely definition (as corpuses contain a great number of words, decisions can be made on their definitions given the frequency of their occurrences in a given context). Ontology-based measures take the definitions that provide the highest level of similarity when comparing words (Li et al. 2006). In spite of the many methods that have been used to deal with disambiguation, it remains a difficult problem to solve.

2.3.7. Developing a Fuzzy Sentence Similarity Measure

It was decided that an ontology-based approach should be used alongside corpus statistics to create the new similarity measure. This is to allow the accurate representation of the relationships between pairs of fuzzy words (creating a word similarity component to the sentence similarity measure) and as a result, allow the measure to represent the levels of similarity between short sets of text. This allows the system to more accurately represent natural language used by humans who normally speak in the form of short sentences. The two fuzzy measures that used ontologies were STASIS and OMIOTIS. Of the two measures that were available, it was decided that the methodology from STASIS should provide a framework for

the system. This is because although OMIOTIS showed a higher correlation with human similarity ratings than STASIS by Tsatsaronis et al. (2010), STASIS is substantially more widely established. It has also been tested with a wider range of data than OMIOTIS (as the STSS-131 dataset had not been developed when OMIOTIS was created). This makes it more reliable as a the reference when creating a framework.

2.4. An Overview of Ontologies

2.4.1. Defining an Ontology

An ontology is a structure that can be used to describe the hierarchical relationships between the entities that are contained within it (Gruber 1995). They differ from other forms of taxonomic relations in that not only does it specify that two entities are related to each other but also defines the nature of their relationship. Ontologies can be used to classify different entities and from this classification determine how closely related they are to each other. This is through looking at their ontological distance between them and the location of their nearest common ancestor. This has allowed them to be used in the field of text and word similarity (as has been explored in Section 2.3). This section provides a detailed exploration of the area of ontologies and presents their importance in terms of developing a fuzzy sentence similarity measure.

2.4.2. Development of Ontologies

From the inception of ontological structures in computer science, they have played an important role in the area of knowledge representation and computer reasoning. In an early paper on the subject, the use of ontology structures in showing hierarchical entity relations was presented (Clancey 1985). This was put forward as a method through which the relations between various concepts could be discovered and represented. The ontology, through use of a directed graph, classified various concepts into categories and sub-categories. Through ontologies, all the properties that an individual entity possessed could be determined based on the categories that it belonged to. The ontology presented by Clancey (1985) allowed for an

entity to have more than one unrelated property (to belong to more than one concept) noting that, in such cases, it was rare that both properties were considered at the same point. The ontology structure also noted that all entities could be considered either as a singular or as part of a greater collective (in the example that was presented a collective of cows was a herd). To deal with this, the paper presented that idea of parallel ontological schemas, one for the entities as individuals and another that dealt with collectives that entities could belong to.

Important work on the use of ontologies in computer science (particularly in terms of knowledge representation was done in a widely cited paper by Thomas Gruber (Gruber 1991)). The paper discussed the very important role that ontologies had so far played in the field of artificial intelligence in terms of knowledge sharing. In terms of that role, the paper proposed a set of design criteria for ontologies to better facilitate their usefulness upon creation. The criteria put forward by the paper were “clarity”, “coherence”, “extendibility”, “minimal encoding bias” and “minimal ontological commitment”. A series of case studies that were presented in the paper (Gruber 1993) illustrated the importance of the criteria. This was the first case of consolidating experience in creating ontologies. These criteria have since played a crucial role in the wider area of ontological creation (Gruninger and Fox 1995) (Corcho et al. 2003) (Duong et al. 2008). It is, therefore, important that they be regarded in terms of any future ontology that is created. This work was further expanded on by Uschold and Gruninger (2004) who proposed steps in creating a unified methodology for creating ontologies. This was through analysing existing methods that could potentially be merged into a single method. An ontology was defined as “An explicit account or representation of some part of a conceptualisation”. This is an adaptation of the definition presented by Uschold and Grunninger (1996). Importantly the paper added two further criteria onto Gruber’s initial ones regarding building an ontology, “Go middle-out” and “Handle Ambiguity”. These two further criteria need to also be considered in terms of future ontology creation. In addition Uschold and Grunniger presented by Uschold and Gruninger (2004) a skeletal method for ontology development,

which acts as a useful reference. Furthermore, they also discussed the practical situations in which ontologies are useful and, importantly, discussed the need for formal languages in both the development and the evaluation of ontologies and presented the barriers to be overcome in ontology design.

An adapted exploration of the meeting points between the various elements of ontology development (including the tools used to develop them) was presented by Uschold and Grunninger (1996). This paper also discussed how the area of ontology development had evolved from the early work in the field (as was described earlier in this section). It demonstrated the increased level of maturity of the field, the fact that the number of tools and methodologies had increased, and how the basic working definitions of ontologies had continued to evolve and diverge with the growth of the field. The papers goal was to determine the level of common ground between the continually developing methodologies, tools and languages. Through its analysis the paper reached a series of important conclusions, two of which were of particular importance. Firstly, even though many of the tools in use were functionally similar, there was very little interoperability between them making them difficult to use in conjunction for ontology development. Furthermore, the paper pointed out that there was increasingly less need for manual ontology development with the current trend in languages for automated tools for the purpose. Given the lack of interoperability however, an issue that remains is that selecting the right framework to develop an ontology remains of importance.

2.4.3. Ontologies in Text Similarity

The development of ontologies has played an important role in the field of semantic similarity. This is particularly shown in terms of determining the level of semantic similarity between pairs of words. It has allowed the creation of new measures to determine the level of similarity between entity classes in either the same or different ontologies (Rodriguez and Egenhofer 2003) (Budanitsky and Hirst 2006) (Li et al. 2003). This work stemmed from the early work done on information retrieval from ontological structures. This

was explored in depth by Rada et al. (1989) where the authors looked at conceptual distance as a method of information retrieval. Specifically it proposed a system of determining the conceptual closeness between Boolean queries and documents. The paper noted that the method that it used returned results that were very close to human results. This work was later expanded upon in Reznik's seminal paper (Resnik 1995). Reznik approached the problem of determining semantic similarity between entities in a taxonomy structure through information retrieval techniques. The system applied a method that deviated from the edge counting method between pairs of words that were connected in the hierarchy that had previously been applied (Rada et al. 1989). Instead, Reznik took a probabilistic approach. This was based on assigning probabilities to individual entities in the ontology based on their frequencies of occurrence in a corpus. The specific lexical ontology that Reznik used was adapted from the WordNet database (which is described in detail in the following section). Subsequent tests of the system showed it to perform well against human results. Resnik later expanded on the work that he had done (Resnik 2011)

Determining levels of similarity through ontologies is based on the fact that entities being more closely related ontologically to each other implies a higher level of similarity (Baldauf et al. 2007) (Miller et al. 1990). Therefore, word and text similarity measures work through taking information about how closely related words are to determine a semantic similarity value between them. As word similarity measures using ontologies have been shown to be successful in representing word similarity (Li et al. 2003), ontological structures present a framework through which the level of similarity between pairs of fuzzy words could be determined. If ontological structures that contained fuzzy words were created, their relatedness could be calculated and through that an overall similarity value could be found. Use of these values could then be applied to expand word similarity to determine the overall level of similarity between pairs of texts.

2.5. WordNet

It is in this context (of the development of ontologies) that the WordNet lexical database needs to be considered. WordNet is a large, widely used lexical database that was described by Miller et al. (1990). WordNet was created to deal with the lack of machine readable lexical databases. This was an issue at the time given the development of more advanced computer system that would be suited to processing lexical information far more efficiently than the dictionary systems of the time allowed. The problem to be addressed was how to create a machine readable database in the most effective and accurate manner possible. To address the issue, the creators took work that had been done in the field of Psycholexicology. This field, proposed by Miller and Johnson-Laird (1976), dealt with the structure of linguistic knowledge and as such was important in developing a system where such knowledge could be represented to a computer system. From the work by Miller (1986) in 1985, WordNet was first proposed. The idea was a linguistic database that could represent words conceptually rather than alphabetically.

While previous work in this area dealt with relatively small samples of words, WordNet would contain a substantially larger number. Specifically, the WordNet database contained a total of 95600 word forms in total (Miller et al. 1990). These words were furthermore organised into sets of synonyms (synsets) based on their shared meanings. This was achievable through the concept of a lexical matrix illustrating multiple word forms with a common meaning or a single word form that encompassed multiple meanings. A distinct feature of WordNet was in the lexical categories that contained the words. Specifically, it used Nouns, Verbs, Adjectives and Adverbs. An issue that was mentioned by Miller et al. (1990) was that words could be present in more than a single category potentially leading to confusion. WordNet also categorised the different relations between words based on synonymy, antonymy, hyponyminy, meronymy and morphological relations. A synonym relation exists between two words if they share the same meaning (belong to the same synset), an antonym relationship exists between two words that have diametrically opposed meanings. Hyponym/hypernym (or conversely

ISA) relations are transitive relations wherein one of the words is a subset of another word (for example car and vehicle), these are discussed in further detail in the next paragraph. Meronym/holonym relations (or HASA) relations are transitive relations where one word is part of a grouping defined by the other. For example “dog” and “pack” would be an example of such a relationship. Morphological relationships are defined as the relationships between the different morphological forms of a particular word for example “car” and “cars”. This categorisation of words makes WordNet far easier for a computer to extract information from than other systems.

One of the most important features of WordNet, particularly in terms of ontological structures and the wider field of word similarity is what was accomplished with nouns and their relations. Of the 57000 nouns that were present in WordNet, a lexical inheritance system was introduced. This categorizes the nouns in a vast lexical tree based on their lexical relationships with others. Superordinate (ISA) relations for each of the nouns towards single points were created. This gave definition to the inter-relatedness of the huge number of nouns that were present. This is a hierarchy that has been defined as an inheritance system. This is because an entity inherits the various properties of all its superordinate words. Therefore, every word is assumed to have not only the properties from its own definition, but also the properties contained in the definitions of its superordinates. For an example see *Figure 2*. This was built through creating a set of 25 different broad categories with “beginner” nouns that large numbers of other nouns inherited values from. For each of these 25 words, a hierarchy was created containing all the nouns that were subordinate to them. Most of the nouns in WordNet inherited from one of those beginner words. It was later observed that there were some common properties among the beginner words that could be described by a small set of nouns. This led to the creation of a small “Tops” file which contained those relations (as well as a central point of “Entity” or “Thing”).

From (Miller et al. 1990) the decision to use an inheritance based system came from work that was done in psycholexicology (continuing towards the initial goal of representing human lexical memory). That

research indicated that human lexical memory operated on an inheritance based system and that people were quicker to ascertain attributes from a closer superordinate than a more distant one. Therefore through usage of the inheritance based model, the WordNet system worked towards effectively emulating the naturalness of human thought (allowing computers to process information in a similar manner to the human mind). This gives WordNet further strength in terms of its use with human-computer dialogue communication systems. This is the reason it was such a suitable candidate to form the basis for the STASIS word similarity measure. *Figure 2* shows the hierarchical relations between the word “car” and all its parent nodes.

car, auto, automobile, machine, motorcar -- (a motor vehicle with four wheels; usually propelled by an internal combustion engine; "he needs a car to get to work")

=> motor vehicle, automotive vehicle -- (a self-propelled wheeled vehicle that does not run on rails)

=> self-propelled vehicle -- (a wheeled vehicle that carries in itself a means of propulsion)

=> wheeled vehicle -- (a vehicle that moves on wheels and usually has a container for transporting things or people; "the oldest known wheeled vehicles were found in Sumer and Syria and date from around 3500 BC")

=> vehicle -- (a conveyance that transports people or objects)

=> conveyance, transport -- (something that serves as a means of transportation)

=> instrumentality, instrumentation -- (an artefact (or system of artefacts) that is instrumental in accomplishing some end)

=> artefact, artefact -- (a man-made object taken as a whole)

=> whole, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")

=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")

=> physical entity -- (an entity that has physical existence)

=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Figure 2 WordNet Ontology hypernyms for word “Car”

Due to the nature of its development, WordNet can play an important role in semantic similarity. This is due to it being structured in such a way as to allow the relationships between the nouns to form a lexical ontology. There have been a number of different semantic similarity measures that have made use of the structure to determine the relatedness of words in the database (Patwardhan and Pedersen 2006) (Duong et al. 2008) (Suchanek et al. 2008). Furthermore, as has been previously discussed, it was the basis for the substantial work that was done by Resnik. Taking Resnik's approach to WordNet, the word similarity measure for STASIS was created (Li et al. 2006) (this has been discussed further in Chapter 2, the literature review). Through using the inheritance based system that was provided with WordNet, word similarity measures were able to determine the relatedness of entities based on how closely they were related.

2.6. Conclusions

In conclusion, the background research provides a clear look at the frameworks that are in place to deal with perception based words using the CWW approach and an overview of how it can be used to attribute quantities to sets of fuzzy words and thus scale them against each other. The review of sentence similarity measures provides a look at the different methods available and illustrates the effectiveness of the ontology-based approach. The further exploration of ontologies demonstrates how ontological structures are able to represent the differences between different words in terms of their semantic meanings (through looking at their distances in the structure). The nature of these structures, therefore, provides a method through which fuzzy words could be scaled against each other and have representations of their differences mapped. Prior to this, the words need to be quantified through consideration of the techniques that are available in fuzzy and CWW. With the completion of the background research, the implementation of the project can begin.

From the research that was done in the background work some important foundations have been created for building FAST. The work done by Zadeh and Mendel with fuzzy sets and CWW illustrated how values could be given

to fuzzy words that could then be represented to a computer system. Furthermore, the work that was done on Granularity illustrated how different words could have different levels of relevance to a particular concept (Zadeh 1996) (Zadeh 1999). In addition, the work that was done on fuzzy type-2 showed how the fact that different individuals have different perceptions about the values of words could be represented (Mendel 2007a) (Liu and Mendel 2008). The research on sentence similarity measures and the further work that was done on ontologies showed how an ontology-based approach could be used to determine the level of similarity between pairs of words.

3. Collecting and Quantifying Fuzzy Words

3.1. Overview

Prior to any work on creating a new text similarity measure the issue of word similarity had to be resolved. This is because calculating the level of similarity between fuzzy words was required to accurately the effect that these words had on sentences that contained them. For example before the level of semantic similarity between the sentences “This is a big tree.” And “That is a small house” could be accurately calculated, the relationship between the words “big” and “small” needed to first be determined. If the quantitative relationships within sets of fuzzy words can be calculated then these relationships can be factored into a new semantic text similarity algorithm. In order to determine these relationships, the words need to be scaled against each other. To accomplish this goal, a methodology needed to be developed to quantify sets of fuzzy words on particular scales. In this chapter, such a methodology is presented. It is based on the work done by Mendel and Zadeh (as was discussed in the literature review). The methodology involves creating a set of categories to contain fuzzy words, populating those categories with words and then, quantifying the fuzzy words against each other based on their level of association within a particular category. This will result in a set of fuzzy words with quantities on a given scale, thus demonstrating the differences between them. This provides a framework from which fuzzy words can be integrated into a text similarity measure.

3.2. Chapter Aims:

- Describe the creation a set of categories to hold sets of fuzzy words
- Describe the generation a set of fuzzy words for each category
- Describe the Quantification each of the fuzzy words on scales related to the categories

3.3. Overview of Quantifying Fuzzy Words

When discussing the concept of Granularity Zadeh stated that different entities could be associated with a larger concept or “granule” (Zadeh 1999).

From his work on fuzzy theory, it was illustrated that different entities could have different levels of association with a given granule (just as entities within a fuzzy set could have different membership functions) (Zadeh 1996) (Zadeh 1997). With his work on creating a codebook, Mendel showed explicitly that Zadeh's work on granularity and CWW could be used to generate quantities to represent words on a given scale (in that particular case a limited set of words related to Size) (Liu and Mendel 2008). As was discussed in the literature review Mendel also proposed a methodology for doing this. This provides a framework for a large set of words to be quantified over a set of categories. Through use of the concepts introduced by Zadeh and Mendel, a set of categories can be created to serve as granules with sets of associated fuzzy words. For example, a category such as "Temperature" could act as a granule and have words such as "Hot", "Cold" or "Lukewarm" associated with it. Therefore, once a method was created and utilized for generating a set of granules they could then be populated with fuzzy words. A granule is illustrated by *Figure 3*

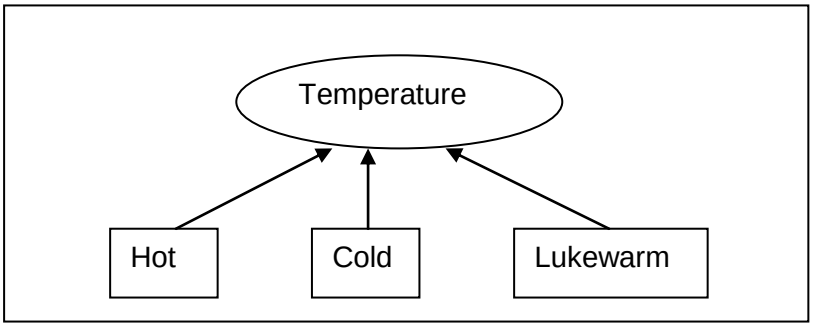

Figure 3: Example of a granule of temperature, a fuzzy concept.

The next stage after that is the method of populating the granules. The objective of this exercise is to determine the relationships between words that can be used in a new sentence similarity measure. For this to be useful, the selected words had to occur in natural human dialogue and so the method used to collect them needed to reflect this. A questionnaire based approach was taken to generate a set of words for each category based on responses from a set of participants. The methodology used in generating these words was based on previous work that had been done by a number of different individuals (Rubenstein and Goodenough 1965) (Miller and Charles 1991) (Charles 2000) (O'Shea 2010). Through processing the

results of these questionnaires, sets of words were then available for all the categories that could be quantified in terms of a common scale (for each category).

3.4. Creating a Set of Fuzzy Categories

As was discussed in the literature review, a codebook for fuzzy words was created by Liu and Mendel (Liu and Mendel 2008). This codebook provided an illustration of how taking Mendel's centroid based method of defuzzification could be used to determine the levels of association a set of words had toward a particular concept. It was based on a type-2 fuzzy set being created for each word and then reduced and defuzzified. Given that Liu and Mendel's experiment only contained one category, for it to be sufficiently expanded a wider range of categories will need to be created to hold a large range of fuzzy words that cover a series of different concepts. An important factor when determining what the categories are is to ensure that they are broad enough to hold a large range of fuzzy words. Furthermore it is important that the category can permit related fuzzy words to each be scaled in terms of their level of association with the category

A limited number of six categories were chosen to ensure that the collection of results was not excessive in scope. This is because each of the categories needed to be populated with words and that these words had to be generated from human experimentation. As such asking respondents to deal with too many categories could have proved an overly complex task for them. When Zadeh first described CWW (1999), he talked in detail about three categories (size, distance and age) as granules and so it was decided that these categories should be used. Size and Distance were then merged into a single one due to the large level of overlap between them. It was, therefore, decided that six categories should be used. This was a due to the large overlap between their members. Four other categories were selected due to the large number of frequently used words that can be contained within them. They are Goodness, Frequency, Temperature and Level of Membership. This is a large expansion on the number of words that were dealt with during Liu and Mendel's codebook paper, which focussed solely

on the category of size. Once the categories had been determined, the next phase was the population of the categories with fuzzy words.

3.5. Generating a Set of Fuzzy Words

With the categories having been created, the next stage involved populating them with fuzzy words. It was important that a wide range of fuzzy words be collected to create the FAST algorithm which is discussed in later chapters. It was also important that the words that were collected be representative of natural language. Specifically it was important that the words be commonly used by English speakers. Through making sure the categories were populated by commonly used words, it ensured that they could be used in any future natural language processing system that required them. Therefore, an experiment needed to be done to determine what words would be contained in the categories.

With their codebook, Liu and Mendel demonstrated the quantification of words, but did not specify why that specific set of words was chosen. This is a problem given the importance of collecting natural language words. As was discussed by O'Shea et al. (2008) if words were arbitrarily chosen there is a risk of selective bias in terms of the person who determines the words. This has 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 (Mendel 2007a) (Liu and Mendel 2008). For these reasons, it was not sufficient to simply expand on Mendel's codebook and populate the other categories based on an individual's opinions. Therefore, whatever the experiment to populate the sets of categories, it needed to incorporate the opinions of a wide range of people.

In terms of creating datasets, a great deal of work was done by James O' Shea in the creation of the STSS-65 benchmark sentence similarity dataset where methods for collecting data were presented (O'Shea et al. 2008a) (O'Shea et al. 2008b). What he created was a robust methodology to return

results from human participants in an unbiased manner. The results of this dataset were used in testing the STASIS algorithm (as was discussed in Chapter 2) and comparing it to the LSA system (This was later expanded to create the STSS-131 dataset that was presented in James O'Shea's doctoral thesis (O'Shea 2010)). The dataset's methodology detailed how sentences could be generated from groups of people. Adapting the methods that he used allows for an experiment to generate a list of words for each category. This experiment involved asking a group of twenty native English speakers to return questionnaires (Appendix 1) that asked them to write down as many words as they could think of from the different categories. The reason that native English speakers were selected was to remove the risk of a participant having a hugely different notion of the meaning of a word based on English being a second language. Being a native English speaker was the sole criterion and participants covered a wide range of demographics. As an illustration of the experiment, on the category of "temperature", participants were asked to write down all the adjectives that they could think of that related to levels of temperature. To ensure that that there was a wide range of words with different values across the categories, a series of guide words were used across each category (for example, with the size/distance category, the guide words were 'near', 'far', 'tiny, 'small', 'medium' and 'large'). Guide words played an important role in the creation of the benchmark dataset (O'Shea et al. 2008a). When considering which guide words would be used, it was important to factor in two things. Firstly it was important that the guide words not be selected in such a way so as to bias the results. Secondly it was important that the guide words serve their intended purpose and not mislead participants. Therefore, careful selection was applied when the words were selected. Once the questionnaire had been completed it was distributed. After the experiment was conducted, it returned a large number of different words that could be used in the creation of a quantification experiment.

Once the words were collected, it presented an opportunity to get an approximation of the impact of fuzzy words on the English language. Specifically an estimation of the frequency with which these words occurred

could be determined. This could provide further justification regarding the importance of the project. 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 (Francis 1965). The Brown Corpus is 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. It has been widely used in a number of different areas (Brill. 1995) (Charniak 2000). Critically, it is also the corpus that the STASIS similarity uses in determining semantic similarity (specifically in terms of finding corpus statistics). Therefore, it can serve as a useful indicator of natural language usage. Looking at the presence of the limited set of words in the Brown Corpus, it was determined that they represented 1.6 percent of all the words within the corpus. Then looking at the presence of the words within sentences within the corpus it was determined that 24% of all the sentences in the corpus contained at least one of the fuzzy words. 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. Therefore this stands as further evidence of the importance of integrating fuzzy words into a text similarity algorithm.

3.6. Quantifying The Fuzzy Words in the Categories

Once the words had been collected in the various categories, the remaining objective was to determine crisp quantities to represent each of the words on a given scale. Doing so would allow the relationships between them to be determined based on their relative values on that scale. Therefore, a new experiment had to be designed that would allow for the accurate quantification of these words. As has been discussed, Liu and Mendel provided an illustration of how fuzzy words could be scaled against each other in their codebook paper (2008), using methods that were described by Zadeh (1999) and Mendel (2007a). As was discussed in the literature review, Mendel's codebook worked on the premise that different people have different perceptions about the meanings of different words and as such these words would have to be represented in a type-2 fuzzy set (allowing for the representation of the uncertainty of the boundaries). At that

point, the centroid based approach could be used to type-reduce the fuzzy set to a type-1 fuzzy set and from there defuzzify it to return a crisp quantity. The concept of defuzzifying a fuzzy set with a set of different people perceptions around a word forms the basis of the experiment to quantify them. It could be considered an option to integrate the values determined by Liu and Mendel in their codebook into the wider set of values. However, in their paper they argued against this and stated that the codebook could not be used as a general purpose set. Furthermore, the results from Liu and Mendel's experiment came exclusively from employees of the Jet Propulsion Lab, and as such was not necessarily a representative population sample.

The main difference in terms of how this methodology differs from Liu and Mendel's codebook methodology is that while Mendel asked for a range of values from each participant in his experiment for each word, in this experiment only a single value is being collected from each participant. In their methodology, when type reducing a fuzzy type-2 set to a type-1 set, Liu and Mendel took the centroids of each element of the type-2 set to form the type-1 set. In this experiment, participants were instead asked to provide a single value that is representative of the point where the membership function of that word would be highest (Appendix 2). It is through taking these values instead of a set of centroids that a type reduced fuzzy set is created for the different words in this experiment. As with the codebook experiment, the standard deviation of these points reflects the level of uncertainty. The reason for this difference is that given that there are a large number of words being covered (because of the number of categories), asking individuals for separate ranges for each word could prove too onerous a task for them to complete successfully. It also reduces the steps needed to deal with potential problems arising from cases where an individual's centroid or a range is not characteristic of the point that they consider having the highest membership function. Comparing the results after the experiment with common words from Liu and Mendel's codebook would provide an example of whether or not this method would provide significantly different results (something that could prove problematic).

When looking at the words that had been returned for the experiment to collect them, it was noted that there were too many to be used in a quantification experiment, within the planned scale. This was because using them all in the questionnaire based approach may have overwhelmed respondents and corrupted the results (from such problems as participants losing interest). Therefore, to reduce the overall number of words that would be used, only words that had appeared in more than one response were maintained. This ensured that while there were a sufficient number of words in each category, none of the categories was excessive in size (specifically none of the categories contained significantly more words than Liu and Mendel's codebook). This also ensured that each category was comparable in scale to the "size" category from Liu and Mendel's benchmark experiment (though given the number of categories the total number of words was substantially higher). Furthermore, this method of sorting the words ensured that the words that remained were the ones that were more frequently used in natural language. Once the sizes of the categories had been narrowed down, the remaining words now allowed experimentation into their quantification and scaling through use of human participants and the application of fuzzy concepts.

There were two important previous papers in terms of word similarity that provided a good framework for a methodology to acquire numeric values for the words from human participants. Aside from Mendel's methodology, a methodology for acquiring levels of similarities between sets of words was put forward by Rubenstein and Goodenough in their seminal paper (Rubenstein and Goodenough 1965) (as was discussed in the literature review), where a dataset of word pair similarities was developed. The Rubenstein and Goodenough set contained 65 sets of word pairs from which human similarity ratings were collected. This dataset has been used in a number of areas (Budanitsky and Hirst 2001) (Budanitsky and Hirst 2006). Furthermore, it was used as a benchmark for the word similarity measure that STASIS was based on. This methodology involved a group of undergraduate students comparing a set of words on a scale of 0 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. Another important and widely used dataset was developed at Technion) and called the Wordsim-353 dataset (Finkelstein et al. 2001. The Wordsim-353 dataset covered a far wider number of word pairs than the Rubenstein and Goodenough dataset and used a 0 to 10 scale to determine similarity, as opposed to a 0 to 4 scale. The Wordsim-353 dataset used 353 different word pairs giving a much larger number of word similarities (though it should be noted that the Wordsim-353 set also encompassed recalculated similarity ratings for the Rubenstein and Goodenough word pairs). An important point to note about all the existing datasets however 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. This is an issue that was addressed in this experiment (in terms of specific types of words) to ensure that the words were representative of natural language. The datasets showed how scales going from low to high can be used to represent levels of association. It was decided that the questionnaire would ask respondents to rate words in each category on a scale of 0 to 10 (in keeping with the Wordsim-353 and codebook methods). The words would be rated based on their levels of association with the highest point in that category (for example, in the temperature scale words would be rated best on their level of association with the maximum possible temperature).

To acquire values for the words in the categories, a questionnaire was created that asked respondents 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 with an option to add a decimal if required (Appendix 2). This represented what the participants considered to be a point where the membership function was highest. It was these results that would be taken in lieu of centroids to form a fuzzy set. The next issue that needed to be addressed was the size of the test group that would quantify the different words. This was determined by looking at the sizes of the different groups that were used to build the other datasets. Acknowledging the numbers used in the codebook and the Wordsim-353 datasets which are more relevant to

this experiment, it was decided that the total number of participants be 20. It was also important that the people who were tested gave answers representative of natural language. To do this, the test group was restricted to native English speakers. This is because, in cases where individuals had English as a second language, different words might have had largely different meanings from a native speaker and could risk distorting the results. This was also a risk in terms of regional dialects. However it was decided that the latter was not a large enough factor to be considered. Of the twenty questionnaires that were sent out, two were spoilt and as such had to be discarded. This left a total of eighteen completed questionnaires. This was sufficient to obtain the desired results from.

The union of these results, per word, creates a fuzzy set. This set can then be defuzzified to create a single value to be used that is representative of that word. To defuzzify the results the mean average of each of the sets will be used. This shall return, for every result, a single crisp-value. The usefulness of this value is then determined by looking at the standard deviation of the members of the set. If a low level of standard deviation exists, the implication is that there is a tendency towards that value containing the highest membership function. If on the other hand, the standard deviation is high, the implication would be that there is no such tendency and taking the centroid of a range would have been better for that word and that other defuzzification methods would need to be considered..

3.7. Results

From the experiment and the subsequent defuzzification, the following results were returned, for each of the categories.

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

Table 1 Size/Distance Category

Word	**Defuzzified Value**	**Standard Deviation**
Arctic	1.06	2.13

Baking	8.17	1.10
Biting	2.11	0.90
Bitter	2.11	1.08
Body-temperature	5.00	0.59
Boiling	8.72	0.83
Brisk	3.28	1.27
Burning	8.67	0.91
Chilly	2.78	1.22
Cold	2.67	1.03
Cool	3.17	1.04
Freezing	1.17	0.99
Frigid	1.50	1.04
Frosty	1.67	1.03
Frozen	1.28	1.07
Hot	7.67	0.91
Icy	1.44	0.70
Lukewarm	4.89	0.83
Mild	4.44	0.86
Nippy	3.06	0.73
Roasting	8.39	1.09
Scalding	9.39	0.78

Scorching	9.00	0.77
Spicy	7.18	2.24
Steaming	7.94	1.11
Sub-zero	1.11	1.68
Sweaty	6.78	0.81
Sweltering	7.89	1.23
Temperate	5.00	0.35
Tepid	4.50	0.99
Warm	5.22	1.31

Table 2 Temperature

Word	**Defuzzified Value**	**Standard Deviation**
Acceptable	4.83	0.86
Alright	5.06	0.87
Amazing	8.17	0.86
Appalling	1.50	0.86
Average	5.00	0.35
Awful	2.39	0.98
Bad	2.17	1.34
Boring	3.24	1.25
Brilliant	7.83	1.95

Dire	2.33	1.88
Dreadful	1.50	0.79
Enjoyable	6.78	1.99
Excellent	8.56	1.10
Fair	5.39	0.85
Fantastic	8.28	1.27
Fine	6.22	1.26
Good	6.56	0.86
Great	7.56	0.86
Inadequate	3.22	1.11
Marvellous	8.06	1.80
Mediocre	4.72	1.67
Middling	4.89	0.32
Nice	5.67	0.84
Ok	5.22	0.73
Passable	4.72	0.75
Pathetic	1.83	0.99
Pleasant	6.00	0.84
Poor	2.56	0.70
Rotten	1.11	0.68
Splendid	8.22	0.73
Superb	9.00	0.97

Terrible	1.22	0.73
Unacceptable	1.17	1.20
Unbearable	0.44	0.62
Unsatisfactory	1.78	1.17
Useless	1.11	1.02
Wonderful	8.78	0.65

Table 3 Goodness Category

Word	**Defuzzified Value**	**Standard Deviation**
Adolescent	4.00	0.97
Adult	6.50	1.10
Aged	8.33	0.59
Ancient	9.83	0.38
Antiquated	9.11	1.08
Antique	9.39	0.61
Archaic	8.11	3.38
Baby	0.83	0.62
Babyish	1.39	1.04
Child	2.50	0.71
Childish	3.00	1.14
Child-like	3.39	1.24

Decrepit	6.28	2.82
Elderly	8.28	0.83
Experienced	6.78	0.88
Full-grown	6.17	1.29
Grown-up	6.33	1.33
Immature	3.94	1.43
Infantile	2.61	0.92
Juvenile	3.61	0.92
Mature	7.06	0.94
Middle-aged	6.06	1.43
New	1.00	0.59
Old	8.22	1.00
Pensionable	8.28	0.83
Pre-historic	6.50	4.74
Pre-pubescent	3.67	0.97
Recent	2.56	1.42
Young	3.11	0.68
Youthful	3.83	0.92

Table 4 Age Category

Word	Defuzzified Value	Standard Deviation
Always	8.89	2.14
Barely	1.67	0.69
Commonly	6.39	1.61
Consistently	7.61	1.46
Constantly	8.33	2.38
Daily	6.89	1.91
Frequently	6.89	1.49
Habitually	6.17	1.47
Hardly	2.17	0.86
Infrequently	2.50	0.99
Never	0.06	0.24
Normally	5.72	1.07
Occasionally	3.89	1.23
Often*	6.61	1.04
On-Occasion	3.89	1.37
Periodically	4.28	1.41
Rarely	1.89	0.90
Regularly	6.17	1.34
Repeatedly	7.17	1.47

Scarcely	1.72	0.83
Seldom	1.94	1.35
Somewhat	3.83	0.92
Uncommonly	2.83	0.79
Unpredictably	3.65	1.69
Usually	7.06	1.70

Table 5 Frequency Category

Word	**Defuzzified Value**	**Standard Deviation**
Adequate	6.12	1.54
Almost	8.22	1.11
Average	5.33	0.77
Barely	2.33	1.37
Bit	2.44	0.92
Generally	6.00	1.28
Greatly	8.06	0.73
Halfway	4.83	0.71
Hardly	2.67	0.59
Just	6.33	2.54
Largely	8.11	0.83
Little	2.33	0.84

Mainly	7.06	1.39
Middling	5.11	0.47
Mostly	7.50	0.99
Partially	4.33	1.18
Rather	6.76	1.86
Scarcely	2.11	1.18
Scraping	2.53	2.07
Somewhat	5.18	1.24
Sufficient	6.76	1.71

Table 6 Level of Membership

The results present a crisp defuzzified value for each word in each of the categories. It is important to assess the value of the results that were collected. Through a review of the standard deviations of the values within the sets that the words were derived from, it can be observed that, in a vast majority of cases, the standard deviation was less than 2.00. This was true for each of the different categories. Looking at the words from within the size/distance category and comparing them with the common words with those collected by Mendel (*Table 7*).

Word	**Our Method**	**Codebook**
Tiny	1.72	0.635
Little	3.17	2.13
Small	3	2.315
Medium	4.67	5.19
Sizeable	7.11	7.155

Large	7.17	8.125
Substantial	7.33	7.9
Huge	8.39	9.34

Table 7 Comparison of common words with codebook

It can be observed that there is a very high correlation (0.99) between the results collected using this method and the means of the centroid collected by Mendel, there is also a very small average standard deviation of 0.51. A T-Test was conducted to test the hypothesis

H1: There is a difference between the values returned by our method and Mendel's

With the resultant null hypothesis

H0: There is no significant difference between the values returned by our method and Mendel's

This returns a p-value of 0.98, strongly suggesting that there is no significant difference between the results returned by the two methods. This indicates that the method used to determine the points of highest membership was successful. As such it can give a good representation of the results that would have to be determined through using Mendel's centroid-based approach. This means that in cases where fuzzy words are quantified on a scale, it is, based on the results from this experiment, sufficient to ask for the single point with the highest membership function rather than collecting ranges for each participant. The results show that there is unlikely to be much difference between what the points with the highest membership functions and centroids are. Significantly, these results show that humans tend towards particular values as the points of highest membership for various words. This implies that this method can be used to acquire greater numbers of fuzzy words and as such, expand the existing categories that way as well as populate wholly new categories through further work with human participants.

3.8. Conclusions

These results give us a series of words across a number of categories that have been scaled against each other on individual scales pertinent to each category. This is important to note as the scaling is solely restricted at present to being within the categories and as of yet the words are not scaled between the categories. Developing a method for doing this is a potential area of future work. The accuracy of this can be demonstrated by the low standard deviations. It shows that we can take these values as representative of human perceptions. This also shows that, with groups of humans, perceptions regarding fuzzy words tend towards certain values. As the words are scaled against each other, a clear picture can then be acquired regarding their relationships with one another. Significantly then, these results could be used in the creation of a new fuzzy ontology. Given that it is now possible to numerically represent these words in terms of each other, an ontology structure could be created to map these relations. Through doing this, these relations can be integrated into an ontological sentence similarity measure. Through mapping the relations between the words in such a structure like WordNet does, a word similarity measure could be created that can represent the levels of similarity between them in the same manner as the word similarity component of STASIS. This would effectively allow a sentence similarity measure to accurately determine the effect of fuzzy words on the overall level of a sentence's similarity. Therefore, the next stage of the project was to develop such and ontological structure based on the information that was collected during this experiment and then use that structure to implement a new fuzzy word similarity measure (that would then be used in a new sentence similarity measure). The work done in this chapter also has further uses. What is now presented is a set of results, as well as a methodology to collect further results in any future work based around the quantification of fuzzy words. This can include the expansion of the existing categories or the creation of further new categories, with wholly different sets of words.

4. A Methodology for building FAST

4.1. Chapter Overview

This chapter describes the new algorithm called FAST (Fuzzy Algorithm for Similarity Testing). The purpose of this algorithm is to take two sentences as input and return a similarity value for them. The difference between FAST and existing semantic similarity measures is that FAST can show the effect that fuzzy words have on the overall level of similarity between pairs of sentences. An important source of inspiration FAST is the STASIS measure which is an existing and well recognized similarity measure that has been discussed at length in the literature review. A brief overview of STASIS will be provided in Section 4.3, which describes the importance of Word Similarity in the context of Sentence Similarity. The first step in the development of FAST was to use the words that had been quantified in Chapter 3 to create a fuzzy ontology (described in Section 4.6) for each category of words. These ontologies could be used to determine the relations between words within the same category. Adapting the STASIS formula to these relationships delivers a similarity value for pairs of fuzzy words. Furthermore, the effect that fuzzy words have on non fuzzy words can be determined through the relationships between fuzzy words (a separate algorithm has been created that determines what these associated words are which is discussed in Section 4.7).This chapter describes the methodology used to build FAST and the development its main components. This includes the creation of fuzzy ontologies and an ontology-based fuzzy word similarity measure; the development of an algorithm that determines the association of non fuzzy words with fuzzy words and a method to determine the effect of fuzzy words on non fuzzy words.

4.2. Chapter Aims

- Describe the creation of ontologies of fuzzy words for the categories
- Describe the development of a fuzzy word similarity measure
- Describe the implementation of the FAST sentence similarity measure

4.3. Relevance of Word Similarity to Sentence Similarity

The STASIS (Li et al. 2006) algorithm was discussed in detail in Chapter 2 (The Background Chapter). The algorithm takes two sets of texts as input and returns a total level of similarity based on semantic and syntactic values. The semantic value is calculated first through calculating the levels of similarity between pairs of words in the two sentences through their relations in the WordNet ontology and then through statistics based on Corpus Statistics. The measure returns a final overall level of similarity between the two sets of text. This illustrates the key importance of an integrated Word Similarity measure to the ability of the STASIS measure to return levels of sentence similarity. Therefore, creating a word similarity component is vital in the creation of an ontological sentence similarity measure.

4.4. Evaluation of an existing Word Similarity Measure

The creation of the Word Similarity Measure that had been implemented in STASIS (Li et al. 2006) has been discussed in detail, in the Literature Review (Chapter 2). However before proceeding it was important to review the effectiveness of the word similarity utilized by STASIS against previous datasets. This is to give a clear impression of the usefulness in extending the ontology-based approach that was taken into determining the level of similarity between fuzzy words. It is also important to review as a factor when evaluating the FAST similarity measure against human ratings.

The word similarity measure was first evaluated in by Li et al. (2003), where it was tested against human similarity ratings. The ratings in question were based on the sets of words that had been collected and quantified by Rubenstein and Goodenough (Rubenstein and Goodenough 1965) and later by Miller and Charles (Miller and Charles 1991). This evaluation involved running the word pairs through the word measure (calculating the total ontological distances between words in WordNet and the depth of their lowest common subsumer) and then running those values through the formula (2) and taking the correlation of these results with the results returned from the Rubenstein and Goodenough datasets (different α and β values were tested to determine which ones offered the best correlation).

After the words were run through the dataset, the α and β values that were determined for formula (2) were 0.2 and 0.6 respectively (This was later changed in by Li et al. (2006) to 0.45 due to a higher correlation being produced). The results showed that the correlation between the words was 0.9015, showing a high level of correlation. This would on the surface demonstrate a good argument for usage of that formula. However looking at the results themselves raises some concerns. Specifically it shows that the values that were entered into the formula do not present the values that the paper states that the formula returned. Furthermore the values that the paper presents as output would present a correlation of 0.88 instead. *Table 8* shows a comparison between the results from the paper against the actual results.

Word 1	**Word 2**	**Similarity Rating**	**Original Value**	**Actual Value**
Cord	smile	0.02	0	0
Rooster	voyage	0.04	0	0
noon	string	0.04	0	0
Glass	magician	0.44	0	0
Monk	slave	0.57	0.355	0.375
Coast	forest	0.85	0.17	0.162
Monk	oracle	0.91	0.168	0.206
Lad	wizard	0.99	0.355	0.375
Forest	graveyard	1	0.132	0.132
Food	rooster	1.09	0	0
Coast	hill	1.26	0.366	0.425
Car	journey	1.55	0	0
Crane	implement	2.37	0.366	0.425
brother	lad	2.41	0.355	0.375
Bird	crane	2.63	0.472	0.546
Bird	cock	2.63	0.779	0.815
Food	fruit	2.69	0.17	0.425
brother	monk	2.74	0.779	0.815

Asylum	madhouse	3.04	0.779	0.818
furnace	stove	3.11	0.585	0.559
magician	wizard	3.21	0.999	0.984
journey	voyage	3.58	0.779	0.815
Coast	shore	3.6	0.778	0.805
implement	tool	3.66	0.778	0.805
Boy	lad	3.82	0.778	0.805
automobile	car	3.92	1	1
midday	noon	3.6	1	1
gem	jewel	3.94	1	0.999

Table 8 Results in Paper Compared to Actual Results

This instead shows a level of correlation of 0.91. This level of correlation is slightly higher than the value that the paper initially put forward despite the errors in the formula and higher than the value the erroneous values would have presented. The paper in fact, therefore, underestimated the strength of the algorithm. Whatever the cause of the error the issue now arises is the validity of the formula as a whole, given the risk of other errors in collecting the data. Therefore, the Word similarity measure (Li et al. 2006) needed to be reassessed. This is further necessitated by the fact that WordNet has evolved since the original implementation of the formula and as such its efficacy may have changed (and the formula itself has been improved). It also allowed for the formula to be tested on an additional set of similarity ratings taken from Rubenstein and Goodenough that were not covered in the original paper.

Through running the words with a new implementation of the word similarity measure (this time using the current version of WordNet), gave the following results. As shown in Tables 9 and 10.

Word 1	Word 2	RG Similarity	Algorithm Similarity
Cord	smile	0.02	0.097
Rooster	voyage	0.04	0.097
noon	string	0.04	0.097
Glass	magician	0.44	0.216
Monk	slave	0.57	0.445
Coast	forest	0.85	0.445
Monk	oracle	0.91	0.445
Lad	wizard	0.99	0.445
Forest	graveyard	1	0.445
Food	rooster	1.09	0.445
Coast	hill	1.26	0.445
Car	journey	1.55	0.445
Crane	implement	2.37	0.445
brother	lad	2.41	0.445
Bird	crane	2.63	0.549
Bird	cock	2.63	0.819
Food	fruit	2.69	0.819
brother	monk	2.74	0.819
Asylum	madhouse	3.04	0.819
furnace	stove	3.11	0.819
magician	wizard	3.21	0.991
journey	voyage	3.58	0.991
Coast	shore	3.6	0.991
implement	tool	3.66	0.991
Boy	lad	3.82	0.991
automobile	car	3.92	1
midday	noon	3.6	1
gem	jewel	3.94	1

Table 9 Word Similarity Measure Applied to First Set of Rubenstein and Goodenough Ratings

Word1	Word 2	RG Similarity	Algorithm Similarity
Fruit	furnace	0.05	0.197
autograph	shore	0.06	0.197
automobile	wizard	0.11	0.197
Mound	stove	0.14	0.295
Grin	implement	0.18	0.295
Asylum	fruit	0.19	0.295
Asylum	monk	0.39	0.295
graveyard	madhouse	0.42	0.295
Boy	rooster	0.44	0.295
cushion	jewel	0.45	0.295
Asylum	cemetery	0.79	0.295
Grin	lad	0.88	0.295
Shore	woodland	0.9	0.393
Boy	sage	0.96	0.393
automobile	cushion	0.97	0.393
Mound	shore	0.97	0.425
cemetery	woodland	1.18	0.425
Shore	voyage	1.22	0.425
Bird	woodland	1.24	0.425
furnace	implement	1.37	0.425
Crane	rooster	1.41	0.425
Hill	woodland	1.48	0.425
cemetery	mound	1.69	0.425
Glass	jewel	1.78	0.425
magician	oracle	1.82	0.425
Sage	wizard	2.46	0.425
Oracle	sage	2.61	0.425
Hill	mound	3.29	1
Cord	string	3.41	1
Glass	tumbler	3.45	1

Grin	smile	3.46	1
Serf	slave	3.46	1
autograph	signature	3.59	1
Forest	woodland	3.65	1
cock	rooster	3.68	1
cushion	pillow	3.84	1
cemetery	graveyard	3.88	1

Table 10 Word Similarity Measure Applied to Second Set of Rubenstein and Goodenough Ratings

From the results shown in *Table 10*, the words in dataset of non fuzzy words have correlations of 0.948 and 0.950 against the human test results. This shows that despite the erroneous results from Li et al. (2003) the formula still produces results that are good enough for it to be used in conjunction with the fuzzy word similarity measure. Therefore, the STASIS word similarity formula and WordNet relations shall be used to determine the relationships between non fuzzy words. Furthermore, the success of the formula shows that applying the formula to the newly developed fuzzy ontologies is not problematic.

4.5. FAST

This section provides an overview of the FAST algorithm. The psuedocode for FAST can be found in *Figure 3*.

1. *Let T1 and T2 be two sentences*
2. *Tokenize every word in the T1 and T2*
3. *Pair every combination of Tokenized words*
4. *For every word pair (A,B):*
5. *If A and B are both fuzzy words:*
6. *If A and B are in the same category:*
    i. *Reference Subsumer Depth from Fuzzy ontology*
    ii. *Reference Length between words from Fuzzy ontology (described in section 4.6)*

b. *Using these values, calculate Level of Similarity with formula (1), the STASIS word similarity formula*
c. *Return Level of similarity (on a scale of 0 to 10)*

7. *Else:*
   a. *Apply STASIS word similarity measure (Li et al. 2003)*
8. *End If*
9. *Return Level of similarity*
10. *Else*
11. *Apply STASIS word similarity measure.*
12. *Determine presence of fuzzy words associated with the non fuzzy words (described in section 4.7).*
13. *If Associated Fuzzy Words are Present:*
14. *Calculate Subsumer Depth and length modifications using the process (described in section 4.7).*
15. *Recalculate Word Similarity*
16. *Return Level of Similarity*
17. *Else:*
18. *Return level of similarity.*
19. *End If*
20. *End If*
21. *Apply Corpus statistics (O'Shea 2010)*
22. *Next*
23. *Determine Syntactic similarity (O'Shea 2010)*
24. *Determine Total similarity using formula (2)*

Figure 4. Pseudocode for FAST Algorithm

4.5.1. Overview of FAST

This pseudocode in *Figure 4* describes how the algorithm deals with pairs of words and from that, how overall sentence similarity is calculated. Once the similarities for all the words have been calculated, FAST uses the same method as STASIS to determine overall similarity. For every pair of words, the FAST algorithm determines if they are fuzzy or not (based on their presence in any of the categories). If they are fuzzy but do not belong to the same category the WordNet based method that STASIS uses determines

their level of similarity If they are present in the same category, then the algorithm calculates their level of similarity based on their subsumer depth and distance from each other in that category's ontology using the formula presented by Tsatsaronis et al. (2009). Once the similarities for all the pairs of words is calculated (given the corpus statistics and syntactic similarity are calculated separately, with the methods discussed by Li et al. (2003)) the total level of similarity can be determined using formula (2). The creation of these fuzzy ontologies is described in *Section 4.6*. The similarity between non fuzzy words is calculated with the existing STASIS word similarity measure. If these words have associated fuzzy words, then their level of similarity is amended using the ontological relations between fuzzy words (as is discussed in Section *4.6).* What follows is a detailed description of all the various steps of the algorithm, explaining how they contribute towards calculating the overall level of sentence similarity. The initial step is the input of two sentences. While the system is able to accept sentences of any length it is specifically designed to deal with short sentences (sentences containing 35 words or less). Therefore sentences should ideally be restricted to that length to ensure the maximum effectiveness of the system.

Line 1: Tokenize every word in the two sentences.

The sentence similarity measure works through applying a word similarity measure to every possible pair of words. Therefore, before the sentences can be processed by the algorithm, the individual words within them must first be identified and separated from each other. As with the rest of the measure, the system for doing this is implemented in python. It uses the Natural Language Toolkit (NLTK) (Bird 2006), a powerful set of python libraries that have a variety of different functions in the area of Natural Language Processing (NLP). Its practical applications can be seen by the work done by MacMahon et al. (2006) and Eisele and Chen (2010). Using this a sentence entered as a string ("The cat is in the big hat") is sorted in a list ["the", "cat", "is", "in", "the", "big", "hat"]

Line 2. **Pair every combination of Tokenized words**

As the word similarity measure needs to compare every single possible combination of words within the two sentences a method needed to be generated to allow this to be done easily. Towards this end a “Bag of Words (Li et al. 2006)” (BOW) was used. This was defined as the union between all the words contained within two sets of words A and B. For example if we consider sets A = {“The”, “Big” Car”} and B = {“A”, “Big”, “House”}, the BOW for them would by (A ∪ B) = {“A”, “Big”, “Car”, “House”, “The”}. This allows the words from each set to be easily paired with every word in the sentences sequentially (For example A would first be paired with A, then with big etc.).

Lines 3 to 10

At this point, for each word pair a level of similarity should be determined. In the case of fuzzy words, this is done through the application of fuzzy ontologies (the creation and development of fuzzy ontologies are explored in substantial detail in Section 4.6). If the words in the pair are not fuzzy then the method that is used in the STASIS word similarity measure (Li et al. 2003) (which is discussed in greater detail, in Section 4.5) is applied. This is also the case if the fuzzy words in a sentence are not in the same category or if the pair contains one fuzzy words and one non fuzzy word.

The first objective is, therefore, determining if the words are fuzzy or not and if they are fuzzy, if they belong to the same category. It is important to note that just because a word is contained within a category it does not necessarily imply that it is a fuzzy word. Misidentification in this case could lead to a miscalculation of the level of similarity. For example consider the word “cold”. The word could have a clear fuzzy meaning in terms of reference to temperature, but it also has a widely used non fuzzy meaning in terms of reference to illness. A method that could be used to handle disambiguation involves considering the word type. Given that the words in the sentences have been previously tagged according to type and that the vast majority of fuzzy words are adjectives or adverbs (Zadeh 1996) (Zadeh 1999) (Mendel 2007b); simply ensuring that only those types of words are

considered reduces the problem. This can easily be done, as the word types have already been determined, during the tagging process. It does not solve the problem entirely, however, as there could be rare instances where these words are not adjectives. There is, however, an additional problem with disambiguation, words belonging to more than one category, but having different meanings within each one (the exception being average that has a universal meaning throughout them all). This problem occurs if both the words in a word pair belong to more than one category. In that case, two similarities need to be calculated for the words, one for each category. After that is done, the higher level of similarity between the two is taken to be the similarity value. This is in keeping with the general assumption of higher similarity that already exists in semantic similarity measures.

Assuming the two fuzzy words belong to the same category, information about their relationship can be derived from their relevant fuzzy ontology. This is done through looking at the distances between the entities that contain the words and the distance from their lower common subsumer to the top of the hierarchy (subsumer depth). Through running both those distances into a formula, a single value can be derived. This value stands as the similarity between those two words based on their ontological relations. The specific formula that is applied is (1). Given that, in the case of non fuzzy words, the measure reverts to the STASIS measure, this was needed to ensure uniformity. This technique was developed and applied primarily in the STASIS word similarity measure (Li et al. 2006) (O'Shea et al. 2008b) which is heavily discussed in the literature review as well as in Section 4.5. Given how the ontology is build and structured, referencing the distances can be easily done. For an illustration of this, consider the words "tiny" and "big" which are both in the size category. Tiny would be associated with the "Very Small" entity while "big" would be associated with the "large" entity. For an illustration consider *Figure 5*.

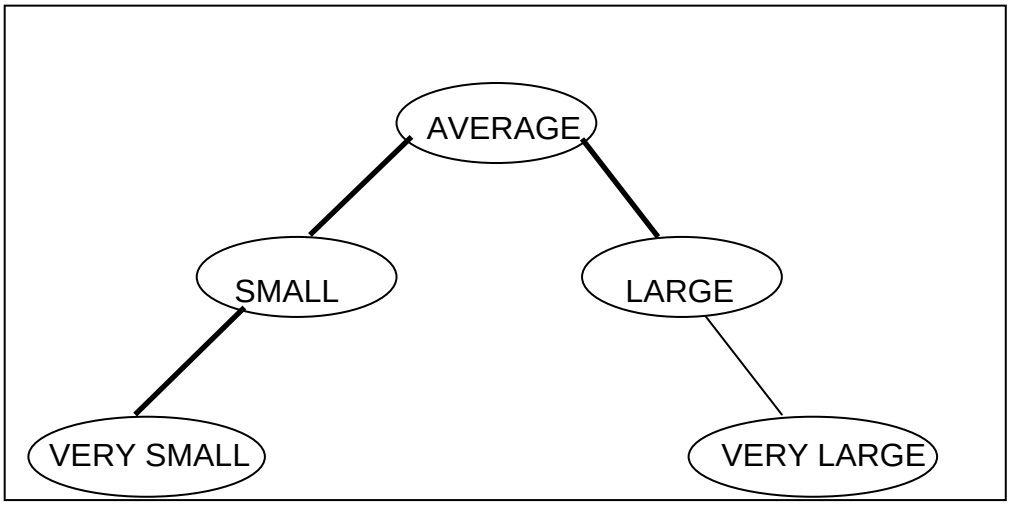

Figure 5 Nodes in Size/Distance Ontology Structure

The distance between the “Very Small” and the “Large” can be seen as 3 while the subsumer depth can be seen as being 1 (with 1 taken as the minimum possible subsumer depth). These are principally similar to the values that the STASIS similarity measure uses when calculating relationships between words in the WordNet ontology.

There is however another adjustment that must be made. As is discussed in *Section 4.6* with an illustration shown in *Table 12*, each of the individual entities acts as its own scale (for example “big” and “large” are both part of the same overall entity but nonetheless don’t have the exact same value). Though the differences in value between words contained in an entity are likely to have a less substantial impact on the overall level of similarity than the inter-entity relations, they still must nonetheless be considered. These scales within entities are discussed in detail in *Section 4.6*. What effectively exists is, for each entity, a -1 to 1 scale that all its words are contained within. The reason that a -1 to 1 scale was used was to better represent the fuzziness of the boundaries between the categories. These distances cause adjustments to be made to the distances and subsumer lengths of the words. As an illustration consider the words “tiny” and “microscopic”. They are both positioned on a scale within the entity “very small”. In this particular case, “tiny” would have a value of 0.31 whereas Microscopic would have a value of -1. The midpoint between them would be -0.34, moving that

distance further away from the beginning entity. Therefore if the words were being compared, their distance and subsumer length values would change from the initial, 0 and 3 to 0.34 and 3.34. Applying these changes this allows a far more accurate representation of the relationships between pairs of words.

With values having been acquired for the word pairs, the next stage was to apply the STASIS word similarity formula to them to allow a semantic value for the ontology to be collected. To illustrate how this worked consider two fuzzy words being put through the system, “huge” and “miniscule”. The initial comparison of the words positions in the ontology reveals a total distance of 4 with a subsumer depth of 1. Then looking at their positions on their respective scales, it can be seen that huge has a position of -0.75, bringing it that closer to the beginning entity while miniscule has a position of -0.69 moving it further away from that point. As a result of this, the total distance between the entities changes to 3.97 while the subsumer depth remains unchanged. These values are then input into the formula (the subsumer depth is represented by *h.* The α and β values are 0.2 and 0.45 respectively.

$$S(huge, miniscule) = e^{-3.97\alpha} \cdot \frac{e^{1.00\beta} - e^{-1.00h}}{e^{1.00h} + e^{-1.00h}} \quad (3)$$

This returns a total level of similarity of 0.19, a low level of similarity between the words. Had this formula not been used a similarity level of 0 would have been returned. This would have not reflected the small level of similarity that would have emerged from their mutual association to a single concept.

If there are no fuzzy words or if the fuzzy words are not from the same category, the measure instead uses WordNet as is the case with the original STASIS measure (Li et al. 2006) (this is discussed further in *Section 4.5*). The measure determines a level of similarity for every possible combination of definitions between the two words and returns the highest level of similarity that can be obtained. While this method can create some problems as the definition pair with the highest level of similarity may not

necessarily be the intended pair, it is nonetheless an efficient method of comparing nouns as was the demonstrated through the success of STASIS.

After this stage, the measure should have returned a set of similarity ratings for the different word pairs. The values can now be used to contribute towards determining the overall level of similarity between two sentences.

Lines 11 -20

After the non fuzzy word pairs had their levels of similarity calculated, it was important that it be determined if they (the words in the pair) had any associated fuzzy words (this is further explored in Section 4.7). For an example of an associated word consider the phrase “The tall man”, the fuzzy word “tall” in that phrase is associated with the word “man”, a similar effect could be observed if the phrase were reworded “The man is tall”. These words affect the meanings of their non-fuzzy affiliates and, as a result change the levels of similarity between non fuzzy words. It is because of this that they had to be considered in terms of word similarity.

Towards the goal of identifying non fuzzy words that were associated to fuzzy words, an algorithm was applied that was able to do this (as is described in Section 4.7). The algorithm, for each fuzzy word, used the tagged words and word types in a sentence and applied grammatical rules and conventions to determine for each fuzzy word, the non fuzzy word most likely to be associated with it. The result of applying this algorithm was a new set of sentence pairs containing a single fuzzy word and a single non fuzzy word. For example if the sentence “The big cat was on the small carpet” was being used, the measure would first identify the words “big” and “small”. After that it would apply the rules to identify “cat” and “carpet”.

Once the associated words had been identified, the changes that the fuzzy words would make to their overall level of similarity could be determined. This was done through amending both the subsumer depth and the distance between the words, which the fuzzy words cause to change. The process of this is discussed in Section 4.7. Once those changes had been made the

level of semantic similarity was recalculated for the words. This was done through inserting the new distance and subsumer depth values (calculated from applying the effect of the fuzzy words on to the original values) into the Word Similarity algorithm. The new values that are returned overwrite the former ones.

Lines 20 -23

After a set of similarity ratings for all the different word combinations had been acquired, it was time for the STASIS similarity measure to use them to determine the total level of semantic similarity between the words. The creation of the STASIS similarity measure is explored in depth in the literature review (*Chapter 2*), this section instead describes how it works in practice to determine the total level of similarity. STASIS uses three components towards this goal, the ontology-based similarity values (which play the largest role), values that are based on corpus statistics and similarities in syntax (which are affected by levels of semantic similarity).

As the word pair similarities have already been established, the next issue is the effect of corpus statistics on the total level of similarity. These statistics are determined from the Brown Corpus, which is a large corpus of texts (Francis 1965). Through looking at the frequency distributions of words within the corpus information weights based on their probabilities of occurrence can be derived through formula (4).

$$i(w) = 1 - \frac{\log(n+1)}{\log(N-1)} \quad (4)$$

Where *N* is the total number of words in the Corpus and *n* is the words frequency. Once they are calculated the weights are applied to the ontological levels of similarity. This application gives the final semantic values.

At this point the syntactic locations of the words in the sentence need to be considered for their effect on sentence similarity. To achieve this goal the system looks for matching words between sentences and determines their levels of similarity. If there is no matching word, the system then has to find

the location of the word with the closest meaning. This is done though looking at the location of the word with the highest level of semantic similarity (which has been previously established by the system). This does unfortunately constrain the level of accuracy of the syntactic component by the accuracy of the semantic component (if the semantic component for example considers a word pair to have a higher level of similarity than they actually do, they may also be incorrectly syntactically paired. However, it should be noted that the syntactic component has a substantially smaller effect on the overall level of similarity than the semantic component.

With all the parameters having been determined, the total level of similarity can now be calculated through formula (2). What is finally returned is a single value. A higher value indicates a higher level of similarity.

4.6 Creating A Fuzzy Ontology

4.6.1 General Discussion of Ontologies

The first step in determining the level of similarity between pairs of fuzzy words was to create a structure that was able to represent the relationships between them. This would allow the algorithm to use these relationships to determine a similarity for the two words (serving as their level of similarity) by performing calculations based on them. The basis for the relationships between fuzzy words was based on the work done in the previous chapter where a range of fuzzy words had been quantified through human relationships. Therefore, this information needed to be the main reference point from which the relationship structures were built. Therefore an ontological structure would be used to determine the relationships between the words. An ontology is a structure where the entities within it are connected by sets of rules. They were described in detail in Chapter 2. Also described was WordNet (Miller et al. 1990) which was a lexical database that could serve as lexical ontology .This provided the background wherein the relationships between fuzzy words could be represented through generating rules that are based on the differences in quantities that the different words have had allocated to them. The success of a WordNet ontology-based method was demonstrated by Li et al. (2006) where it was

used as the basis for the original word similarity measure that STASIS was based on (Li et al. 2003)

The usefulness of ontologies in the field of sentence semantics made them ideal structures to contain the different fuzzy words that had been quantified. Furthermore, they provide a structure through which the relationships between the words can be represented into a form that could be computer-readable. This would allow them to be easily integrated into an automated word similarity measure. The success of ontology-based similarity measures such as Resnik's work and the word done with STASIS provide a proof of concept of their effectiveness. It also demonstrates the value of WordNet as a structure for a lexical ontology. The next section provides a detailed overview of how fuzzy ontologies were created and how relationships within them were represented.

4.6.2 Building A Fuzzy Ontological Structure

In the previous chapter, six categories of fuzzy words had been created, populated and, through human experimentation, the fuzzy words were quantified on a given scale. The issue at that point then became how to create a structure that could show the inter-relatedness of the words. In the previous sections, ontological structures (with a particular focus on WordNet) were identified as suitable candidates. Therefore, the objective was to create an ontological structure that was able to show the relationships between the fuzzy words in a category. The aim of this section is to describe how these structures will be built and describe the nature of the relationships of the entities contained within. This ontology structure would fill a role akin to the WordNet ontology was used in Resnik's similarity measure in terms of being 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 could be built specifically around determining the level of similarity between pairs of fuzzy words.

When the fuzzy words were collected and quantified, the work done by Zadeh on granularity and the vital role that it played was discussed by Zadeh (1999) and Mendel (2007a) (2007b). As had been expanded on in the

background reading, this was how a group of smaller entities could be associated with a single larger concept (or a granule) and in terms of the concept of fuzzy; different entities could have different levels of membership. This concept fits in very well with the overall nature of ontological structures wherein different entities can be related to a common concept. This is particularly represented in the inheritance based system used in WordNet (that was discussed extensively in the last section), where entities are identified as having a set of characteristics defined by the concept that they inherit from. This serves to strengthen the argument of why ontologies are the most suitable structure to serve as the backbone of the similarity measure. In terms of the incorporation of fuzzy elements in ontologies themselves, substantial work was done (Parry 2004) (Lee and Huang 2005) (Reformat and Ly 2009) (Bobillo and Straccia 2011) .

In creating a fuzzy ontology, the first step was to divide each category into nodes that were related to each other through subsumer relations. With the division of categories in this manner, this allows for sets of words from the categories to be stored within these nodes. This would allow for the relations between these words to be represented by their distances and subsumer depths. It was decided that each category be divided into five nodes with the central subsumer being representative of the area around the midpoint of the range.

The issue, therefore, remained as to how many classes should exist within the domains and whether a greater or smaller number of classes would provide better results. Therefore, two different ontological structures were designed with a different number of classes in each one (called Structure 1 and Structure 2 respectively). Given that sets of fuzzy words had been quantified on various scales in Chapter 3, the creation of classes based on areas of the scale would allow them to be easily populated with fuzzy words. It was decided that each domain in Structure 1 and Structure 2 would contain five and ten classes respectively. The nature of Structure 1 ensured that each class would contain relatively equal number of fuzzy words but contained a risk that some of the nuances in the different quantities between the fuzzy words would be lost (a method for dealing with this is presented later in this chapter). The nature of Structure 2 on the other hand, created a

risk that there would be empty classes but ensured (by the fact classes covered smaller ranges of values) that the fuzzy words contained within each class was close to each other in terms of the quantities they represented. What is common to both these structures is that their form is in keeping with Mendel and Zadeh's work of objects or classes of objects being through some form of relationship, subsumed by other classes of objects. For Structure 1, for each of the domains the following classes were created.

$$Size = \{Very\ Small, Small, Average, Large, Very\ Large\}$$

$$Goodness = \{Very\ Bad, Bad, Average, Good, Very\ Good\}$$

$$Age = \{Very\ Young, Young, Average, Old, Very\ Old\}$$

$$Temperature = \{Very\ Cold, Cold, Average, Hot, Very\ Hot\}$$

$$Frequency = \{Very\ Often, Often, Average, Rarely, Very\ Rarely\}$$

$$Membership = \{Nearly\ Empty, Hardly, Average, Mostly, Almost\ Full\}$$

For Structure 2, the following structure was applied to all the domains (It shows 5 evenly spaced classes with values below a centre point and 5 evenly spaced classes with values above the centre point.

$$Domain = \{Neg\ 5, Neg\ 4, Neg\ 3, Neg\ 2, Neg1\ Centre, Pos\ 1, Pos\ 2\ Pos\ 3, Pos\ 4, Pos5$$

Structure 2 is divided into "Neg" classes that contain words with values progressively lower on the scale than 0 and "Pos" classes that contain words with values progressively greater than 0 with the centre point representing a single point on the scale where the value of 0 would be taken. For example for the size domain, the class Neg2 contains (among others) the words "minute" and "tiny" while Neg1 contains the words "Microscopic" and "Miniscule" which have smaller values. Both of these structures needed to be implemented and assessed independently to see which one delivered better results. This is done through determining which one enables better performance of the FAST algorithm. This evaluation is further discussed in Chapters 5 and 6.

With the creation of the detailed domains (categories), that now contained classes, the next step of the methodology required determining the relationships between the classes. Given the nature of the classes that are being considered here, it was apparent that standard subsumer relations (ISA/HASA Relations) could not be used to map the relations between the classes. This is because of the nature of the classes occupying areas on a scale as opposed to one of them being a type or a property of another. Instead the relations that were used needed to reflect their differences in scale. Therefore another novel approach was required to represent their relationships. What is proposed is instead a “Surpasses” relationship. This represents a class surpassing all of the criteria that would be needed to be a member of a surpassed class. For example, consider an evaluation of an exam paper. Whether a candidate performs better or worse than average (being classified as good or bad) their performance would exceed the requirements for being average. Then if the grade were very good then not only would it surpass average but also surpass good along that value. Similarly on another value, bad would also surpass average and be surpassed by very bad. Through the implementation of these relations the ontologies can provide a clear picture therefore of the differences and similarities between the classes in terms of the ontological distances between them and their subsumer depths.

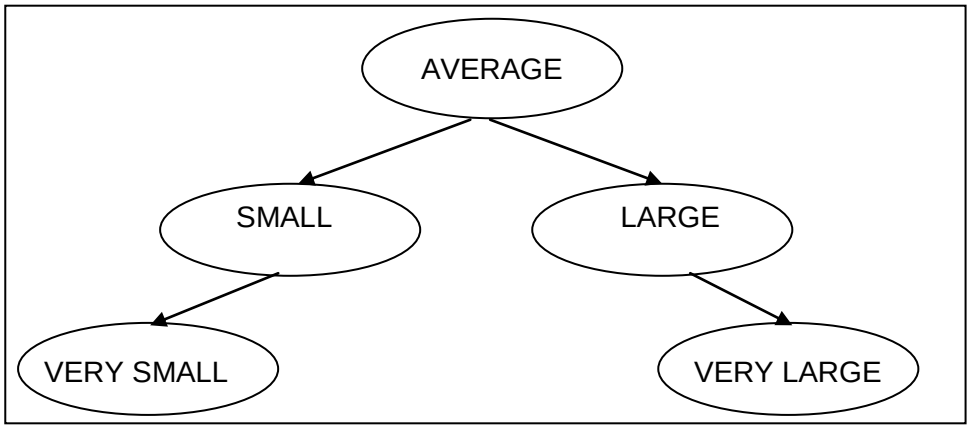

Figure 6 Size/Distance Category (Structure 1)

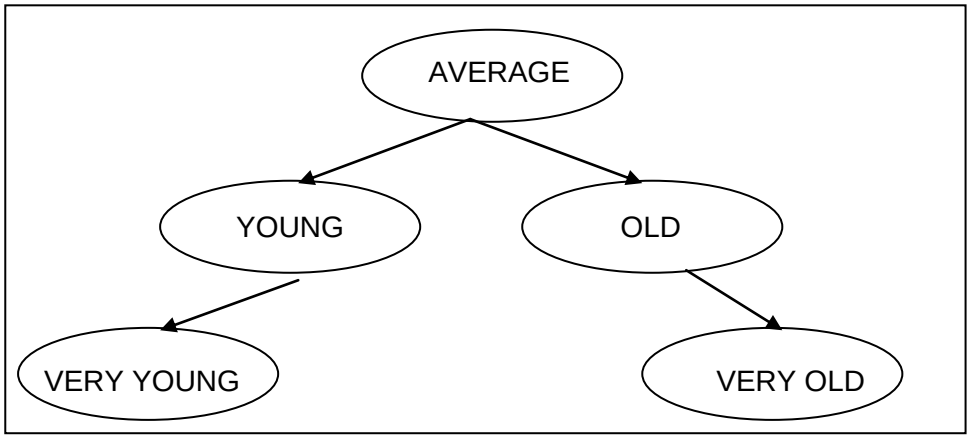

Figure 7 Age Category (Structure 1)

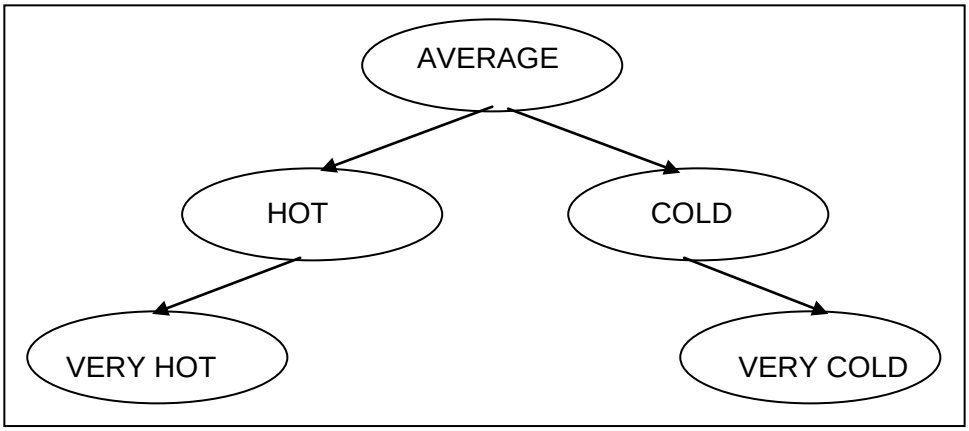

Figure 8 Temperature Category (Structure 1)

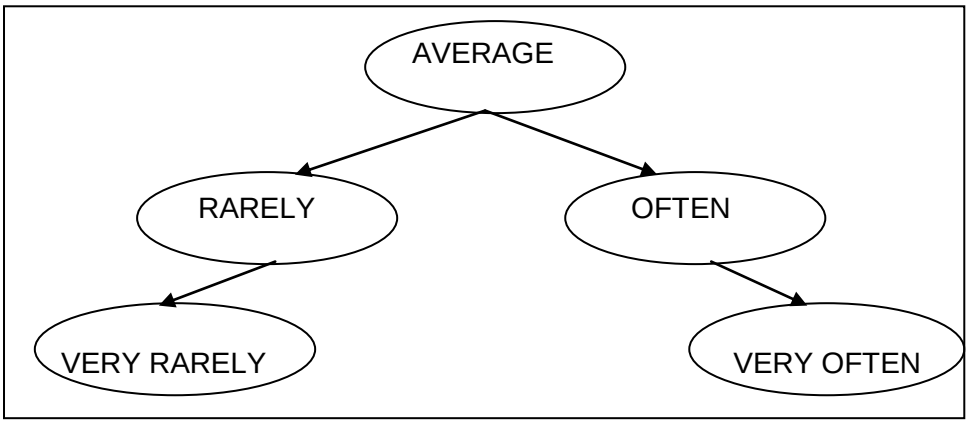

Figure 9 Frequency Category (Structure 1)

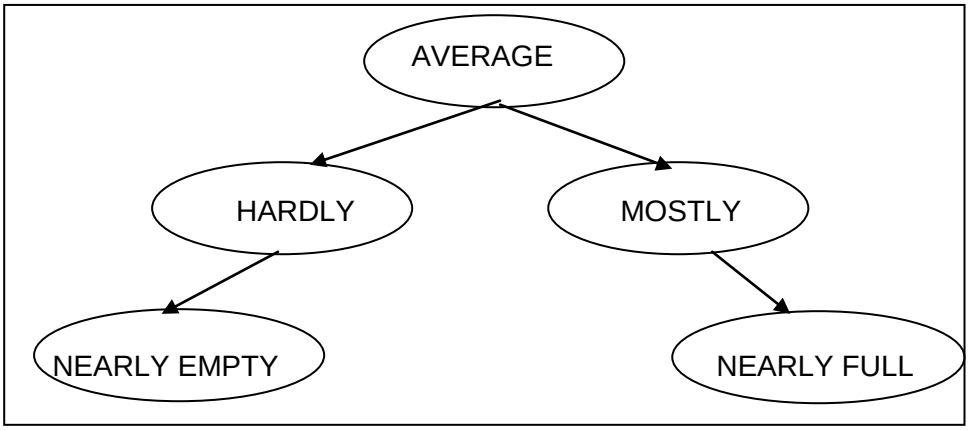

Figure 10 Membership Category (Structure 1)

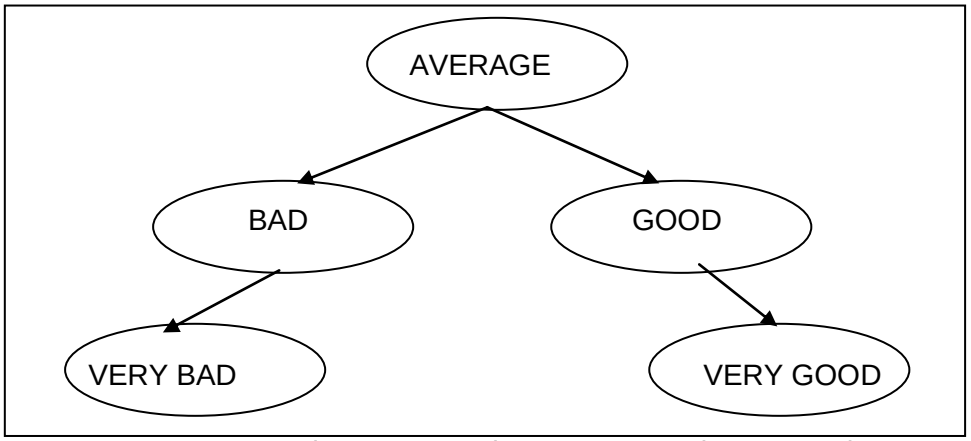

Figure 11 Goodness Category (Structure 1)

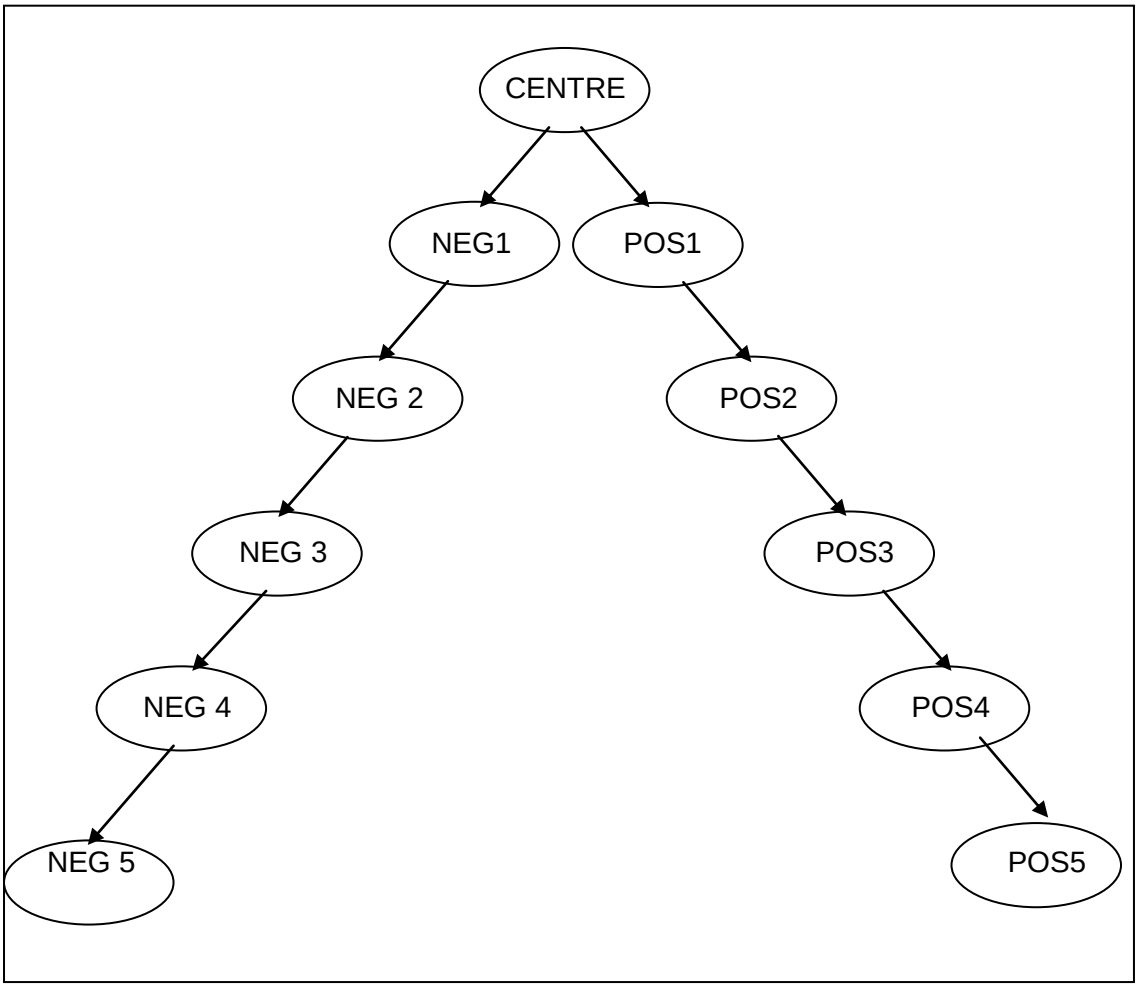

Figure 12 General Template for All Categories (Structure 2)

It is in the nodes in these diagrams that fuzzy words are stored. At this point the issue is then, how the set of fuzzy words are classified into the correct blocks (For example, whether the word “Excellent” should be stored in “Good” or “Very Good”.

To solve the classification issue, there were two stages. Firstly, given that both the structures proposed required that the fuzzy words be classified according to ranges they occupied on a -1 to 1 axis, they had to be rescaled along it (from the 0 to 1 axis that they were previously on). This was because some of the words were positively orientated while others were negatively oriented. From that point they were classified into the appropriate classes (for either of the proposed structures for the domain) within their given domains based on the locations on the axis that they occupied. For example consider the word “tiny” which on the size scale takes a value of -

0.76. This would allow it to be classified into the “Very Small” class in Structure 1 and the “Neg4” class in Structure 2. The following two tables show the how the words were classified in the size domain using both Structure 1 and Structure 2.

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

Table 11 Classification of the Size/Distance category (Structure 1)

Neg1	Microscopic
	Miniscule
Neg2	Minute
	Tiny
	Alongside
	Insignificant
	Diminutive
	Petite
	Adjacent
Neg3	Close
	Near
Neg4	Nearby
	Small
	Thin
	Proximal

	Proximate
	Little
Neg5	Regular
	Standard
Centre	
Pos1	Medium
	Normal
	Middle
Pos2	Average
	Sizeable
	Large
	Thick
	Big
	Substantial
Pos3	Distant
	Massive
	Remote
	Long
	Great
Pos4	Far
	Huge
Pos5	Enormous
	Giant
	Gargantuan
	Gigantic

Table 12 Classification of the Size/Distance category (Structure 2)

However there was another issue that needed to be addressed that was pertinent to Structure 1. How could the differences in quantities between the words within a given node be represented? As each node covered words that had a range of values, it was essential to allow them to be factored in. For example “Gargantuan” and “Immense” both belong to the same category (Very Large) but both had different values returned from human testing. This was not an issue with Structure 2 where all the nodes contain far smaller ranges of values but could show a difference in the level of similarity between words (though to a lesser level than the inter category similarity) in

Structure 1. Therefore to be able to deal with this issue, each node in itself needed to represent a small scale, with the word with the middle value representing the midpoint. Through this, the distances between the words could be used to pull the words either towards or away from the nearest nodes of a distance of up to 1 in either direction in terms of their ontological relations with each other. For an example consider a scale for the very small class in the size category.

Microscopic	-1
Miniscule	-0.81818
Minute	-0.27273
Tiny	-0.27273
Alongside	-0.18182
Insignificant	-0.09091
Diminutive	-0.09091
Petite	0.090909
Adjacent	0.181818
Close	0.363636
Near	0.636364
Nearby	0.909091
Small	0.909091
Proximal	1
Proximate	1

Table 13 Scale of Words in Very Small Class

To solve the issue of fuzzy words which were not in the domains appearing in sentences, WordNet Synsets were used. Specifically for any given adjective or adverb that was not contained within any of the domains, a review of its Synsets was done. If any word was found that was within one of the domains and of a similar type (i.e. adjective), its value in the ontology structure was taken instead. While this does expand the total number of words it is not a suitable replacement for human quantification of the words, which is the ultimate goal.

Through the application of the Word Similarity component of FAST on to these new ontological relations (by determining the distances between

classes and the subsumer distances) the relationships between fuzzy words can be incorporated into the STASIS measure to create a new sentence similarity measure that accounts for fuzzy words. Furthermore, the new approach to differentiating between words held on a particular node on a certain scale allows for the intricacies of the differences between the words to be demonstrated (See *Figure 1* for an example). The evaluation of the FAST measure with both of these two structures (and thus the evaluation of the two structures) is presented in the Results and Discussion chapter (Chapter 6). This evaluation is done through determining which of the ontological structures allows FAST to return a higher level of similarity with human results.

4.7 Determining the Effect of Fuzzy Words on Non Fuzzy Words

The fuzzy ontology allows for comparing the relations between fuzzy words. However, there is another area wherein fuzzy words affect overall sentence similarity. In a large number of cases, fuzzy words have associated non fuzzy words whose meanings they can affect. For example if comparing the words “Mountain” and “Hill” there is a given semantic value between them. However, instead, through the addition of associated fuzzy words “Small Mountain” and “Big Hill” were being compared, then there would be a difference in the level of similarity. This is because the addition of the two fuzzy words has altered the level of semantic similarity between the non-fuzzy ones. Therefore, the system needs to be able to present a representation of this alteration when the semantic is being calculated.

The first stage in representing the effect of fuzzy words on non fuzzy words was determining which pairs of words were associated. The system was implemented for this purpose. Firstly the system, upon taking a sentence as input, tagged each of the words according to type (e.g. noun, verb, adjective, etc.). The system to do this was built into the NLTK that was applied in creating the system. Given that the vast majority of fuzzy words that could affect other words are by their nature either adjectives or adverbs, the system was designed to find associated words to these word types. The system was also designed to determine when a non fuzzy had multiple

words associated (for example “A freezing cold day”) and was built to find all the fuzzy words associated with a non fuzzy word and represent their cumulative effect on its meaning. The system found the associated word based on locations within the sentence through running a series of rules based on grammatical traits. This was done by first identifying a fuzzy word which was also an adjective (so as to differentiate, for example, someone feeling cold to someone having a cold) and then applied the rules to determine the noun or adverb that was most likely to be associated with it. After the implementation of the system, it was tested for its accuracy. This was done through taking three random articles from The Independent newspaper, manually determining associated fuzzy words with nouns in each sentence in each article and running all the sentences from them into the system. This allowed the system to look through a large group of sentence with a substantial number of fuzzy descriptors of other words. The results of this showed that the system was able to correctly identify the associated fuzzy words with non fuzzy words in 80% of the sentences. This, therefore, allowed for this system to be used in the sentence similarity measure.

To represent the impact of a fuzzy word on a non fuzzy word the quantities the fuzzy words had on a -1 to 1 scale are used (they were scaled in an earlier stage when they were classified in their categories. Each of the fuzzy words exerts a pull on the non fuzzy word’s similarity with other words. This pull is represented in the subsumer depth and total distance when a word is compared to another word. When two words with associated words from the same category are compared to each other, the difference between the associated words on the scale is added to both to the distance between the words and the subsumer depth. For example consider the words Car and Ship and their ontological relationship (as is illustrated in *Figure 13*). From taking their ontological distance and subsumer depth a similarity value, X can be determined for them.

S(Car, Boat) = X

If however, fuzzy words are associated with them for example, *Small Car* and *Big Boat*, then the semantic similarity between the two original words is changed by the addition of new fuzzy words (S(Car,Boat) = X $\pm$ Y) with Y representing the total change brought about by these new words. This would be demonstrated by a change in their relative ontological positions (as is illustrated in *Figure 13*)

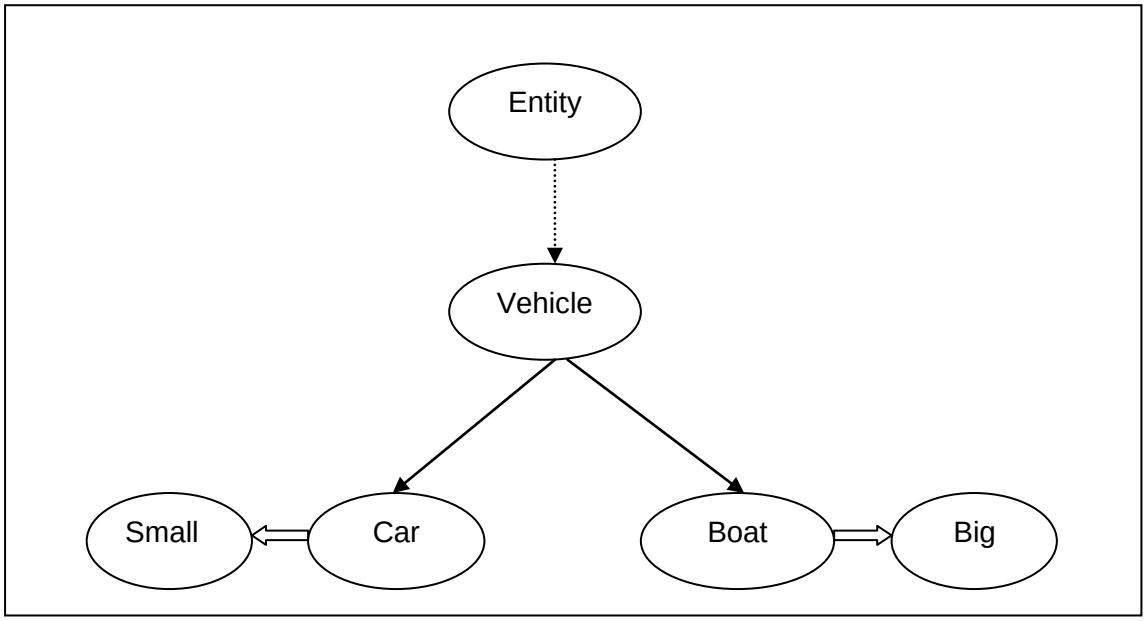

Figure 13. Ontological Relationship between Car and Boat

In cases of words where one of these words was being compared to another word without an associated fuzzy word from the same category it would instead add the distance from the word's position on the scale and 0 to the subsumer depth and the distance. Although through this method, fuzzy words only have a small influence on the overall level of semantic similarity between the non fuzzy word and others, it could nonetheless have a noticeable effect, particularly if taken cumulatively in sentences where there are multiple fuzzy words.

The establishment of this component and the method behind it allowed for another area of the experiment to be tested. Specifically, it could now be determined if fuzzy words that were inherent in the definitions of non fuzzy words could affect their similarity with others. This was tested using the Rubenstein and Goodenough dataset (Rubenstein and Goodenough 1965)

and the WordSim-353 dataset(Finkelstein et al. 2001). The process of testing involved determining the fuzzy words present in the WordNet descriptions of the fuzzy word pairs, applying the effect of the fuzzy words and then determining the levels of similarity. The results did not show a significant difference in similarity and as such it was determined that it was not necessary to look for inherent fuzzy words in the definitions of non fuzzy words.

4.8 Conclusion

This chapter has discussed how the FAST sentence similarity measure has been constructed. This involved creating new fuzzy ontologies for each of the categories based on the results of the quantification experiments in the previous chapter. This also involved looking at the STASIS word similarity measure and Sentence similarity measure and determining their suitability to be used in conjunction with FAST. The application of a formula based on distances and subsumer relations between these words allowed for a semantic similarity value to be returned for any pair of words within a given category. Through incorporating these new ontologies into the wider STASIS word similarity measure, a new word similarity measure was created. With the new measure having been created, the issue that now remained was how it could be evaluated. The systems effectiveness at determining the levels of similarity between fuzzy sentences needed to be measured against a set of human results. It also needed to be benchmarked against other similarity measures..

5 Building an Evaluation Dataset

5.1 Introduction

The previous chapter (Chapter 4) described the creation and implementation of FAST, a fuzzy semantic similarity measure. This measure was built to address the problem that was discussed in Chapters 1-3. Specifically that no existing sentence similarity measures were able to accurately represent the effect that fuzzy words have on sentence similarity. The FAST algorithm is able to take in two sets of texts with fuzzy components in them and return a similarity value for the two of them while factoring in the effect of the fuzzy words. With the algorithm having been implemented as a computer system it needed to be fully evaluated. This would mean testing the capabilities of the system against human results to determine the level of correlation. Ideally, this would have involved the existence of a dataset of pairs of sentences with a level of similarity between them determined by human participants (such as Miller and Charles and Rubenstein and Goodenough datasets, in terms of words and in terms of sentences the O'Shea sentence pair dataset (O'Shea et al. 2008a) which was used to evaluate STASIS and the subsequent STSS-131 dataset (O'Shea 2010)). Unfortunately no suitable datasets existed that could be used to evaluate FAST. This is because none of the existing sentence similarity datasets contained enough sentence pairs with fuzzy words in each sentence. The aim of this chapter therefore is to describe the creation of a new sentence similarity dataset that can be used to rigorously evaluate the FAST measure.

There are two aspects of FAST that need to be evaluated. Firstly it needs to be determined if FAST is able to accurately predict the level of similarity between pairs of sentences that contain a single fuzzy word in each. Furthermore, the ability of the system to represent the effect (if there is any) of additional fuzzy words in sentence pairs (i.e. two fuzzy words in each of the sentences in a pair) needs to be evaluated. Therefore there are two parts to the dataset that needed to be created. The first part, a Single Fuzzy Word Dataset (SFWD) needs to contain a set of sentence pairs with levels of similarity between them that contain one fuzzy word in each of the sentences. The second part, a Multiple Fuzzy Word Dataset (MFWD)

requires a set of sentence pairs with multiple fuzzy words from either the same categories or different categories (as were created in Chapter 3) in each of the pairs. In this chapter two methodologies for collecting pairs of sentences with similarity ratings (one for each of the aforementioned sets) are described and justified with their results discussed in further detail.

After the two datasets of human similarity ratings have been created, an evaluation procedure can be made to test the effectiveness of FAST. The evaluation of FAST must determine the measures effectiveness in terms of the datasets and in terms of other sentence similarity measures (which themselves will be assessed against the dataset). The overall goals of the evaluation are:

- To clearly determine the effect that fuzzy words have on the abilities of short text semantic similarity measures to represent sentence similarity
- To determine whether FAST can more successfully represent the levels of similarity between fuzzy sentences than other measures.

Therefore the two main items that needed to be considered regarding the evaluation procedure was which specific sentence similarity measures would be used to benchmark against FAST, what criteria would be tested and what methods would be used to analyse the results and draw conclusions. The development of this evaluation procedure is discussed in this chapter.

The following structure will be used in this chapter. *Sections 5.2 and 5.3* discuss the creation of the dataset that will be used to evaluate FAST. Section 2 explores how a set of sentence pairs with a single fuzzy word in each sentence is created. Section 3 further discusses how a set of sentences pairs with two or more fuzzy words in each sentence is created. Section 4 then proceeds to examine what the overall evaluation procedure would be and how it was developed. Section 5 contains a review of the chapter.

5.2 Chapter Aims

- Create an evaluation dataset of sentence pairs with one fuzzy word in each sentence.
- Create an evaluation dataset of sentence pairs with two fuzzy words in each sentence.

5.3 Building A Set of Sentence Pairs with One Fuzzy Word

5.3.1 Overview

This section describes the creation of the SFWD dataset. The SFWD 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 set there were two different steps that had to be completed to ensure that it was accurate and representative of human dialogue.

- A set of fuzzy sentence pairs had to be generated and paired, and a methodology to do this had to be generated.
- Another method would be needed to return human similarity ratings for the sentence pairs.

This would allow the sentence pairs to be put through the algorithm to be compared against these ratings do determine the algorithms' accuracy. In keeping with other datasets (Rubenstein and Goodenough 1965) (Liu and Mendel 2008) a total of 30 sentence pairs will be created.

The creation of the Rubenstein and Goodenough (Rubenstein and Goodenough 1965), Miller and Charles (Miller and Charles 1991), and O'Shea datasets (O'Shea et al. 2008a) (O'Shea 2010) were explored in detail in Chapter 2. The creation of those datasets provided a framework for which other datasets could be created. Specifically by Turney and Littman (2003) and crucially by O'Shea (2010) which explored in detail what the factors that needed to be taken into consideration when creating a sentence similarity dataset were, and methods regarding how they could be addressed. Expanding on the work done by Miller and Charles and

Rubenstein and Goodenough, O'Shea et al. created a dataset of quantified pairs of sentences (SPSS-65) (O'Shea et al. 2008a) (O'Shea et al. 2008b) and subsequently James O'Shea created the SPSS-131 dataset (O'Shea 2010), with a more substantially more detailed methodology for its creation detailed. Given that these sentences had previously been collected through "Gold Standard" methods (O'Shea 2010), incorporating them into the SFWD would ensure that the same level of quality is retained in the new dataset.

As aforementioned, the problem with using the previous datasets was the lack of fuzzy sentence pairs (sentence pairs with fuzzy words in each sentence) that had been present in any of them. Therefore, if sentences from an existing dataset were to be used, they would need to have fuzzy components added to them and then be re-quantified through human participants re-evaluating them. For existing sentences to be used what needs to be developed, therefore, is a set of sentence pairs to be fuzzified (to have a fuzzy component added to them). It was important that these news 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. After pairing them a method to quantify them (using human participants) needed to be created. It was important that the method to quantify the fuzzy sentence pairs was robust, unbiased and would not lead human participants towards specific answers (O'Shea 2010). This was a problem that was also addressed in Chapter 3 where fuzzy words had to be quantified on a specific scale.

5.3.2 Fuzzifying Sentences through the use of Linguistic Experts

Given the two options of datasets to fuzzify sentences from (James O' Shea's first and second datasets) it was decided that sentences from the second dataset (STSS-131) (O'Shea 2010) be taken. This is because of the Gold Standard (O'Shea 2010) (O'Shea et al. 2010) methodology that was

used to create the dataset and because this dataset involved the creation of new natural language sentences while the first dataset instead simply provided definitions for the words contained in the Rubenstein and Goodenough datasets to create sentences (pairing the sentence definitions of the word pairs). As was discussed in Chapter 2, the second dataset instead involved creating sentence completely from scratch using human participants and guide words to direct them towards particular themes. The justification for its Gold Standard consideration was put forward by O'Shea (2010), which presented the methodology for creating Gold Standard datasets.

With the issue of which existing dataset to take sentences from established, the next problem was how to structure the sentence pairs and how many sentence pairs would be used in total. This number would be equal to one part of a two part dataset. Given the numbers of sentence pairs that were present in each STSS datasets as well as the numbers of word pairs present in the Miller and Charles datasets (Miller and Charles 1991) (Charles 2000), the Rubenstein and Goodenough dataset (Rubenstein and Goodenough 1965) and the Wordsim-353 dataset (Finkelstein et al. 2001) it was decided that a total of 30 sentence pairs be used for this part of the dataset (practical use of both the Rubenstein and Goodenough dataset and the Wordsim-353 dataset was demonstrated by Agirre (2009)). This would by its nature mean that there would need to be 60 unique sentences which, when paired, give a complete set. The solution to the issue of pairing sentences was to fuzzify the selected sentences multiple times (adding in words from a given category) and then pair different instances of them.

In terms of fuzzifying the sentences, paraphrasing (Dolan et al. 2004) is a method that was considered. That is to rewrite the sentence while 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 (Dolan and Brockett 2005). This is a large corpus of pairs of paraphrased sentences with human similarity ratings for each pair. The widely used nature of the corpus (Callison-Burch 2008) (Androutsopous and Malakasiotis 2009) (Das and Smith 2009), 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. As a result, the numbers of fuzzy sentences pairs were substantially lower than would be required to make a dataset in keeping with the size of existing datasets).

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 (Ide et al. 1998) (Gildea and Jurafsky 2002) (Hu and Liu 2004) (Pang and Lee 2008), the effect the orientation of fuzzy words could have on a words semantic meaning was discussed. It was stated that fuzzy words could be either positively or negatively oriented (this was further demonstrated in Chapter 4, with the creation of 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 the 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. This is how the SFWD will be built.

With the concept behind fuzzifying sentences having been decided the next issue to decide, would be who would be responsible for the fuzzifying the sentences. This is important as there were two different factors that needed to be balanced. Firstly there was the issue of avoiding biases (O'Shea 2010). This was important as failing to do so could result in an inaccurate appraisal of FAST, either through over or underestimating its performance. It was for this reason that the fuzzification would have to be done through human participants (as was the case with the generation of words for fuzzy categories in Chapter 3). The next issue was that the sentences 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 many other sentence similarity measures (Islam and Inkpen 2008) (Ho et al. 2010) (Tsatsaronis et al. 2009)), and this had to be represented in the evaluation. As a result of this, some selectivity was required regarding which group of participants would fuzzify the sentences.

In his work), which followed on from work by Miller and Charles (1991) and Rubenstein and Goodenough (1965), James O' Shea discussed both the importance and usefulness of the use of linguistic experts in the generation of natural language sentence datasets (O'Shea et al. 2008a) (O'Shea et al. 2008b) (O'Shea 2010). He stated that experts, through their in depth knowledge of the English language and sentence construction, could be relied upon to, if given a sentence construction task, construct natural language sentences. As they are also impartial to the project, the risk of biases within their responses is also reduced (O'Shea et al. 2008a) (O'Shea 2010). 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. An extensive discussion of how this could be accomplished was done by O'Shea (2010), which serves as a very useful reference in the construction of this experiment.

For the purpose of creating the SFWD, in the experiment, 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 (with 20 sentence pairs selected for high levels of similarity, 5 for medium and five for low. This was to ensure the distribution of results across a range of possible similarity levels) from the O'Shea dataset (O'Shea 2010) using the definitions of high, medium and low from that dataset and asked to fuzzify using the method of amplifying or diminishing a particular aspect (Appendix 3). For example when given the instruction;

Increase or diminish, if possible, the level of delay

For the Sentences;

-When I was going out to meet my friends there was a delay at the train station
-The train operator announced to the passengers on the train that there would be a delay.

The returned Fuzzified sentences were; -

-When I was going out to meet my friends there was a significant delay at the train station
-The train operator announced to the passengers on the train that there would be a brief delay

This method therefore, returned for the sentence pair, a pair of fuzzy sentences that could be used to evaluate FAST. Through this method a total of 90 pairs of sentences (180 unique sentences in total) were created. This was enough sentences to form a dataset of natural language fuzzy sentence pairs that could be quantified. Given that the sentence pairs that were selected were already distributed in such a way as to represent the whole spectrum of similarity, no further work was needed to ensure that this was considered. There was another step, however, that was needed in the

building of the SFWD. 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 14* contains the acquired sentence pairs.

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 utmost 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.
S	-I always like to have a tiny slice	-I like to put a large wedge of lemon

P 8	of lemon in my drink, especially if it's coke.	in my drinks, especially cola.
S P 9	-The key always never works, can you give me another?	-I dislike the word quay, it confuses me every time, I always think of the thing for locks, there's another one.
S P 1 0	-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.
S P 1 1	-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.
S P 1 2	-Does this soggy sponge look dry to you?	-Does pleasant music help you to relax or does it distract you too much?
S P 1 3	-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.
S P 1 4	-Global warming is what everyone is really worrying about greatly today.	-Global warming is what everyone is mildly worrying about today.
S P 1 5	-Midday is 12 o'clock in the midpoint of the day.	-Midday is 12 o'clock in the centre of the day.
S P 1 6	-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.
S	-I am sorry but I can't go out as I	-I've a gargantuan heap of things to

P25	have loads of work to do.	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 14: SFWD Sentence Pairs

5.3.3 Quantification of Words in the SFWD

With the sentence pairs having been collected, the next stage was to quantify them. This required further human experimentation. There had been a number of different methodologies already established for quantifying both word (Rubenstein and Goodenough 1965) (Finkelstein et al. 2001) and sentence similarity (O'Shea et al. 2008a) (O'Shea 2010). Most recently, there was the work that was done by O' Shea et al. (O'Shea et al. 2008a) and by James O'Shea (O'Shea 2010). Furthermore, work was done in terms

of attributing quantitative values based on semantic meaning in Chapter 3, where sets of fuzzy words were quantified. This provides a solid background in what is needed to accurately quantify the level of semantic similarity between texts and allows for the creation of a methodology to do so. As was the case in the construction of all previous sentence similarity datasets, the collection of the similarity data is questionnaire based. The next issue would be the total number of people who results would be collected from. The number that was selected was 20. The selection criterion for participants was that they were native English speakers. While this number was lower than the number used in some sentence similarity experiments (O'Shea 2010), it is nonetheless greater than the number that was used in others (Finkelstein et al. 2001). Furthermore, it is also in keeping with the number used in the initial quantification of fuzzy words in Chapter 3, where 20 participants were used.

Having established the parameters that were for quantification a suitable questionnaire had to be designed. As with the previous stage, it was important that the approach that was taken not lead or bias the respondents' answers. There were some common parameters to all previous sentence similarity experiments that aided in addressing this problem (O'Shea et al. 2008a). They illustrated that examples could be used (just as was the case in the initial collection of sentences), to clearly given 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, (O'Shea et al. 2008a) 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.

Taking these factors into consideration, a sentence similarity questionnaire was developed (Appendix 4). This asked participants to rate pairs of sentences based on their level of similarity on a scale of 0 to 10. Notably, this scale differs from the scales used in both the O'Shea sentence similarity datasets (O'Shea et al. 2008a) (O'Shea 2010) and the Rubenstein and Goodenough and Miller and Charles datasets (Rubenstein and Goodenough

1965) (Miller and Charles 1991). However, it was in keeping with the scales that were used in the WordSim-353 word similarity dataset (Finkelstein et al. 2001) and the Mendel Codebook (Liu and Mendel 2008), which was discussed heavily in Chapters 2 and 3. The reason that this scale is used is because of the nature of FAST, which returns results on this scale. Using the 0 to 4 scales that had been used in previous datasets (even with the caveat that allowed respondents to add in a decimal if they wanted), and rescaling the results run the risk of diminishing the level of accuracy of the comparisons. Crucially, the 0 to 10 scale had been used in Mendel's codebook which was specifically geared towards fuzzy quantification. Given that the WordSim-353 provided proof of concept, there were no hindrances to using the 0 to 10 scale. With the questionnaire having been developed, it was distributed to participants using both e-mail and hard copy formats. *Table 15* shows the similarity results that were collected, in terms of average values and standard deviations (with SPs. Corresponding to *Table 14*).

SP	Average Human Rating	Standard deviation
SP 1	3.833	2.021
SP 2	0	0
SP 3	7.3	1.995
SP 4	7.952	1.85
SP 5	1.281	2.43
SP 6	8.719	1.002
SP 7	7.095	1.737
SP 8	6.719	1.762
SP 9	0.952	1.8
SP 10	8.248	1.008
SP 11	4.957	1.489

SP 12	0.529	0.978
SP 13	3.286	2.57
SP 14	6.371	1.827
SP 15	9.138	0.892
SP 16	6.781	1.81
SP 17	3.229	2.386
SP 18	2.11	1.995
SP 19	6.757	2.212
SP 20	8.986	0.784
SP 21	3.548	3.24
SP 22	8.852	1.45
SP 23	7.043	1.623
SP 24	3.833	2.296
SP 25	8.857	0.964
SP 26	7.583	1.835
SP 27	8.919	1.076
SP 28	6.914	2.016
SP 29	1.295	2.211
SP 30	6.624	2.398

Table 15: SFWD Human Ratings

As has been aforementioned the results in *Table 15*, when used in sentence similarity evaluation make two key contributions;

- They demonstrate whether or not fuzzy words have any effect on the accuracy of sentence similarity measures (i.e. whether or not they

return a lower level of correlation with human ratings), and therefore by proxy, if they have any effect on the overall semantic meanings of sentences. If the answer is negative, then the need to use a fuzzy sentence similarity measure is redundant as existing similarity measures would suffice under the existing circumstances.

- The results demonstrate the usefulness of the FAST sentence similarity measure (whether or not it can produce a high correlation with human similarity ratings). Furthermore through benchmarking FAST against other sentence similarity measures, it can be determined whether the methodology used by FAST to represent the effect of fuzzy words on sentence similarity is successful in its role.

5.4 Building a Set of Sentence Pairs with Multiple Fuzzy Words

5.4.1 Overview

With a set of sentence pairs containing one fuzzy word per sentence having been created further work needed to be done to create an additional evaluation dataset. Having created a set of words to establish whether or not fuzzy words affect sentence similarity and determine if FAST can represent that effect, another set of words was required to determine if increasing the number of fuzzy words in a sentence further affects similarity. Furthermore it also neeeded be determined whether or not FAST could maintain an improvement over existing measures with an increased number of fuzzy words in sentences. Therefore a new method needed to collect and quantify a set of sentence pairs with at least two fuzzy words per sentence per pair. The addition of this set of fuzzy sentences to the SFWD completes the total fuzzy dataset. This section describes the creation of the Multiple Fuzzy Word Dataset (MFWD).

5.4.2 A Corpus Based Method of Building a Fuzzy Dataset

Given the complexity of the sentences that would now be required, the expert method that was used to build the first half of the dataset could no longer be used. Therefore a new method with a different approach would be needed to create a new set of fuzzy sentences and then pair them. For the purposes of expediency developing an automated method was considered. Specifically the issue of whether a system of extracting sentences with fuzzy components from a corpus, fuzzifying them and then pairing them could be implemented to create the set. There has been substantial work that has been done in terms of extracting semantic information from corpuses), with some components of FAST already interfacing with the Brown and WordNet corpuses (Li et al. 2006) (Islam and Inkpen 2008) (Tsatsaronis et al. 2009). The existing work done in information extraction from Corpuses provides a framework from which an algorithm that can create a range of sentence pairs through extracted corpus information that fit the required criteria. A problem that does exist is that sentence pair that could be created in this manner would necessarily be as representative of natural language as sentences that were created using the expert method as they are artificially generated. However, an automated method would be much faster than the expert method 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 (Francis 1965), using them even after further fuzzification is not likely to significantly reduce their naturalness.

5.4.3 Selecting a Corpus

Before the extraction methodology could be designed, the corpus that would be used needed to be considered. It was possible to have had the extraction algorithm taking sentences from multiple corpuses simultaneously (randomly picking sentences from each one) but given the size and the variety of different sources available in each individual corpus this would have been unnecessary and a single corpus would suffice. Of the available corpuses, the Gutenberg Project corpus was selected (Hart 1971). This corpus contains a wide variety of texts from a number of different sources. It has

been used extensively used in a number of different Natural Language Processing projects (Madnani 2007) (Shmidt and Colomb 2009) 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 focussed 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. With the base corpus to collect sentences from having been determined, the next stage was to implement an algorithm to build the fuzzy sentence set.

5.4.4 The Sentence Pairing Algorithm

The algorithm takes as input the maximum length of a sentence (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. A given category is defined as C. These values were selected to ensure that a suitable range of results was returned by the algorithm. The sentence length of 30 was selected as that was considered to be the maximum length a set of text could be for it to be considered as a sentence (as was discussed in Chapter 4) As with the FAST algorithm, the Sentence Pairing Algorithm was coded entirely using Python. It also used the NLTK library to tag words in sentences and interface with and extract sentence from the Gutenberg corpus.

1) Let T = Sentences in Gutenberg Corpus
2) Let F = List of all fuzzy words in categories
3) Tag each Sentence (S) in T
4) For S in T :
    a. If Length of S < Ln :
        i. If Number of words W in S and F = Fz:
            1. Add S to List Sf
5) Apply High Similarity Algorithm
6) Apply Medium Similarity Algorithm
7) Apply Low Similarity Algorithm

High Similarity:

1) Let Fp = List of all positively oriented fuzzy words
2) Let Fn = List of all negatively oriented fuzzy words
3) Select SP random sentences in Sf AS Sr
    a. For Sentence S in SP
        i. Clone S as S1
            1. For word W in S1:
                a. If W in Fp:
                    i. Replace W with Random word W1 in Fp where W and W1 in C
                b. Else If W in Fn
                    i. Replace W with Random word W1 in Fn where W and W1 in C
        ii. Add S and S1 as pair to List TSet
4) Return TSet

Medium Similarity:

1) Fp = List of all positively oriented fuzzy words
2) Fn = List of all negatively oriented fuzzy words
3) Select M random sentences in Sf AS Sr
    a. For Sentence S in SP Where S not in TSet
        i. Clone S as S1
            1. For word W in S1:

a. If W in Fp:
    i. Replace W with Random word W1 in Fn where W and W1 in same domain
b. Else If W in Fn
    i. Replace W with Random word W1 in Fp where W and W1 in same domain

ii. Add S and S1 as pair to List TSet

4) Return TSet

Low Similarity

1) Select L * 2 Random Sentences in Sf Where Sentences not in TSet. AS Sr
2) Randomly Pair all sentences in Sr
3) Add all Sentence Pairs to TSet

Figure 14 Sentence Pairing Algorithm

5.4.5 Overview of the Sentence Pairing algorithm

Step 1: Step one specifies all the sentences in the Gutenberg corpus as a single set. Collecting this set is done through interfacing with the corpus using the NTLK. Once the list has been collected the sentences can be dealt with and parsed as individual entities.

Step 2: Step 2 specified a list of all the fuzzy words across all the various domains. This list is referenced to determine the presence of fuzzy words in any of the sentences.

Step 3: This step involves tokenizing each of the sentences (separating them into their individual constituent words). Essentially the sentence now becomes 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 4: This step involves generating a list of all fuzzy sentences where there are two fuzzy words in each of the sentences. It also determines which sets of text fit the criterion of a “Sentence” (having 35 words or less). For all the given sentences, the algorithm first looks at the length of the sentence (the number of words) and 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 the list F, the algorithm determines the presence of fuzzy words in the sentences. The measure is specifically looking for sentences that contain a number of fuzzy words equal to the Fz parameter which is specified beforehand. If the sentence does have the correct number of fuzzy words, it is then added to another list of sentences Sf. The sentences within this list are used for the purpose of generating sentence pairs.

Step 5: The high similarity sub-algorithm is applied **(see High Similarity Algorithm)**

Step 6: The medium similarity sub-algorithm is applied **(see Medium similarity Algorithm)**

Step 7: The low similarity sub-algorithm is applied **(see Low similarity Algorithm)**

High similarity Algorithm

Step 1: All the positively oriented 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.

Step 2: A similar procedure is then applied to all the negatively oriented words (that have a value of less than 0 on the classification scale.

Step 3: Step 3 involves the actual generation of sentence pairs. This is done 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 original sentence and the one with replaced words). 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. A majority of the sentence pairs used in the dataset are created at this stage.

Medium similarity Algorithm

Step 1: This step is identical to the first step in the high similarity algorithm.

Step 2: This step is identical to the second step in the high similarity algorithm.

Step 3: At this stage, words with a low level of similarity are generated. This is done through a similar procedure to the one used in step 3 of the high similarity algorithm. Firstly, before any sentences are selected however, the algorithm checks to ensure that instances of the sentence do not already exist in the TSet list. This is to prevent repetition of the sentences from occurring. 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

Step 1: 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.

Step 2: All the sentences that have been collected are now randomly paired with each other. Given the vast range of different sentences that are present in the corpus, this makes is 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.

Step 3: The sentence pairs that have been generated using this method are added to the TSet list.

The list of sentence pairs that have been collected through use of the sentence pairing algorithm are returned in the form of the TSet list. *Table 16* shows the complete list of sentence similarity pairs with two fuzzy words that was collected.

SP	**Sentence 1**	**Sentence 2**
SP1	How marvellous 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	A great watcher and very

	very dietetically pathetic is Dr Bunger	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	We will not say how near for fear of shocking the young ladies

	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

Table 16: MFWD Sentence Pairs

With the second set of sentence pairs having been generated, the next stage would involve quantifying them. Following this they could, along with the first set of sentence pairs (the generation and quantification of which was discussed in Section 5.2) form a complete sentence similarity dataset.

5.4.6 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 than for the earlier component. More pertinently, time was also now a factor as collecting a second set of human ratings through the handed out questionnaire method that was utilized with the first set of sentence pairs could delay the evaluation procedure unnecessarily. These issues presented an opportunity to utilize a novel approach to collecting test data. A method that had been used in a number of areas for collecting data from human participants was crowdsourcing (Snow et al. 2008). Crowdsourcing refers to, in this particular instance, collecting information from a group of people who volunteer to participate through a common interface (such as a website) for a small monetary reward. Crowd sourcing has had multiple applications in the fields of computer science and Natural Language Processing (Kittur et al. 2008) (Munro et al. 2010) (Tellex et al. 2011)

One major tool for crowdsourcing was the Crowdflower system (Carvalho 2011). This allows for users to complete a survey (or questionnaire) for a monetary reward (or optionally none at all) that is specified by the survey's creator. It also allows the 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 the 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 second set of data, two sources would be used. The collection of results would be divided between small numbers of direct surveys to participants (as had been done with the SFWD) 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. If the answer were negative, it would mean that the crowdsourced answers could be used in conjunction with the direct ones collected in the first set of sentences. The

survey was created using the same methodology that was used to create the SFWD (as described in Section 5.2), with the use of a 0 to 10 scale and examples to clarify instructions to the users.

Through this, a total of 26 responses were collected from participants (22 of these results were from crowdsourced participants).A Student's t-test to test the hypothesis

H1: Non-Crowdsourced results will be different from Crowdsourced results

With the ensuing null hypothesis

H0: Non-Crowdsourced results will be the same as Crowdsourced results

Returns a p-value of 0.96. This very strongly suggests that there is no significant difference between Non-Crowdsourced and Crowdsourced results. What this illustrates is the similarity of the two sets of standard deviations from the crowdsourced and non-crowdsourced results. This, therefore, opens a new avenue in terms of data collection for any future work. It furthermore means that, in terms of the evaluation, the crowdsourced results can comfortably be used with the directly collected results that made up the first part of the dataset (Section 2). The similarity ratings and standard deviations for the set of sentence pairs with two fuzzy words are presented in *Table 17*. The Sp-values in *Table 17* correspond to the Sp-values in *Table 16*.

SP	Average Human Rating	Standard Deviation
SP 1	5.623	2.934
SP 2	1.715	2.059
SP 3	3.769	2.27
SP 4	0.75	1.621
SP 5	3.708	2.748
SP 6	8.35	1.906

SP 7	5.677	2.616
SP 8	3.842	2.815
SP 9	4.873	2.594
SP 10	6.865	2.156
SP 11	1.223	2.373
SP 12	7.127	2.366
SP 13	5.285	2.62
SP 14	5.938	2.144
SP 15	7.381	1.949
SP 16	3.238	2.844
SP 17	4.312	2.88
SP 18	1.446	2.391
SP 19	7.792	2.609
SP 20	7.815	1.974
SP 21	2.112	3.37
SP 22	6.25	2.719
SP 23	8.162	1.911
SP 24	7.215	2.43
SP 25	7.485	1.916
SP 26	6.331	2.482
SP 27	3.842	2.564
SP 28	1.269	1.87

SP 29	6.069	2.656
SP 30	6.488	2.615

Table 17: MFWD Human Ratings

With the two sets of fuzzy sentence pairs having now been quantified, they together form a single complete fuzzy sentence similarity dataset. The combination of them (with a total of 60 sentence pairs between them) can now be used to evaluate the FAST sentence similarity measure and generally test the effect of fuzzy words on sentence similarity.

5.5 Conclusion

In conclusion, this chapter featured the creation of a new fuzzy dataset that can be used in the evaluation of any future sentence similarity measures. It alongside the STSS sentence similarity datasets is one of the very few dedicated sentence similarity datasets available. The mains contributions of this chapter are

- A methodology for collecting sentence pairs using linguistic experts and then quantifying them.
- A pairing algorithm for the automated creation of fuzzy sentence pairs from any given corpus.
- Proof of concept for the efficacy of crowdsourcing as a mechanism for quantifying the level of similarity between sentence pairs.
- A dataset of 60 pairs of fuzzy sentences (30 of which contain one fuzzy word per sentence, 30 of which contain 2) that can be used to evaluate any sentence similarity measure.

With the construction of the dataset, the FAST algorithm (and others) could now be completely evaluated against it and the results of its evaluation analysed to determine the value of FAST (which is done in Chapter 6).

6 Experimental Results

6.1 Introduction

After the FAST measure was created there needed to be a method of evaluating it to determine its accuracy and thus how useful its application would be. This required the creation of an evaluation sentence similarity dataset as, prior to the creation of FAST, there were no suitable datasets because none of existing ones contained a significant number of fuzzy sentence pairs (pairs of sentences with one or more fuzzy words in each of the sentences). A new dataset was created in Chapter 5. This dataset was a concatenation of two smaller datasets that were created in that chapter that were known as the Single Fuzzy Word Dataset (SFWD) and The Multiple Fuzzy Word Dataset (MWFD). The SFWD contained sentence pairs with a single fuzzy word in each sentence in the pair. The MFWD contained sentence pairs with multiple fuzzy words in each sentence per pair. Two different methodologies were used to create these datasets, both of which are described in Chapter 5. The SFWD and MWFD are the basis for the evaluation of FAST described in this chapter,

The goal of this chapter is to describe the evaluation of the FAST system. It contains the different techniques that will be used, the parameters that will be tested, a presentation of the results of the evaluation and discussion of the results and the implications of those deductions. The aim is to demonstrate whether FAST can accurately represent the effect of fuzzy components in terms of sentence similarity, what the best ontology structure (of the two structures that were created in Chapter 4) to use with FAST and to what level FAST could contribute to the field of sentence similarity. In giving FAST a thorough assessment, there are a number of different areas that need to be explored before its evaluation could take place.

The aim of this chapter, therefore, is to give a clear picture of the levels of effectiveness of FAST and to identify any areas of weakness that could allow FAST to be improved in future implementation. To accomplish this, the following tasks were required to be successfully completed.

- Determine if fuzzy words have a significant effect on sentence similarity (and the level of similarity between sentence pairs by proxy). This is done through determining if fuzzy words in sentence pairs affect the performance of previously existing sentence similarity measures (i.e. Would their performance diminish if faced with a sentence pair that contained fuzzy words?).
- Determine which of the two possible ontological structures that could be used to construct FAST (as was described in Chapter 3), would allow the algorithm to more accurately represent semantic similarity. This can be determined through analysing results produced by the two structures against the sentence similarity dataset. This allows a particular structure to be used in the general implementation of FAST (which can then be used in all future experiments with the system).
- Determine whether or not FAST is able to accurately represent the level of sentence similarity between pairs of sentences that contained fuzzy words in each sentence. This is determined through examining the similarity levels that were returned by FAST (using the better ontological structure) against human ratings and determining their level of correlation .
- Benchmark FAST against existing sentence similarity measures. This was to determine whether or not FAST was able to show an improvement over existing similarity measures and if it was, how significant these improvements were. This process involves comparing the results that FAST returned with the results from other selected similarity measures. This involves an analysis not just of whether FAST was closer to human results, but how significant that increased closeness was.

For all the aspects of the system to be fully evaluated, a robust methodology was needed to ensure that the evaluation procedure was fair, thorough and addressed all the issues that needed to be addressed. Some aspects of evaluating sentence similarity measures had been covered by O'Shea et al. (2008b), and these would serve as valuable references when determining the best methods to use for the procedure. One significant paper that was

used to evaluate sentence similarity measures was (O'Shea et al. 2008b), which actually covered the comparison of one measure (STASIS) against another (LSA) based on their performances when compared against human sentence similarity ratings.

For a complete evaluation, the procedure was divided into three separate experiments.

- Experiment to test whether or not fuzzy words had an effect on the semantic meaning of a sentence. This experiment addresses the first of the chapter's stated objectives
- Experiment to test which ontological structure allows FAST to perform best. This addresses the chapter's second objective
- Experiment to benchmark FAST and test its performance in relation to other sentence similarity measures. This represents the chapter's third and fourth objectives

This chapter describes three experiments (which are covered in Sections *6.2*, *6.3* and *6.4* respectively). Each section contains a discussion on the methodology for the experiment that it covers, a presentation of the results of that experiment and a discussion on what the implications of these results are and how they can be utilized.

6.2 Experiment 1: Measuring the Effect of Fuzzy Words on Sentence Similarity

6.2.1 Methodology

Before any evaluation of FAST could be done, the necessity of the evaluation had to first be established. The justification for the development of a fuzzy sentence similarity measure was based on two concepts.

A: Fuzzy words have an effect on semantic sentence similarity ratings

B: Existing semantic sentence similarity measures are unable to accurate represent the effect fuzzy words have on sentence similarity.

The first concept, (A), (as was explored in more detail in Chapter 2) was based on 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 hypothesis was formed from an analysis of the extensive work done in the fields CWW and fuzzy (Zadeh 1999) (Zadeh 1997) and through exploration of the limitations of existing sentence similarity measures (Li et al. 2006) (Islam and Inkpen 2008). The second concept (B) was that, given fuzzy words had an impact on sentence similarity (assuming A was true), existing semantic similarity measures were not well suited to representing the effect of these fuzzy words. This was derived from observing the limitations that existed from existing ontologies (as was explored in more detail in chapter 2) in terms of their ability to determine the relationships between sets of fuzzy words. For these two concepts to be tested a number of experiments needed to be carried out on the sets of fuzzy sentence pairs that had been collected in Chapter 5.

It needed to be established whether or not the addition of fuzzy words to an existing pair of sentences affected their overall meanings enough to alter their levels of similarity. This test could be done through comparing the sentence pairs from the Single Fuzzy Word Dataset(SFWS) with the corresponding sentences from the STSS-131 dataset (O'Shea 2010) from which the SFWD sentences (before being fuzzified) were derived. Specifically, the difference could be determined through looking at the levels of variation between the quantities from human ratings of the two sets of data. Given the low level of variation among results when the STSS-65 results were collected by O'Shea et al. (2008a) and the STSS-131 results were collected by O'Shea (2010) (a similarly low level of inter-result variance was found among the results for the fuzzy dataset as was shown in Chapter 3) if fuzzy words had no effect on similarity, then there should be a low variance between the SWFD results and their corresponding STSS results

The next issue to be determined was if the addition of fuzzy words had an adverse effect on the abilities of existing sentence similarity measures to accurately calculate a semantic similarity value for the sentences. The

specific semantic similarity measures that were chosen to test the effect of fuzzy words were STASIS (Li et al. 2006) and LSA (Landauer et al. 1998). These measures were chosen because of their wide use (which is particularly true of LSA and because they had been previously evaluated against each other with datasets that were made for that purpose) (Deerwester et al. 1990) (Landauer et al. 1998) (Hofmann 1999). If the measures were able to represent fuzzy words successfully, this fact would be represented in their ability to process the fuzzy datasets and the results that they returned would be in keeping with the results that they returned from non fuzzy datasets. Therefore, the STASIS and LSA measures will be tested against both the Single Fuzzy Words Dataset and the Multiple Fuzzy Word Dataset with the overall performance compared to their performance when dealing with non fuzzy datasets. This also allows the issue of whether or not the presence of multiple fuzzy words increases an potential level of diminishment in the accuracy of STASIS and LSA..

6.2.2 Results and Discussion

Testing of A:

Table 18 shows the differences between the human similarity ratings from the O'Shea dataset (O'Shea 2010) and the similarity ratings from corresponding sentence pairs in the SFWS dataset. Altogether 30 pairs of sentences were compared. The O'Shea dataset was rescaled from the 0 to 4 scale to the 0 to 10 scale. This was to allow the differences between them to be more clearly illustrated. The use of the 0 to 10 scale as a standard was discussed in detail in Chapter 5. Sentence pairs (corresponding between the two datasets) are shortened to SP.

SP	STSS-131	SFWD	Difference
1	7.825	3.974	3.851
2	0.4	0	0.4

3	7.1	7.437	0.337
4	9.15	8	1.15
5	0.225	1.416	1.191
6	9.775	8.795	0.980
7	8.95	7.053	1.897
8	9.525	6.953	2.572
9	1.8	1.053	0.747
10	7.65	8.168	0.518
11	8.05	5.058	2.992
12	0.25	0.532	0.282
13	3.625	3.158	0.467
14	7.85	6.568	1.282
15	9.9	9.047	0.853
16	9.625	7.021	2.604
17	8.975	3.411	5.564
18	2.625	2.279	0.346
19	9.825	6.995	2.830
20	9.7	8.984	0.716
21	5.525	3.921	1.604
22	9.6	8.837	0.763
23	8.4	7.047	1.353
24	5.45	4.079	1.371

25	9	8.895	0.105
26	8.975	7.645	1.330
27	8.9	8.963	0.063
28	9.575	7.063	2.512
29	1.25	1.432	0.182
30	9	6.637	2.363

Table 18: Comparison of SFWD and O'Shea dataset

As can be seen from comparing the values from the two datasets (as is illustrated by the level of difference between them), there are a number of cases where a large difference exists between the human ratings that were collected for the SWFD dataset and the O'Shea et al. dataset. Between the two datasets there exists an average difference of 1.44 (an 11.4% difference). While this is not a large difference it does show that the fuzzy words do exert an effect on sentence similarity, showing that fuzzy words change the meanings of sentences. Conducting a paired Student's T-Test of the means to test the hypothesis

H1: Adding fuzzy words to a pair of sentences affects their level of similarity

With the null hypothesis

H0: Adding fuzzy words to a pair of sentences does not affect their level of similarity

Returns a probability of 0.004% of accepting H0 (a p-value of 0.00004) and thus it is rejected. This strongly suggests that the addition of fuzzy words affects the level of similarity. As a result of this, the addition of the fuzzy words then the addition of further fuzzy words could further increase the difference. More importantly, in a number of cases a high level difference exists in the level of similarity between the STSS-131 and SFWD data. This can be demonstrated in Sentence Pairs such as SP30 and SP28. This indicates that fuzzy words can affect sentence similarity at different levels

and in some cases can have a large impact. This indicates that fuzzy words do have an impact upon sentence similarity. This provides strong evidence that H1 can be accepted and serves as a justification for creating a fuzzy sentence similarity measure such as FAST. The next issue to be addressed is if existing sentence similarity measures are affected by the addition of fuzzy words.

The following table (*Table 19)* shows the performance of both STASIS and LSA when they were measured against the SFWD (their performance in terms of the O'Shea et al. dataset had already been discussed by O'Shea et al. (2008b)). This allows for their differences in performance, when dealing with fuzzy sentences to be gauged and for H2 to be tested.

SP	SFWD	LSA	STASIS
1	0.397	0.48	0.750
2	0	0.01	0.468
3	0.744	0.26	0.671
4	0.8	0.84	0.747
5	0.142	0.02	0.555
6	0.880	0.95	0.627
7	0.705	0.63	0.854
8	0.695	0.81	0.78
9	0.105	0.49	0.616
10	0.819	0.46	0.708
11	0.506	0.49	0.413
12	0.053	0.32	0.488
13	0.316	0.05	0.565

14	0.657	0.93	0.924
15	0.905	1	1
16	0.702	0.7	0.844
17	0.341	0.59	0.321
18	0.228	0.61	0.497
19	0.699	0.79	0.779
20	0.898	0.36	0.824
21	0.392	0.28	0.545
22	0.884	0.42	0.882
23	0.705	0.8	0.859
24	0.408	0.39	0.707
25	0.889	0.72	0.742
26	0.764	0.96	0.867
27	0.896	0.71	0.708
28	0.706	0.88	0.862
29	0.143	0.16	0.385
30	0.664	0.48	0.534

Table 19: LSA and STASIS tested against SFWD

From the results in *Table 19*, it can be seen that LSA and STASIS have differing levels of accuracy in terms of representing sentence similarity given the dataset. Using the Pearson's Correlation (which was also used in the initial assessment of STASIS against LSA (O'Shea et al. 2008b)) it can be seen that STASIS has an overall correlation of 0.708 while LSA has a correlation of 0.65. In both these cases, this is substantially lower than the

values that were returned when they were compared against the O'Shea dataset (O'Shea 2010). There are two main implications that can be taken from this. Firstly it can be seen that the presence of fuzzy words diminished the level of accuracy of both the existing similarity measures. Secondly, it can be seen that STASIS outperformed LSA by a significant margin. This serves as an indication that the ontology-based approach is more successful than a purely corpus based approach in terms of dealing with fuzzy datasets. Therefore B is accepted and the case for a new sentence similarity measure that can deal with fuzzy words is presented.

What this portion of the experiment has illustrated was that the presence of fuzzy words adversely affected the success of sentence similarity measures. This therefore implies that to improve success, a sentence similarity measure needs to explicitly deal with the fuzzy words that are present in a sentence. This serves as the justification for the creation of FAST, which is built to achieve that goal

6.3 Experiment 2: Empirical Determination of FAST Ontological Structure

6.3.1 Methodology

The aim of this experiment was to empirically determine which of the two ontological structures used in FAST implementations (Structure 1 and Structure 2) was to be used in a general implementation of FAST. This was through determining which of them was able to more accurately represent the level of similarity between pairs of fuzzy sentences.

The methodology to determine the best structure needed to focus specifically on correlation with human similarity ratings in a sentence similarity dataset. However, when in future work the set of quantified fuzzy words is expanded (beyond the current set of words and added synonyms that are currently used) other factors beyond accuracy would need to be considered. There are factors such as the computation time which could increase.. The methodology to select the correct ontology involved the

following steps. It made use of the fuzzy sentence similarity dataset that was presented in chapter 5.

There are two stages in determining the best ontological structure to utilize with FAST.

1. **-Testing the ontological structures against the SFWD**
2. **-Testing the ontological structures against the MFWD**

The first stage of testing is to determine which of the ontological structures enables the sentence similarity measure to perform better (return values with a higher to human ratings) when a single fuzzy component whatsoever is factored. To do this both of the implementations needed to be tested against the SFWD. As the SFWD dataset contains a set of human ratings for each of the sentence pairs, the similarity ratings that each of the FAST ontology implementations returns for each of the sentence pairs can be compared to the human ratings. Therefore the structure that is able to return results that are closer to the human ratings can be taken as more accurate and able to return results that are more representative of human perceptions of sentence similarity.

The next stage in determining what the best structure is to use in the implementations with the MFWD. If increasing the number of fuzzy words in a fuzzy sentence pair increases or diminishes the level of similarity between the sentences in a fuzzy pair then the issue of whether the level of accuracy of the structures from stage 1 remained consistent needed to be addressed. Through running the sentence pairs in the MFWD through the FAST ontology implementations, correlations are returned for both the implementations. Then through looking at the overall from the correlations returned by the two ontologies a decision can then be made on which structure to use based on which one returned higher correlations

6.3.2 Results and Discussion

Tables 3 and 4 illustrate the different performances of the two ontology structures when compared with the human results from the fuzzy datasets.

Table 20 shows a comparison between FAST implementations with Structures 1 and 2 and the Single Fuzzy Word Dataset (SFWD).

SP	SFWD	Structure 1	Structure 2
1	3.833	0.719	0.71
2	0	0.474	0.468
3	7.3	0.778	0.796
4	7.952	0.744	0.744
5	1.281	0.555	0.555
6	8.719	0.627	0.627
7	7.095	0.848	0.845
8	6.719	0.779	0.771
9	0.952	0.616	0.608
10	8.248	0.825	0.821
11	4.957	0.406	0.404
12	0.529	0.477	0.469
13	3.286	0.605	0.612
14	6.371	0.891	0.88
15	9.138	1	1
16	6.781	0.898	0.879
17	3.229	0.501	0.499
18	2.11	0.514	0.498
19	6.757	0.782	0.765

20	8.986	0.836	0.836
21	3.548	0.545	0.545
22	8.852	0.902	0.902
23	7.043	0.891	0.858
24	3.833	0.713	0.71
25	8.857	0.769	0.758
26	7.583	0.919	0.894
27	8.919	0.795	0.804
28	6.914	0.862	0.862
29	1.295	0.385	0.385
30	6.624	0.574	0.576

Table 20: Ontology Structures 1 and 2 tested against SFWD

Using Pearson's correlation, it can be seen that Structure 1 has a correlation with the human similarity ratings of 0.77. Structure 2 had a correlation of 0.78. The first thing that can be noted from this experiment was that a high correlation with the human similarity ratings was shown by both of the structures. The correlation was similar in scope to the differences between STASIS, LSA and human similarity ratings that were tested by O'Shea et al. (2008b), where the difference was determined as being significant. This therefore indicates the improvement of FAST over STASIS. The results showed a level of improvement of Structure 2 over Structure 1. This was however not a substantial improvement, with structure 2 only showing a 1.2% improvement, which is not significant. This is further illustrated through conducting a Fisher r to z transformation, which returns a pvalue of 0.91, strongly indicating that any difference between the two structures in terms of the SFWD is down to chance. Therefore the Multiple Fuzzy Word Dataset

(MFWD) was necessary to determine the best structure. *Table 21* shows the comparison of Structures 1 and 2 against the MFWD.

SP	MFWD	Structure 1	Structure 2
1	5.623	0.904	0.897
2	1.715	0.588	0.656
3	3.769	0.944	0.898
4	0.75	0.21	0.198
5	3.708	0.901	0.892
6	8.35	0.997	0.997
7	5.677	0.937	0.92
8	3.842	0.978	0.973
9	4.873	0.822	0.808
10	6.865	0.969	0.962
11	1.223	0.577	0.575
12	7.127	0.996	0.967
13	5.285	0.97	0.93
14	5.938	0.967	0.94
15	7.381	0.943	0.923
16	3.238	0.76	0.826
17	4.312	0.965	0.934
18	1.446	0.362	0.329
19	7.792	0.975	0.968
20	7.815	0.792	0.593

21	2.112	0.625	0.625
22	6.25	0.993	0.989
23	8.162	0.996	0.996
24	7.215	0.844	0.845
25	7.485	0.854	0.732
26	6.331	0.859	0.809
27	3.842	0.967	0.961
28	1.269	0.438	0.435
29	6.069	0.913	0.909
30	6.488	0.965	0.967

Table 21: Ontology Structures 1 and 2 tested against MFWD

The results in *Table 21* show that while the correlation for Structure 1 remains high at 0.765, the correlation for Structure 2 drops to 0.679. A Fishers R to Z transformation test returns a p-value of 0.5 showing no significant between the results. However, at this point the difference in similarity between the two correlations has increased to 11.7%. Therefore it is potentially the case that the accuracy of the Structure 2 implementation drops as more fuzzy words are added (falling below a point at which it would be inadequate even with two fuzzy words per sentences).

Therefore the best option for a general implementation of FAST was Structure 1. This meant that all future work done by FAST would incorporate this structure (the creation of which is described in Chapter 4). Therefore whenever FAST is referred to henceforth it is to be assumed it is the implementation with that structure. The work that has been done at this stage of the implementation now allows for a full evaluation of FAST against other similarity measures.

6.4 Experiment 3: Benchmarking FAST

6.4.1 Methodology

With methodologies in place to determine the overall usefulness of a fuzzy measure in theory and to determine which the best structure to use the aim of Experiment 3 is to evaluate the actual performance of the FAST measure. This stage of the evaluation process served two key goals. Firstly it would determine the usefulness of FAST as a similarity measure on its own. It would also demonstrate whether FAST is more useful than other existing similarity measures when dealing with sentence pairs that have fuzzy components. The measures that FAST would be evaluated against would be STASIS (Li et al. 2006) and LSA (Landauer et al. 1998). These measures were chosen because of how widely used they are (K.O'Shea et al. 2008) (Osathanunkul 2011) (Yang et al. 2007) and for the fact that they have previously both been benchmarked against a human sentence similarity dataset.

The testing of the FAST sentence similarity measure against STASIS and LSA had to be done in two Stages.

Stage 1- Benchmarking of FAST against LSA and STASIS with the SFWD

Stage 2- Benchmarking of FAST against LSA and STASIS with the MFWD

The first stage 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. As the fact that fuzzy words affected sentence similarity had been established in Section 6.2 of this chapter, this stage of the evaluation allowed for the testing of how well FAST, STASIS and LSA were able to deal with the effect. To conduct this experiment the results that each of the measures returned when they took each sentence pair in the SFWS as input was considered. This, therefore, returned for each of the sentence pairs in the SFWD a similarity rating from FAST, STASIS and LSA. Each of the sets of results for each measure

would have a level of correlation with the human similarity ratings from the dataset. 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.

After the level of success FAST had at representing a single fuzzy component had been determined, the next stage was to determine if FAST's level of accuracy varied when it was presented with sentence pairs with multiple fuzzy words. The main goal of the second stage, therefore, was to determine if STASIS' performance remained consistent when presented with a greater number of fuzzy words or if it improved or was diminished. This also presented an opportunity to determine if increasing the number of fuzzy words further affected the accuracy of either STASIS or LSA. For this test, the MFWD was used. As with testing the SFWD, this procedure involved running each of the sentence pairs through from the dataset through each of the sentence similarity measures. Once that is done the correlations can then be examined. This also allows for the observation of the level of difference if any in terms of accuracy when compared to the results from the Single Fuzzy Word Dataset.

The ultimate goal of these tests was to clearly demonstrate the usefulness of FAST. If FAST was shown to have demonstrated a significant improvement over STASIS and LSA, then it serves as a strong candidate similarity measure for dealing with any pairs of fuzzy sentences. This was done through comparing the results of FAST against human ratings and comparing them to the results that were returned from FAST and STASIS .

6.4.2 Results and Discussion

Table 22 shows the comparison of FAST, STASIS and LSA in terms of the SWFD. It contains the average human ratings for each sentence pair, the standard deviation for each pair, and the similarity ratings for each pair returned by LSA, STASIS and FAST. In this evaluation as was the case in *Section 6.2*, the standard version of LSA was used. This was because of its

wide use and the fact that it was used in the initial benchmarking of STASIS (Deerwester et al. 1990) (Li et al. 2003).

SP	Average Human Rating	Standard Deviation	LSA	STASIS	FAST
1	3.833	2.021	0.48	0.75	0.719
2	0	0	0.01	0.468	0.474
3	7.3	1.995	0.26	0.671	0.778
4	7.952	1.85	0.84	0.747	0.744
5	1.281	2.43	0.02	0.555	0.555
6	8.719	1.002	0.95	0.627	0.627
7	7.095	1.737	0.63	0.854	0.848
8	6.719	1.762	0.81	0.78	0.779
9	0.952	1.8	0.49	0.616	0.616
10	8.248	1.008	0.46	0.708	0.825
11	4.957	1.489	0.49	0.413	0.406
12	0.529	0.978	0.32	0.488	0.477
13	3.286	2.57	0.05	0.565	0.605
14	6.371	1.827	0.93	0.924	0.891
15	9.138	0.892	1	1	1
16	6.781	1.81	0.7	0.844	0.898
17	3.229	2.386	0.59	0.321	0.501
18	2.11	1.995	0.61	0.497	0.514
19	6.757	2.212	0.79	0.779	0.782

20	8.986	0.784	0.36	0.824	0.836
21	3.548	3.24	0.28	0.545	0.545
22	8.852	1.45	0.42	0.882	0.902
23	7.043	1.623	0.8	0.859	0.891
24	3.833	2.296	0.39	0.707	0.713
25	8.857	0.964	0.72	0.742	0.769
26	7.583	1.835	0.96	0.867	0.919
27	8.919	1.076	0.71	0.708	0.795
28	6.914	2.016	0.88	0.862	0.862
29	1.295	2.211	0.16	0.385	0.385
30	6.624	2.398	0.48	0.534	0.574

Table 22 FAST, LSA and STASIS tested against SFWD

From these results it can be seen the FAST has an overall Pearsons Correlation level of 0.773 with human similarity ratings in the SWFD. STASIS and LSA correlation levels had been previously calculated at 0.707 and 0.644 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 therefore show that FAST can return sentence similarity from sentences in the SWFD with a high level of accuracy and that in terms of these sentences it can show a substantial improvement over both STASIS and 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.

The second part of the experiment involved testing FAST with the MFWD. Therefore the test used with the SWFD was repeated with the MFWD, the results of this are presented in Table 21.

SP	Average Human Rating	Standard Deviation	LSA	STASIS	FAST
1	5.623	2.934	0.66	0.868	0.904
2	1.715	2.059	0.72	0.402	0.588
3	3.769	2.27	0.82	0.726	0.944
4	0.75	1.621	-0.01	0.237	0.21
5	3.708	2.748	0.84	0.878	0.901
6	8.35	1.906	0.99	0.997	0.997
7	5.677	2.616	0.98	0.898	0.937
8	3.842	2.815	0.9	0.946	0.978
9	4.873	2.594	0.73	0.794	0.822
10	6.865	2.156	0.92	0.899	0.969
11	1.223	2.373	0.08	0.545	0.577
12	7.127	2.366	0.72	0.5	0.996
13	5.285	2.62	0.16	0.86	0.97
14	5.938	2.144	0.59	0.843	0.967
15	7.381	1.949	0.18	0.921	0.943
16	3.238	2.844	0.71	0.667	0.76
17	4.312	2.88	0.86	0.812	0.965
18	1.446	2.391	0.06	0.337	0.362

19	7.792	2.609	1	0.975	0.975
20	7.815	1.974	0.93	0.734	0.792
21	2.112	3.37	0.06	0.625	0.625
22	6.25	2.719	0.78	0.946	0.993
23	8.162	1.911	0.97	0.999	0.996
24	7.215	2.43	0.93	0.843	0.844
25	7.485	1.916	0.92	0.854	0.854
26	6.331	2.482	0.68	0.746	0.859
27	3.842	2.564	0.92	0.956	0.967
28	1.269	1.87	0.07	0.44	0.438
29	6.069	2.656	0.47	0.714	0.913
30	6.488	2.615	0.79	0.748	0.965

Table 23 FAST, LSA and STASIS tested against MFWD

From the results shown in Table 21 it can be seen that the correlation between FAST and the human ratings in the MFWD is 0.765 which is very close to its correlation with the SFWD. On the other hand the correlation between STASIS and the MFWD drops down to 0.685 while the level of correlation between LSA and the MFWD drops to 0.627. The decreases in the levels of accuracy from both STASIS and LSA were not significant implying that the increase in the number of fuzzy words in the sentence pairs did not substantially diminish their performance. As was the case from the first stage of the evaluation where the SFWD was tested, FAST continued to show a strong performance if using the work done by O'Shea et al. (2008b) as a benchmark for success. 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 measures that is dealing with them. 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 main conclusions that can be drawn from this experiment is that FAST shows a high level of accuracy in terms of dealing with fuzzy words and a notable improvement over both STASIS and LSA whose performances dropped substantially when compared to non-fuzzy words. This was demonstrated over both the SFWD and the MFWD. This therefore proves that the inter-relatedness between fuzzy words in a hierarchical structure must be considered if their effect on sentence similarity is to be clearly represented. This is strongly suggested, given the level of improvement shown by FAST which utilized such a hierarchical system. The results showed that the structure that held these relations that was used in FAST was able to accomplish this goal. The experiment therefore demonstrated that FAST can and should replace STASIS when sentences that contain fuzzy words are dealt with.

6.5 Conclusions

In conclusion, there were a number of different important points that were established during the evaluation process. Experiments were conducted to explore the different areas of the project and a number of different goals were accomplished. The final goal of the evaluation was to determine the overall effectiveness of FAST which was done through the evaluation process.

The first accomplishment of the evaluation was determining the effect (if any) that fuzzy words had on sentence similarity. Experiment 1 showed that there was a level of impact that was caused when fuzzy words were added to sentences. This was through both, a difference in human ratings between from the SFWD and the corresponding unfuzzified pairs in the O'Shea et al. Dataset and deterioration in the level of accuracy from both the STASIS and

LSA similarity measures. The two main points that this experiment proved were

- Concept A was demonstrated as being correct. That the presence of fuzzy words changed the semantic meanings of sentences enough to change human perceptions of the levels of similarity between them. This was demonstrated through a comparison of the differences in levels of human similarity between corresponding results in the SWFD and the O'Shea et al. Dataset
- Concept B was demonstrated as being correct. It was shown that the presence of fuzzy words in sentences affected the ability of existing semantic similarity measures to accurately represent the level of similarity between them. This was shown in the large differences in the performance of LSA and STASIS when dealing with non-fuzzy sentence pairs and when dealing with fuzzy sentence pairs

The second accomplishment was determining which of the fuzzy ontology structures (created in chapter 4) was best suited to a general implementation of FAST measure. Experiment 2 tested both of the structures through both the SWFD and the MWFD and determined which structure was able to return a higher correlation with similarity ratings. Both the structures returned very similar results in terms of the SWFD, with Structure 2 slightly outperforming Structure 1. However when the number of fuzzy words was increased with the MFWD, the ability of Structure 2 to accurately represent similarity dropped and Structure 1 had a higher of correlation. Therefore Structure 1 was determined to be the best option for a general implementation of the FAST sentence similarity measure. A potential reason for the improvement of Structure 1 over Structure 2 is the possibility that an excessive number of nodes can have an adverse effect on the usefulness of an ontology in terms of determining word similarity.

The final and most important accomplishment of the evaluation was to benchmark the FAST sentence similarity measure with the SFWD and the MFWD and to test it against both STASIS and LSA. The experiment showed

that FAST was able to outperform both STASIS and LSA in terms of both the datasets. Using the work that was done in the initial benchmarking of LSA and STASIS (O'Shea et al. 2008b) FAST was shown to have returned a high level of correlation with human similarity ratings while this was not that case with STASIS or LSA in both the datasets. There was shown to be no substantial difference in the performances of the similarity measures when presented with the MFWD and a greater number of fuzzy words per sentence. This was shown by their correlations remaining almost the same. While the accuracy of FAST remained high, the accuracy of STASIS began to slightly drop and the accuracy of LSA remained comparatively low. This therefore showed that FAST was a suited replacement to existing non fuzzy semantic similarity measures in the area of fuzzy sentences.

With FAST having now been fully assessed and having shown to be successful in terms measuring similarity between sentences with fuzzy words, it now has the potential to be used in practical applications. The next goal is to determine what areas FAST can make a contribution (such as Conversational Agents and Expert Systems (Ball and Breese 2000) (Fernandez et al. 2009) (Lee and Wang 2011) and how the measure can be implemented in those areas. This is discussed in detail in Chapter 7, where a full review of the project is presented.

7. Conclusion and Further Work

7.1. Overview

The research aims that were discussed in the introduction (Chapter 1) were met.

- A set of fuzzy words were collected and quantified using a methodology that was described in Chapter 3.
- An ontology-based fuzzy word similarity measure was created with a methodology described in Chapter 4.
- The FAST fuzzy sentence similarity measure was created using a methodology described in Chapter 4
- An evaluation dataset was created using a methodology described in Chapter 5
- At the end of the project a new sentence similarity measure (FAST) was built and was fully evaluated (the evaluation procedure was described in Chapter 6). This measure was able to accurately determine the level of similarity between two sets of text that contained one or more fuzzy words.

7.2. Summary of Work

This section contains an overview of the project, illustrating what its overall contributions to the field of Sentence Similarity and the overall field of computer science were. Furthermore this chapter also identifies areas of potential future work. These are areas where aspects of the research that has been done to build FAST (as well as the implementation of FAST itself) may be used for further scientific research or for implementation in practical applications.

The decision that the project required was the number of categories of fuzzy words that would be required. Based on a variety of factors (that are explored in Chapter 3), six categories were chosen. These categories would

hold all of the fuzzy words that would be quantified during the project. After the number of categories had been chosen the next issue that needed to be addressed was how to populate them with relevant fuzzy words. It was decided, based on previous work that had been done in word quantification, that they would be collected through human experimentation. After the words had been collected, further human experimentation was used to quantify the words. The end result at this stage was a set of six categories of quantified fuzzy words.

The next issues to be addressed in the project were how to create and implement a fuzzy word similarity measure and then how to use that word similarity measure to create a fuzzy sentence similarity measure (FAST). This process is covered in detail in Chapter 4. From the research that had been done into similarity measures it was decided that an ontology-based approach would be used (creating fuzzy ontologies for each of the fuzzy categories). The next decision that was required related to the nature of the ontology that would be used. Two different candidate ontology structures were created, with the goal of determining the one that allowed FAST to return a greater level of accuracy at the evaluation stage. Using the ontology structures, two implementations of the word similarity measure were created. The next issue to be addressed was implementing FAST using the word similarity measure. Using methods inspired by the STASIS algorithm, two implementations of FAST were created, one using each of the two implementations of the Fuzzy Word Similarity Measure.

At this point, was necessary to create an evaluation dataset for FAST as there was no suitable existing one. This required pairs of sentences with fuzzy components with human ratings on how similar to each other they were. It was decided that this new dataset would consist of two sub-datasets. They were a dataset of sentence pairs with a single fuzzy word in each sentence per pair, the Single Fuzzy Word Dataset (SFWD) and a sentence pair dataset with multiple fuzzy words per sentence per pair, the Multiple Fuzzy Word Dataset (MFWD). The SFWD was created through human experimentation based on work that was previously done in sentence similarity dataset creation. The MFWD was created through use of a novel

method of pairing that involved the automated extraction and fuzzification of sentence pairs and a novel method of quantification of sentence pairs through the use of crowd sourcing. With crowdsourcing there is a risk of manipulation of the data. These are risks such as users registering multiple accounts and performing the tests multiple times, unsuitable candidates taking tests and participants entering random answers (which is a particular risk given that there is an incentive to finish quickly). This necessitates precautions to be taken (as was done in the case of the MFWD development, where a Gold Standard was used and through using a crowdsourcing system that is able to determine that participants fulfil a set of criteria.

With FAST having been implemented and datasets having been created, it could now be evaluated. In the evaluation process a decision was made to determine the effect of fuzzy words on sentences and on the ability of pre-FAST sentence similarity measures to process sentence pairs that contained fuzzy sentences. This was done through comparing the SFWD to the dataset it was based on and then testing the STASIS and LSA measures against the SFWD. The next stage of the evaluation involved determining which ontological structure to use by testing both implementations of FAST against the SFWD and the MFWD. The final stage in the evaluation was to benchmark FAST against STASIS and LSA. This was done through testing all three measures against the MFWD and the SFWD.

7.3. Summary of Contributions

- **Creation of FAST**
- **Creation of Suitable Evaluation Dataset**
- **Evaluation and Benchmarking of FAST**

7.3.1. Quantification of a set of fuzzy words

Prior to the creation of a fuzzy sentence similarity measure, the relationships between different perception based words had to be established. This was because the measure would have needed to use these relationships to determine the level of similarity between any pair of words (which in turn was needed to determine sentence similarity). The first

main accomplishment of this research the generation of six categories of perception based words, with the words within each category quantified on a given scale. This was detailed in Chapter 3 with the most significant contributions being the generation of two important robust methodologies. The first methodology detailed how a fuzzy category could be populated with a set of fuzzy words through human experimentation (this methodology took inspiration from a number of sources (Rubenstein and Goodenough 1965) (Charles and Miller 1989) (Finkelstein et al. 2001) (Liu and Mendel 2008) with a particular focus on the work of James O'Shea (O'Shea et al. 2008a) (O'Shea 2010). The second methodology described how the fuzzy words within the categories could be quantified on a given scale. This allowed the words to be scaled against each other and as a result compared to each other. Beyond their use within the project, these methodologies have potential wider uses. If any future work requires the creation of a set of words that are related to a particular concept through human experimentation, the first methodology describes how this can be done. If a set of fuzzy words related to a given subject need to be quantified by human participants, the methodology presented in Chapter 3 shows how this can be done. This also includes cases where existing categories need to be expanded with more fuzzy words.

7.3.2. Creation of FAST

The second accomplishment was the actual creation of the FAST sentence similarity measure. This measure allowed a user to input two sentences (idealised towards 35 words or less but that is not a restriction) and it then returns a single value representing their overall level of similarity. This level of similarity based on a combination of the texts' semantic and syntactic levels of similarity. To build the FAST algorithm, two fuzzy ontology structures were created that contained the sets of fuzzy words and relations between them (based on the quantities that have been obtained in Chapter 3). These relationships were able to be used to find the similarity between pairs of fuzzy words in the structures (which did not exist in the WordNet ontology). These structures were each used in implementations of the

overall FAST sentence similarity measure. The FAST measure used ontological relations between words and corpus semantics to determine the total level of semantic similarity. The methodology used to implement it is described in Chapter 4. Furthermore, an additional algorithm was created to show the effect of fuzzy words on the level of similarity between the levels of similarity between non fuzzy words (described in Chapter 4) was implemented into FAST.

7.3.3. Creation of a suitable evaluation dataset

The third accomplishment of the project was the creation of a fuzzy sentence pair dataset that was needed to evaluate FAST. The reason that this was required was because that there had not been any previous sentence pair datasets that contained a suitable number of fuzzy words in the sentence to conduct a comprehensive test. Taking inspiration from and expanding on the pioneering work done in dataset creation by O'Shea et al. (2008a) and by James O'Shea himself in (2010), two methodologies for creating fuzzy datasets were made. This was to allow an overall dataset to be built from two sub datasets. The first sub dataset, called the Single Fuzzy Word Dataset (SFWD) was to contain sentence pairs with a single fuzzy word in each sentence pair (to determine if any effect occurs if fuzzy components are added to sentences). The second sub dataset is called the Multiple Fuzzy Word Dataset (MFWD), this is to determine the level of similarity between pairs of sentences with multiple fuzzy words in each sentence. Aside from allowing for the testing of FAST, the overall dataset (combining the MFWD and the SFWD) has wider applications. It can be used for the testing of any future sentence similarity measure which requires sentence pairs that contain fuzzy components. Furthermore the methodologies that have used to create the datasets can be used in future dataset construction. This is useful if any future research occurs which requires a new fuzzy dataset to be built for the task.

7.3.4. Evaluation and Benchmarking of FAST

The final accomplishment for the project was the thorough evaluation of FAST. Firstly it was shown through experimentation that not only do fuzzy

words change the semantic meanings of pairs of sentences enough to change their levels of similarity to each other but also that pre-FAST sentence similarity measures are unable to accurately represent the level of similarity between fuzzy pairs of sentences. It was then determined which of the two ontological structures that were created in the implementation of FAST would return the most accurate results. This allowed for a general implementation of FAST to be created. Finally, FAST was benchmarked against both the STASIS and LSA measures with both the SFWD and the MFWD. From the results FAST was shown to be more accurate than both STASIS and LSA and to have a high level of correlation with the results from the dataset. Therefore the main overall contribution from this stage of the project was evidence that the FAST similarity measure could accurately represent the effect of fuzzy words on sentence similarity.

7.4. Future Work

There are a number of different areas that the research done in creating FAST can be expanded into. This is most true in cases of systems that are required to make decisions based on human linguistic input. There are two particular areas of relevance in this case. They are conversational agents (Li et al. 2004) (Ruttkay and Pelechaud 2004) (K.O'Shea et al. 2008) and expert systems (Lee and Wang 2011) (Fernandez et al. 2009).

7.4.1. Improvements to FAST

Further work can be done to improve the overall performance of FAST. There are two areas that specifically can be addressed to allow FAST to return more accurate results, when tested against datasets of human similarity ratings (such as the SFWD and the MFWD). Firstly the number of domains that are represented in the ontology structures that FAST uses can be increased beyond its current six. Examples of these further domains are

- Brightness (Example words: Dark, Dim and Bright)
- Strength (Example words: Puny, Weak and Strong)
- Speed (Example words: Slow, Fast and Speedy)

This can be done through the methodologies that are presented in Chapter 3 and Chapter 4, which deal with the population of categories of fuzzy words, and implementing the words within them into FAST. This will allow FAST to represent words from the new subject areas as accurately as it does those from the current subject area. The expansion of FAST into the new domains allows it to represent the similarity between sentence pairs that contain words from these new domains. This allows FAST to represent the similarity of a substantially wider range of natural language sentences.

The second potential improvement to FAST comes from increasing the number of fuzzy words that are quantified through human ratings from the current set. While many words were quantified using the methodology that was used in Chapter 3, many more fuzzy words exist. The most accurate method of determining human perceptions for the words is through human experimentation. Therefore, to improve FAST, the Chapter 3 methodology could be used to collect and quantify more words for the categories. This could give more accurate values than the synonym based approach that was taken. After new words have been collected, they can be added to the existing ontological structures that FAST uses to determine similarity. This can be done through the methodology that was demonstrated in Chapter 4. Through this expansion of the categories, FAST could return higher similarity ratings in sentences that contain the new words that were added to the category. While the current manual method of data collection does produce good results, it can be very time consuming, particularly if exhaustive data about domains is to be collected. Therefore, an important area of future research is the automation of the process of collecting data and sorting it into its appropriate ontology. The development of a system that could generate fuzzy values for words with a comparable level of accuracy to human results could greatly increase the speed at which new values are developed for words and as such the speed at which new categories are built or existing ones expanded.

7.4.2. Utilising FAST Within Applications

There are a number of different areas that can be expanded into due to the research done in creating FAST. This is most true in cases of systems that are required to make decisions based on human linguistic input. There are two particular areas of relevance in this case. They are conversational agents (Li et al. 2004) (Ruttkay and Pelechaud 2004) and expert systems (Fernandez et al. 2009) (Lee and Wang 2011). Each will now be briefly discussed.

7.4.2.1. Expert Systems

Expert systems are systems that are able to provide intelligent answers to questions posed to them by humans that are interfacing with them. This is done through the system processing the input and attempting to derive the most successful response through exploration of a knowledge base (Lee and Wang 2011). Expert systems are generally very domain specific providing high level responses to complex issues in a particular area (Fernandez et al. 2009). This field has grown to encompass a wide range of subjects such as engineering (Wu and Mendel 2010) and medicine (Wyatt and Spiegelhalter 1988) (Lee and Wang 2011). A text similarity system is an integral part of expert systems that allows them to determine the best response to give to human input. Specifically, through matching human natural language input to similar text within its knowledge base, the system determines which response to give. Due to the limitations that existed within text similarity prior to this project, the naturalness of the human dialogue that could be used to interact with these systems was limited. This is because human natural language contains large numbers of fuzzy words. Therefore with the implementation of FAST, newer input mechanisms for expert systems can be developed, allowing users to interface with them using perception based words. As a result of this, the level naturalness of language that a person could successfully use in a FAST based expert system could be greatly increased over existing ones. This reduces the demands on the user and allows the system to return information that more

accurately represent the users' requirements. Therefore the development of fuzzy expert systems is a major potential area of future work.

7.4.2.2. Conversational Agents

Conversational Agents are agent based systems that are able to emulate human natural language dialogue, enabling them to converse with human users or with other conversational agents (Latham et al. a2010) (O'Shea 2012) (Chakraborti and Luger 2012). As was the case with expert systems, the linguistic choices conversational agents make in response to human input is based on results from a text similarity measure that it uses. Substantial work has been done in the development and integration of sentence similarity measures with conversational agents (O'Shea 2012) (Chakraborti and Luger 2012). Recently a framework was presented for the development of conversational agents using semantic similarity measures (O'Shea 2012). This was particularly illustrated through the demonstration of the usefulness of STASIS in this context. Therefore an opportunity for the development of more advanced conversational agents in the future that can converse more naturally with human users comes through the integration of FAST into them. Therefore the development of these new conversational agents which utilise a Fuzzy short text semantic similarity measure is a potent area for future development.

7.5. Concluding Remarks

In conclusion, the project has successfully addressed the issue of the inability of sentence similarity measures to accurately represent fuzzy words. This was done through the completion of a series of experiments that involved both human experimentation, algorithm creation and the development and implementation of working pieces of software. The final developed product, the FAST similarity measure, was thoroughly evaluated in chapter 6. The results of this evaluation demonstrated that fuzzy words do have an effect on sentence similarity and that FAST was able to represent sentence similarity in sentences with fuzzy words. Therefore, as has been discussed in this chapter, the FAST algorithm can be used in any relevant future work (with potential areas of future work having been identified and

discussed in this chapter). This is also true of the methodologies that have been developed in the creation of FAST.

Bibliography

Achananuparp, Palakorn, Xiaohua Hu, and Xiajiong Shen. "The evaluation of sentence similarity measures." Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, (2008). 305-316.

Agirre, Eneko, et al. "A study on similarity and relatedness using distributional and WordNet-based approaches." Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, (2009).

Androutsopoulos, Ion, and Prodromos Malakasiotis. "A survey of paraphrasing and textual entailment methods." arXiv preprint arXiv:0912.3747 (2009).

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. Vol. 463. New York: ACM press, (1999).

Baldauf, Matthias, Schahram Dustdar, and Florian Rosenberg. "A survey on context-aware systems." International Journal of Ad Hoc and Ubiquitous Computing 2.4 (2007): 263-277.

Baldauf, Matthias, Schahram Dustdar, and Florian Rosenberg. "A survey on context-aware systems." International Journal of Ad Hoc and Ubiquitous Computing 2.4 (2007): 263-277.

Ball, Gene, and Jack Breese. "Emotion and personality in a conversational agent." Embodied conversational agents (2000): 189-219.

Beckwith, Richard, Christiane Fellbaum, Derek Gross, and George A. Miller. "WordNet: A lexical database organized on psycholinguistic principles." *Lexical acquisition: Exploiting on-line resources to build a lexicon* (1991): 211-232.

Bigham, Jeffrey P. "Increasing web accessibility by automatically judging alternative text quality." Proceedings of the 12th international conference on Intelligent user interfaces. ACM, (2007).

Bilgin, Aysenur, Hani Hagras, Areej Malibari, Mohammed J. Alhaddad, and Daniyal Alghazzawi. "A general type-2 fuzzy logic approach for adaptive modeling of perceptions for Computing With Words." In *Computational Intelligence (UKCI), 2012 12th UK Workshop on*, pp. 1-8. IEEE, (2012).

Bilgin, Aysenur, Hani Hagras, Areej Malibari, Mohammed J. Alhaddad, and Daniyal 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*, pp. 1-8. IEEE, (2012).

Bilgin, Aysenur, Hani Hagras, Areej Malibari, Mohammed J. Alhaddad, and Daniyal 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*, pp. 1-8. IEEE, (2012).

Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, (2006).

Black, William, Sabri Elkateb, and Piek Vossen. "Introducing the Arabic wordnet project." In *In Proceedings of the third International WordNet Conference (GWC-06*. (2006).

Bobillo, Fernando, and Umberto Straccia. "Fuzzy ontology representation using OWL 2." International Journal of Approximate Reasoning 52.7 (2011): 1073-1094.

Bollegala, Danushka, Yutaka Matsuo, and Mitsuru Ishizuka. "Measuring semantic similarity between words using web search engines." www 7 (2007): 757-766.

Boyd-Graber, Jordan, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. "Adding dense, weighted connections to WordNet."

In *Proceedings of the third international WordNet conference*, pp. 29-36.(2006).

Brill, Eric. "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging." Computational linguistics21.4 (1995): 543-565.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. "Word-sense disambiguation using statistical methods." In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pp. 264-270. Association for Computational Linguistics, (1991).

Budanitsky, Alexander, and Graeme Hirst. "Evaluating wordnet-based measures of lexical semantic relatedness." Computational Linguistics 32.1 (2006): 13-47.

Budanitsky, Alexander, and Graeme Hirst. "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures." Workshop on WordNet and Other Lexical Resources. Vol. 2. (2001).

Budiu, Raluca, and John R. Anderson. "Interpretation-based processing: A unified theory of semantic sentence comprehension." Cognitive Science 28.1 (2004): 1-44.

Cabrerizo, Francisco Javier, Sergio Alonso, and Enrique Herrera-Viedma. "A consensus model for group decision making problems with unbalanced fuzzy linguistic information." International Journal of Information Technology & Decision Making 8.01 (2009): 109-131.

Callison-Burch, Christopher. Paraphrasing and Translation. Diss. University of Edinburgh, (2008).

Cao, Yongzhi, and Guoqing Chen. "A fuzzy petri-nets model for computing with words." Fuzzy Systems, IEEE Transactions on 18.3 (2010): 486-499.

Carvalho, Vitor R., Matthew Lease, and Emine Yilmaz. "Crowdsourcing for search evaluation." ACM Sigir forum. Vol. 44. No. 2. ACM, (2011).

Cassell, Justine, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents." In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 413-420. ACM, (1994).

Chakrabarti, Chayan, and George F. Luger. "A semantic architecture for artificial conversations." Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on. IEEE, (2012).

Charles, Walter G. "Contextual correlates of meaning." Applied Psycholinguistics 21.4 (2000): 505-524.

Charles, Walter G., and George A. Miller. "Contexts of antonymous adjectives."Applied psycholinguistics 10.03 (1989): 357-375.

Charniak, Eugene. "A maximum-entropy-inspired parser." Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics, 2000. Morgan Kaufmann Publishers Inc., (2000).

Clancey, William J. "Heuristic classification." Artificial intelligence 27.3 (1985): 289-350.

Corcho, Oscar, Mariano Fernández-López, and Asunción Gómez-Pérez. "Methodologies, tools and languages for building ontologies. Where is their meeting point?." Data & knowledge engineering 46.1 (2003): 41-64.

Crockett, Keeley A., Zuhair Bandar, and Akeel Al-Attar. "Soft decision trees: a new approach using non-linear fuzzification." Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on. Vol. 1. IEEE, (2000).

Crockett, Keeley, Zuhair Bandar, David Mclean, and James O'Shea. "On constructing a fuzzy inference framework using crisp decision trees." *Fuzzy sets and systems* 157, no. 21 (2006): 2809-2832.

Das, Dipanjan, and Noah A. Smith. "Paraphrase identification as probabilistic quasi-synchronous recognition." In *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: Volume 1-Volume 1*, pp. 468-476. Association for Computational Linguistics, (2009).

David Chandran, Keeley Crockett, David McLean, Zuhair Bandar. "FAST: A Fuzzy Sentence Similarity Measure". IEEE International Conference on Fuzzy Systems (2013)

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. "Indexing by latent semantic analysis."*JASIS* 41, no. 6 (1990): 391-407.

Dolan, Bill, Chris Quirk, and Chris Brockett. "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, (2004).

Dolan, William B., and Chris Brockett. "Automatically constructing a corpus of sentential paraphrases." Proc. of IWP. (2005).

Duong, Trong Hai, Ngoc Thanh Nguyen, and Geun Sik Jo. "A method for integration of wordnet-based ontologies using distance measures." Knowledge-Based Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, (2008).

Eisele, Andreas, and Yu Chen. "MultiUN: A Multilingual Corpus from United Nation Documents." LREC. (2010).

Fellbaum, Christiane. WordNet. Springer Netherlands, (2010).

Fernández, Alberto, María José del Jesus, and Francisco Herrera. "On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets." Expert Systems with Applications 36.6 (2009): 9805-9812.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. "Placing search in context: The concept revisited." In *Proceedings of the 10th international conference on World Wide Web*, pp. 406-414. ACM, (2001).

Foltz, Peter W., Darrell Laham, and Thomas K. Landauer. "Automated essay scoring: Applications to educational technology." World Conference on Educational Multimedia, Hypermedia and Telecommunications. Vol. 1999. No. 1. (1999a).

Foltz, Peter W., Darrell Laham, and Thomas K. Landauer. "The intelligent essay assessor: Applications to educational technology." Interactive Multimedia Electronic Journal of Computer-Enhanced Learning 1.2 (1999b).

Francis, W. Nelson. "A standard corpus of edited present-day American English." College English 26.4 (1965): 267-273.

Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis." IJCAI. Vol. 7. (2007).

Gasir, Fathi, Zuhair Bandar, and Keeley Crockett. "The effect of datasets characteristic in the induction of fuzzy regression trees using Elgasir." Int. Arab Conf. Information Technology. (2010).

Gildea, Daniel, and Daniel Jurafsky. "Automatic labeling of semantic roles."Computational linguistics 28.3 (2002): 245-288.

Godil, Saniya Siraj, Muhammad Shahzad Shamim, Syed Ather Enam, and Uvais Qidwai. "Fuzzy logic: A “simple” solution for complexities in neurosciences?." *Surgical neurology international* 2 (2011).

Gonzalo, Julio, Felisa Verdejo, Irina Chugur, and Juan Cigarran. "Indexing with WordNet synsets can improve text retrieval." *arXiv preprint cmp-lg/9808002*(1998).

Gruber, Thomas R. "A translation approach to portable ontology specifications."Knowledge acquisition 5.2 (1993): 199-220.

Gruber, Thomas R. "The role of common ontology in achieving sharable, reusable knowledge bases." KR 91 (1991): 601-602.

Gruber, Thomas R. "Toward principles for the design of ontologies used for knowledge sharing?." International journal of human-computer studies 43.5 (1995): 907-928.

Grüninger, Michael, and Mark S. Fox. "Methodology for the Design and Evaluation of Ontologies." (1995).

Haase, Peter, Jeen Broekstra, Marc Ehrig, Maarten Menken, Peter Mika, Mariusz Olko, Michal Plechawski et al. "Bibster–a semantics-based bibliographic peer-to-peer system." In *The Semantic Web–ISWC 2004*, pp. 122-136. Springer Berlin Heidelberg,(2004).

Hand, David J. "Principles of data mining." Drug safety 30.7 (2007): 621-622.

Hart, Michael. Project gutenberg. Project Gutenberg, (1971).

Hayes-Roth, Frederick, Donald Waterman, and Douglas Lenat. "Building expert systems." (1984).

Hearst, Marti A. "Automated discovery of WordNet relations." WordNet: an electronic lexical database (1998): 131-151.

Herrera, Francisco, Enrique Herrera-Viedma, and Luis Martínez. "A fuzzy linguistic methodology to deal with unbalanced linguistic term sets." Fuzzy Systems, IEEE Transactions on 16.2 (2008): 354-370.

Herrera, Francisco, Sergio Alonso, Francisco Chiclana, and Enrique Herrera-Viedma. "Computing with words in decision making: foundations,

trends and prospects." *Fuzzy Optimization and Decision Making* 8, no. 4 (2009): 337-364.

Herrera-Viedma, Enrique, Gabriella Pasi, Antonio G. Lopez-Herrera, and Carlos Porcel. "Evaluating the information quality of web sites: A methodology based on fuzzy computing with words." *Journal of the American Society for Information Science and Technology* 57, no. 4 (2006): 538-549.

Herrera-Viedma, Enrique, Francisco Chiclana, Francisco Herrera, and Sergio Alonso. "Group decision-making model with incomplete fuzzy preference relations based on additive consistency." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 37, no. 1 (2007): 176-189.

Higuchi, Shinsuke, Rafal Rzepka, and Kenji Araki. "A casual conversation system using modality and word associations retrieved from the web."Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, (2008).

Ho, Chukfong, Masrah Azrifah Azmi Murad, Rabiah Abdul Kadir, and Shyamala C. Doraisamy. "Word sense disambiguation-based sentence similarity." In*Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 418-426. Association for Computational Linguistics, (2010).

Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, (1999).

Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." Machine learning 42.1-2 (2001): 177-196.

Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews."Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2004).

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. "Improving word representations via global context and multiple word prototypes." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873-882. Association for Computational Linguistics, (2012).

Ide, Nancy, and Jean Véronis. "Introduction to the special issue on word sense disambiguation: the state of the art." Computational linguistics 24.1 (1998): 2-40.

Islam, Aminul, and Diana Inkpen. "Real-word spelling correction using Google Web IT 3-grams." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics, (2009).

Islam, Aminul, and Diana Inkpen. "Semantic text similarity using corpus-based word similarity and string similarity." ACM Transactions on Knowledge Discovery from Data (TKDD) 2.2 (2008): 10.

Jan, Dusan, Antonio Roque, Anton Leuski, Jacki Morie, and David Traum. "A virtual tour guide for virtual worlds." In *Intelligent Virtual Agents*, pp. 372-378. Springer Berlin Heidelberg, (2009).

Jiang, Jay J., and David W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy." arXiv preprint cmp-lg/9709008 (1997).

Joachims, Thorsten. Text categorization with support valuemachines: Learning with many relevant features. Springer Berlin Heidelberg, (1998).

Kacprzyk, Janusz, and Slawomir Zadrozny. "Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation." Fuzzy Systems, IEEE Transactions on 18.3 (2010): 461-472.

Karnik, Nilesh N., and Jerry M. Mendel. "Centroid of a type-2 fuzzy set."Information Sciences 132.1 (2001): 195-220.

Karnik, Nilesh N., Jerry M. Mendel, and Qilian Liang. "Type-2 fuzzy logic systems." Fuzzy Systems, IEEE Transactions on 7.6 (1999): 643-658.

Karnik, Nilesh Naval, and Jerry M. Mendel. "Type-2 fuzzy logic systems: type-reduction." Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on. Vol. 2. IEEE, (1998).

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. "Crowdsourcing user studies with Mechanical Turk." Proceedings of the SIGCHI conference on human factors in computing systems. ACM, (2008).

Kopp, Stefan, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. "A conversational agent as museum guide–design and evaluation of a real-world application." In *Intelligent Virtual Agents*, pp. 329-343. Springer Berlin Heidelberg,(2005).

Kopp, Stefan, Jens Allwood, Karl Grammer, Elisabeth Ahlsen, and Thorsten Stocksmeier. "Modeling embodied feedback with virtual humans." In *Modeling communication with robots and virtual humans*, pp. 18-37. Springer Berlin Heidelberg, (2008).

Kraft, Reiner, Chi Chao Chang, Farzin Maghoul, and Ravi Kumar. "Searching with context." In *Proceedings of the 15th international conference on World Wide Web*, pp. 477-486. ACM,(2006).

Laird, John E., Allen Newell, and Paul S. Rosenbloom. "Soar: An architecture for general intelligence." Artificial intelligence 33.1 (1987): 1-64.

Landauer, Thomas K., and Susan T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological review 104.2 (1997): 211.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.

Latham, Annabel M., Keeley A. Crockett, David A. McLean, Bruce Edmonds, and Karen O'Shea. "Oscar: An intelligent conversational agent tutor to estimate learning styles." In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pp. 1-8. IEEE, (2010).

Latham, Annabel, Keeley Crockett, David Mclean, and Bruce Edmonds. "A conversational intelligent tutoring system to automatically predict learning styles." *Computers & Education* 59, no. 1 (2012): 95-109.

Lauritzen, Steffen L., and David J. Spiegelhalter. "Local computations with probabilities on graphical structures and their application to expert systems."Journal of the Royal Statistical Society. Series B (Methodological) (1988): 157-224.

Leacock, Claudia, and Martin Chodorow. "Combining local context and WordNet similarity for word sense identification." WordNet: An electronic lexical database 49.2 (1998): 265-283.

Lee, Chang-Shing, and Mei-Hui Wang. "A fuzzy expert system for diabetes decision support application." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41.1 (2011): 139-153.

Lee, Chang-Shing, Mei-Hui Wang, and Hani Hagras. "A type-2 fuzzy ontology and its application to personal diabetic-diet recommendation." Fuzzy Systems, IEEE Transactions on 18.2 (2010): 374-395.

Lee, Chang-Shing, Zhi-Wei Jian, and Lin-Kai Huang. "A fuzzy ontology and its application to news summarization." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 35.5 (2005): 859-880.

Lemaire, Benoit, and Philippe Dessus. "A system to assess the semantic content of student essays." Journal of Educational Computing Research 24.3 (2001): 305-320.

Lesk, Michael. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream

cone." Proceedings of the 5th annual international conference on Systems documentation. ACM, (1986).

Lewis, David D., and W. Bruce Croft. "Term clustering of syntactic phrases."Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, (1989).

Li, Yuhua, Zuhair Bandar, David McLean, and James O'Shea. "A Method for Measuring Sentence Similarity and iIts Application to Conversational Agents." In *FLAIRS Conference*, pp. 820-825. (2004).

Li, Yuhua, David McLean, Zuhair A. Bandar, James D. O'shea, and Keeley Crockett. "Sentence similarity based on semantic nets and corpus statistics."*Knowledge and Data Engineering, IEEE Transactions on* 18, no. 8 (2006): 1138-1150.

Li, Yuhua, Zuhair A. Bandar, and David McLean. "An approach for measuring semantic similarity between words using multiple information sources."*Knowledge and Data Engineering, IEEE Transactions on* 15, no. 4 (2003): 871-882.

Lin, Dekang. "An information-theoretic definition of similarity." ICML. Vol. 98. (1998).

Lin, Dekang. "An information-theoretic definition of similarity." ICML. Vol. 98. (1998).

Lin, Feiyu, and Kurt Sandkuhl. "A survey of exploiting wordnet in ontology matching." Artificial Intelligence in Theory and Practice II. Springer US, 2008. 341-350.

Lin, Tsau Young, Yiyu Y. Yao, and Lotfi A. Zadeh, eds. Data mining, rough sets and granular computing. Vol. 95. Springer, (2002).

Liu, Feilong, and Jerry M. Mendel. "Encoding words into interval type-2 fuzzy sets using an interval approach." IEEE Transactions on Fuzzy Systems 16.6 (2008): 1503.

Liu, Feilong. "An efficient centroid type-reduction strategy for general type-2 fuzzy logic system." Information Sciences 178.9 (2008): 2224-2236.

Liu, Jun, Luis Martinez, Hui Wang, Rosa M. Rodriguez, and Vasily Novozhilov. "Computing with words in risk assessment." *International Journal of Computational Intelligence Systems* 3, no. 4 (2010): 396-419.

MacMahon, Matt, Brian Stankiewicz, and Benjamin Kuipers. "Walk the talk: Connecting language, knowledge, and action in route instructions." Def 2.6 (2006): 4.

Madnani, Nitin. "Getting started on natural language processing with Python."Crossroads 13.4 (2007): 5-5.

Maedche, Alexander, and Steffen Staab. Ontology learning. Springer Berlin Heidelberg, (2004).

Marcus, Andrian, and Jonathan I. Maletic. "Recovering documentation-to-source-code traceability links using latent semantic indexing." Software Engineering, 2003. Proceedings. 25th International Conference on. IEEE, (2003).

Marcus, Andrian, Andrey Sergeyev, Vaclav Rajlich, and Jonathan I. Maletic. "An information retrieval approach to concept location in source code." In *Reverse Engineering, 2004. Proceedings. 11th Working Conference on*, pp. 214-223. IEEE, (2004).

Martinez, Luis, Da Ruan, and Francisco Herrera. "Computing with words in decision support systems: an overview on models and applications."International Journal of Computational Intelligence Systems 3.4 (2010): 382-395.

McNamara, Timothy P., James K. Hardy, and Stephen C. Hirtle. "Subjective hierarchies in spatial memory." Journal of Experimental Psychology: Learning, Memory, and Cognition 15.2 (1989): 211.

McNamara, Timothy P., Robert J. Sternberg, and James K. Hardy. "Processing verbal relations." Intelligence 15.2 (1991): 193-221.

Mendel, Jerry M. "Fuzzy logic systems for engineering: a tutorial." Proceedings of the IEEE 83.3 (1995): 345-377.

Mendel, Jerry M. "Advances in type-2 fuzzy sets and systems." Information Sciences 177.1 (2007a): 84-110.

Mendel, Jerry M. "Computing with words and its relationships with fuzzistics."Information Sciences 177.4 (2007): 988-1006.

Mendel, Jerry M. "Computing with words: Zadeh, Turing, Popper and Occam."Computational Intelligence Magazine, IEEE 2.4 (2007a): 10-17.

Mendel, Jerry M. "On KM Algorithms for Solving Type-2 Fuzzy Set Problems." (2012): 1-1.

Mendel, Jerry M. "Type-2 fuzzy sets and systems: an overview." Computational Intelligence Magazine, IEEE 2.1 (2007b): 20-29.

Mendel, Jerry M., and Hongwei Wu. "New results about the centroid of an interval type-2 fuzzy set, including the centroid of a fuzzy granule." Information Sciences 177.2 (2007b): 360-377.

Mendel, Jerry M., and RI Bob John. "Type-2 fuzzy sets made simple." Fuzzy Systems, IEEE Transactions on 10.2 (2002): 117-127.

Mendel, Jerry M., Robert Ivor John, and Feilong Liu. "Interval type-2 fuzzy logic systems made simple." Fuzzy Systems, IEEE Transactions on 14.6 (2006): 808-821.

Mendel, Jerry, et al. "What computing with words means to me [discussion forum]." Computational Intelligence Magazine, IEEE 5.1 (2010): 20-26.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." AAAI. Vol. 6. 2006.

Miller, George A. "Dictionaries in the Mind." Language and Cognitive Processes1.3 (1986): 171-185.

Miller, George A. "Nouns in WordNet: a lexical inheritance system."International journal of Lexicography 3.4 (1990): 245-264.

Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

Miller, George A., and Christiane Fellbaum. "Semantic networks of English."Cognition 41.1 (1991): 197-229.

Miller, George A., and Walter G. Charles. "Contextual correlates of semantic similarity." Language and cognitive processes 6.1 (1991): 1-28.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. "Introduction to wordnet: An on-line lexical database*."*International journal of lexicography* 3, no. 4 (1990): 235-244.

Mitchell, Jeff, and Mirella Lapata. "Value-based Models of Semantic Composition." ACL. (2008).

Mizumoto, Masaharu, and Kokichi Tanaka. "Some properties of fuzzy sets of type 2." Information and control 31.4 (1976): 312-340.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. "Crowdsourcing and language studies: the new generation of linguistic data." In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122-130. Association for Computational Linguistics,(2010).

N Karnik, Nilesh, and Jerry M Mendel. "Operations on type-2 fuzzy sets."Fuzzy Sets and Systems 122.2 (2001): 327-348.

Niles, Ian, and A. Pease. Mapping WordNet to the SUMO ontology. Teknowledge Technical Report, (2003).

Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." (2001).

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) 413 pages.

O'Shea, James, Zuhair Bandar, Keeley Crockett, and David McLean. "A comparative study of two short text semantic similarity measures." In *Agent and Multi-Agent Systems: Technologies and Applications*, pp. 172-181. Springer Berlin Heidelberg, (2008a).

O'Shea, James, Zuhair Bandar, Keeley Crockett, and David McLean. "Pilot short text semantic similarity benchmark data set: Full listing and description."*Computing* (2008b).

O'Shea, Karen, Zuhair Bandar, and Keeley Crockett. "A novel approach for constructing conversational agents using sentence similarity measures."Proceedings of the World Congress on Engineering. Vol. 1. (2008).

O'Shea, Karen. "An approach to conversational agent design using semantic sentence similarity." Applied Intelligence 37.4 (2012): 558-568.

Osathanunkul, Khukrit, et al. "Semantic similarity measures for the development of Thai dialog system." Agent and Multi-Agent Systems: Technologies and Applications. Springer Berlin Heidelberg, (2011). 544-552.

Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis."Foundations and trends in information retrieval 2.1-2 (2008): 1-135.

Parry, David. "A fuzzy ontology for medical document retrieval." Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation-Volume 32. Australian Computer Society, Inc., (2004).

Patwardhan, Siddharth, and Ted Pedersen. "Using WordNet-based context values to estimate the semantic relatedness of concepts." Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. Vol. 1501. (2006).

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. "WordNet:: Similarity: measuring the relatedness of concepts." Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics, (2004).

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. "WordNet:: Similarity: measuring the relatedness of concepts." Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics, (2004).

Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections." Proceedings of the 17th international conference on World Wide Web. ACM, (2008).

Rada, Roy, Hafedh Mili, Ellen Bicknell, and Maria Blettner. "Development and application of a metric on semantic nets." *Systems, Man and Cybernetics, IEEE Transactions on* 19, no. 1 (1989): 17-30.

Rajati, Mohammad Reza, and Jerry M. Mendel. "Novel Weighted Averages versus normalized sums in computing with words." Information Sciences (2013).

Reformat, Marek, and Cuong Ly. "Ontological approach to development of computing with words based systems." International Journal of Approximate Reasoning 50.1 (2009): 72-91.

Reisinger, Joseph, and Raymond J. Mooney. O'Shea, James, et al. "Benchmarking short text semantic similarity."International Journal of Intelligent Information and Database Systems 4.2 (2010): 103-120.

Resnik, Philip Stuart. "Selection and information: a class-based approach to lexical relationships." IRCS Technical Reports Series (1993): 200.

Resnik, Philip, and Mona Diab. "Measuring verb similarity." Proceedings of the 22nd Annual Meeting of the Cognitive Science Society. (2000).

Resnik, Philip. "Disambiguating noun groupings with respect to WordNet senses." Natural Language Processing Using Very Large Corpora. Springer Netherlands, (1999). 77-98.

Resnik, Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." arXiv preprint arXiv (2011):1105.5444 .

Resnik, Philip. "Using information content to evaluate semantic similarity in a taxonomy." arXiv preprint cmp-lg/9511007 (1995).

Rodríguez, M. Andrea, and Max J. Egenhofer. "Determining semantic similarity among entity classes from different ontologies." Knowledge and Data Engineering, IEEE Transactions on 15.2 (2003): 442-456.

Rubenstein, Herbert, and John B. Goodenough. "Contextual correlates of synonymy." Communications of the ACM 8.10 (1965): 627-633.

Rubin, Stuart H. "Computing with words." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 29.4 (1999): 518-524.

Ruttkay, Zsófia, and Catherine Pelachaud, eds. From brows to trust: evaluating embodied conversational agents. Vol. 7. Springer, (2004).

Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

Salton, Gerard, and Michael E. Lesk. "Computer evaluation of indexing and text processing." Journal of the ACM (JACM) 15.1 (1968): 8-36.

Schmidt, Desmond, and Robert Colomb. "A data structure for representing multi-version texts online." International Journal of Human-Computer Studies67.6 (2009): 497-514.

Schütze, Hinrich. "Automatic word sense discrimination." Computational linguistics 24.1 (1998): 97-123.

Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks." In *Proceedings of the conference on empirical methods in natural language processing*, pp. 254-263. Association for Computational Linguistics, (2008).

Stoilos, Giorgos, Giorgos Stamou, and Jeff Z. Pan. "Fuzzy extensions of OWL: Logical properties and reduction to fuzzy description logics." International Journal of Approximate Reasoning 51.6 (2010): 656-679.

Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. "Knowledge engineering: principles and methods." Data & knowledge engineering 25.1 (1998): 161-197.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: A large ontology from wikipedia and wordnet." Web Semantics: Science, Services and Agents on the World Wide Web 6.3 (2008): 203-217.

Tanaka, Hideo, Peijun Guo, and H-J. Zimmermann. "Possibility distributions of fuzzy decision variables obtained from possibilistic linear

programming problems." Fuzzy Sets and Systems 113.2 (2000): 323-332.

Tellex, Stefanie, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. "Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation." In *AAAI*. (2011).

Tho, Quan Thanh, Siu Cheung Hui, Alvis Cheuk M. Fong, and Tru Hoang Cao. "Automatic fuzzy ontology generation for semantic web." *Knowledge and Data Engineering, IEEE Transactions on* 18, no. 6 (2006): 842-856.

Toral, Antonio, Oscar Ferrández, Eneko Agirre, Rafael Munoz, Informatika Fakultatea, and Basque Country Donostia. "A study on Linking Wikipedia categories to Wordnet synsets using text similarity." In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*, pp. 449-454. (2009).

Tsatsaronis, George, Iraklis Varlamis, Michalis Vazirgiannis, and Kjetil Nørvåg. "Omiotis: A thesaurus-based measure of text relatedness." In *Machine Learning and Knowledge Discovery in Databases*, pp. 742-745. Springer Berlin Heidelberg, (2009).

Tsatsaronis, George, Iraklis Varlamis, and Michalis Vazirgiannis. "Text relatedness based on a word thesaurus." Journal of Artificial Intelligence Research 37.1 (2010): 1-40.

Turney, Peter D., and Michael L. Littman. "Measuring praise and criticism: Inference of semantic orientation from association." ACM Transactions on Information Systems (TOIS) 21.4 (2003): 315-346.

Turney, Peter. "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL." (2001).

Uschold, Michael, and Michael Gruninger. "Ontologies and semantics for seamless connectivity." ACM SIGMod Record 33.4 (2004): 58-64.

Uschold, Mike, and Michael Gruninger. "Ontologies: Principles, methods and applications." Knowledge engineering review 11.2 (1996): 93-136.

Voorhees, Ellen M. "Using WordNet to disambiguate word senses for text retrieval." Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, (1993).

Wang, L-X., and Jerry M. Mendel. "Generating fuzzy rules by learning from examples." Systems, Man and Cybernetics, IEEE Transactions on 22.6 (1992): 1414-1427.

Wang, Yingxu, Yousheng Tian, and Kendal Hu. "Semantic manipulations and formal ontology for machine learning based on concept algebra." International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 5.3 (2011): 1-29.

Wang, Yingxu. "On concept algebra for Computing with Words (CWW)."International Journal of Semantic Computing 4.03 (2010): 331-356.

Waterman, Donald. "A guide to expert systems." (1986).

Widyantoro, Dwi H., and John Yen. "A fuzzy ontology-based abstract search engine and its user studies." In *Fuzzy Systems, 2001. The 10th IEEE International Conference on*, vol. 3, pp. 1291-1294. IEEE, (2001).

Wu, Dongrui, and Jerry M. Mendel. "Aggregation using the linguistic weighted average and interval type-2 fuzzy sets." Fuzzy Systems, IEEE Transactions on15.6 (2007): 1145-1161.

Wu, Dongrui, and Jerry M. Mendel. "Computing with words for hierarchical decision making applied to evaluating a weapon system." Fuzzy Systems, IEEE Transactions on 18.3 (2010): 441-460.

Wu, Dongrui, and Jerry M. Mendel. "Uncertainty measures for interval type-2 fuzzy sets." Information Sciences 177.23 (2007): 5378-5393.

Wyatt, Jeremy, and David Spiegelhalter. "Evaluating Medical Expert Systems: What To Test, And How?." Knowledge Based Systems in Medicine: Methods, Applications and Evaluation. Springer Berlin Heidelberg, (1991). 274-290.

Yager, Ronald R. "A procedure for ordering fuzzy subsets of the unit interval."Information Sciences 24.2 (1981): 143-161.

Yager, Ronald R. "Fuzzy subsets of type II in decisions." Cybernetics and System 10.1-3 (1980): 137-159.

Yager, Ronald R., and Lotfi A. Zadeh. An introduction to fuzzy logic applications in intelligent systems. Kluwer Academic Publishers, 1992.

Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." ICML. Vol. 97. 1997.

Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods." Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1995.

Yeh, Eric, et al. "WikiWalk: random walks on Wikipedia for semantic relatedness." Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. Association for Computational Linguistics, 2009.

Zadeh, Lotfi A. "From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions." Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on 46.1 (1999): 105-119.

Zadeh, Lotfi A. "Fuzzy logic= computing with words." Fuzzy Systems, IEEE Transactions on 4.2 (1996): 103-111.

Zadeh, Lotfi A. "Fuzzy sets." Information and control 8.3 (1965): 338-353.

Zadeh, Lotfi A. "Is there a need for fuzzy logic?." Information Sciences 178.13 (2008): 2751-2779.

Zadeh, Lotfi A. "Outline of a new approach to the analysis of complex systems and decision processes." Systems, Man and Cybernetics, IEEE Transactions on 1 (1973): 28-44.

Zadeh, Lotfi A. "Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems." Soft Computing-A fusion of foundations, methodologies and applications 2.1 (1998): 23-25.

Zadeh, Lotfi A. "The concept of a linguistic variable and its application to approximate reasoning—I." Information sciences 8.3 (1975): 199-249.

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 (1997): 111-127.

.

Appendix 1:

Words Collection Questionnaire

Words Experiment

INSTRUCTIONS:Please look at the following six categories. For each of the categories, take each word and state all the words that you feel have similar meanings. For example if you feel Cool has a similar meaning to the word Cold, please write it under the section where Cold is. Please 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"). If you need any additional paper to add more words, please ask and some will be provided. With each category, an example is provided to give a clearer picture of context).

CATEGORY 1: DISTANCE/SIZE (e.g. It is far away)

Near

- - -
-

- - -
-

Far

- - - -

- - - -

Tiny

- - - -

- - - -

Small

- - - -

- - - -

Medium

- - - -

- - - -

Large

- - - -

- - - -

Huge

-	-	-	-
-	-	-	-

Category 2: Temperature (e.g. The water is cold)

Freezing

- - -
-

- - - -

- - - -

Cold

- - - -

- - - -

- - - -

Lukewarm

- - - -

- - - -

- - - -

Hot

- - - -

- - - -

- - - -

Boiling

- - - -

- - - -

- - - -

Category 3: Goodness (e.g. That music is average)

Awful

-	-	-	-
-	-	-	-
-	-	-	-

Bad

-	-	-	-
-	-	-	-
-	-	-	-

Mediocre

-	-	-	-
-	-	-	-
-	-	-	-

Good

-	-	-	-
-	-	-	-
-	-	-	-

Excellent

-	-	-	-
-	-	-	-
-	-	-	-

Category 4: Age (e.g. That particular version is quite young)

Infantile

-	-	-	-
-	-	-	-
-	-	-	-

Young

-	-	-	-
-	-	-	-
-	-	-	-

Middle-aged

-	-	-	-
-	-	-	-
-	-	-	-

Old

-	-	-	-
-	-	-	-
-	-	-	-

Ancient

-	-	-	-
-	-	-	-
-	-	-	-

Category 5: Frequency (He comes in here infrequently)

Rarely

-	-	-	-
-	-	-	-
-	-	-	-

Infrequently

-	-	-	-
-	-	-	-
-	-	-	-

Sometimes

-	-	-	-
-	-	-	-
-	-	-	-

Commonly

-	-	-	-
-	-	-	-
-	-	-	-

Often

-	-	-	-
-	-	-	-
-	-	-	-

Category 6: Membership in a group (e.g. This grade is just enough to warrant a pass)

Just

-	-	-	-
-	-	-	-
-	-	-	-

Hardly

-	-	-	-
-	-	-	-
-	-	-	-

Somewhat

-	-	-	-
-	-	-	-
-	-	-	-

Largely

-	-	-	-
-	-	-	-
-	-	-	-

Mostly

-	-	-	-
-	-	-	-
-	-	-	-

Appendix 2:
Words Quantification Questionnaire

PhD Experiments to Test the Quantifiability of Fuzzy Words

By David Chandran

As part of my PhD I am trying to investigate words with no precise meaning. This experiment consists of 7 categories of commonly used words in the English language

Please read the following instructions

For each category:

-Try to imagine the extremes of the category

-Read every word in the list

-Try to imagine each word in a sentence

-With each category is a scale of 0 to 10 with 0 being the minimum and 10 being the maximum, try to place each word on the given scale.

Feel free to give multiple words the same value on any scale, also please note that words may appear more than once over different categories

Please return all questionnaires to your instructor/lecturer

If you have any questions please email D.Chandran@mmu.ac.uk

CATEGORY : Distance

Scale: 0 to 10 with 0 representing the closest possible distance and 10 representing the furthest

Word	**Rating**
Adjacent	
Alongside	
Average	
Big	
Close	
Diminutive	
Distant	
Dwarf	
Enormous	
Far	
Gargantuan	
Giant	
Gigantic	
Great	
Huge	
Insignificant	
Large	
Little	
Massive	
Medium	
Microscopic	
Middle	

Miniscule	
Minute	
Near	
Nearby	
Normal	
Petite	
Proximal	
Proximate	
Regular	
Remote	
Sizeable	
Small	
Standard	
Substantial	
Tiny	

Category: Size

Scale: 0 to 10 with 0 representing the smallest possible size and 10 representing the largest possible size

Word	Rating
Average	
Big	
Close	
Diminutive	
Dwarf	
Enormous	
Gargantuan	
Giant	
Gigantic	
Great	
Huge	
Insignificant	
Large	
Little	
Massive	
Medium	
Microscopic	
Miniscule	
Minute	
Middle	
Normal	
Petite	

Regular	
Small	
Substantial	
Tiny	

Category: Temperature

Scale: 0 to 10 with 0 representing the coldest possible temperature and 10 representing the hottest possible temperature

Word	**Rating**
Arctic	
Baking	
Biting	
Bitter	
Body-temperature	
Boiling	
Brisk	
Burning	
Chilly	
Cold	
Cool	
Freezing	
Frigid	
Frosty	
Frozen	
Hot	
Icy	
Lukewarm	
Mild	
Nippy	
Roasting	
Scalding	

Scorching	
Spicy	
Steaming	
Sub-zero	
Sweaty	
Sweltering	
Temperate	
Tepid	
Warm	

Category: Goodness

Scale: 0 to 10 with 0 representing the worst possible case and 10 representing the best possible case

Word	**Rating**
Acceptable	
Alright	
Amazing	
Appalling	
Average	
Awful	
Bad	
Boring	
Brilliant	
Dire	
Dreadful	
Enjoyable	
Excellent	
Fair	
Fantastic	
Fine	
Good	
Great	
Inadequate	
Marvellous	
Mediocre	

Middling	
Nice	
Ok	
Passable	
Pathetic	
Pleasant	
Poor	
Rotten	
Splendid	
Superb	
Terrible	
Unacceptable	
Unbearable	
Unsatisfactory	
Useless	
Wonderful	

Category: Age

Scale: 0 to 10 with 0 representing the smallest possible age and 10 representing the oldest possible age

Word	Rating
Adolescent	
Adult	
Aged	
Ancient	
Antiquated	
Antique	
Archaic	
Baby	
Babyish	
Child	
Childish	
Child-like	
Decrepit	
Elderly	
Experienced	
Full-grown	
Grown-up	
Immature	
Infantile	
Juvenile	
Mature	
Middle-aged	

New	
Old	
Pensionable	
Pre-historic	
Pre-pubescent	
Recent	
Young	
Youthful	

Category: Frequency

Scale: 0 to 10 with 0 representing and event never happening and 10 representing and the event constantly happening

Word	**Rating**
Always	
Barely	
Commonly	
Consistently	
Constantly	
Daily	
Frequently	
Habitually	
Hardly	
Infrequently	
Never	
Normally	
Occasionally	
Often	
On-Occasion	
Periodically	
Rarely	
Regularly	
Repeatedly	
Scarcely	
Seldom	

Somewhat	
Uncommonly	
Unpredictably	
Usually	

Category: Fullness/Closeness to completion

Scale: 0 to 10 with 0 representing something completely empty or unstarted and 10 representing something complete or full

Word	**Rating**
Adequate	
Almost	
Average	
Barely	
Bit	
Generally	
Greatly	
Halfway	
Hardly	
Just	
Largely	
Little	
Mainly	
Middling	
Mostly	
Partially	
Rather	
Scarcely	
Scraping	
Somewhat	
Sufficient	

Appendix 3:
Sentence Generation Questionnaire

Thank you for participating in this experiment. This experiment is part of a PhD project, researching sentence similarity. Please read each of the sentence pairs individually. For each sentence pair you will be asked to add a single word to either increase or diminish a particular aspect of the sentence, or change an existing adjective/adverb within the sentence to achieve the same effect (if inserting more words is required to ensure the sentence remains grammatically correct please do so but please try to use as few as possible). This is to be done if you feel it is possible to do so. If any further clarification is required please email me (David Chandran) at d.chandran@mmu.ac.uk

Example 1:

Increase or diminish, if possible, the size of the entity in the sentence

-There is a hill in the distance that we have to climb

-There is a large mountain in the Lake District

The size of the hill can be increased by adding a word and the size of the word mountain can be increased through changing a word.

-There is a small hill in the distance that we have to climb

-There is a huge mountain in the Lake District

Example 2

Increase or diminish, if possible the temperatures of the consumables in the sentence

-Earlier this afternoon, I had some soup for lunch

-After I finished jogging, I had a glass of water

The temperatures of the soup and the water can be changed through adding a word

-Earlier this afternoon, I had some hot soup for lunch

-After I finished jogging, I had a glass of cold water

1) Increase or diminish, if possible, the level of delay
-When I was going out to meet my friends there was a delay at the train station
-The train operator announced to the passengers on the train that there would be a delay

2) Increase or diminish, if possible, the severity of the punishment
-You must realize that you will definitely be punished if you play with the alarm
-He will absolutely be punished for setting the fire alarm off

3) Increase or diminish, if possible, the level of humour
-I will make you laugh so hard that your sides ache and split
-When I tell you this you will split your sides laughing

4) Increase or diminish, if possible, the level of regret
-I offer my condolences to the parents of John Smith, who was unfortunately murdered
-I extend my sympathy to John Smith's parents, following his murder

5) Increase or diminish, if possible, the effect of the product
-If you continuously use these products, I guarantee you will look young
-I assure you that, by using these products over a long period of time, you will appear youthful

6) Increase or diminish, if possible, the size of the lime wedge
-I always like to have a slice of lemon in my drink, especially if it's coke.
-I like to put a wedge of lemon in my drinks, especially cola

7) Increase or diminish, if possible, the length of the journey
-We got home safely in the end, though it was a long journey
-Though it took many hours travel on the long journey, we finally reached our house safely

8) Increase or diminish, if possible, the size of the diamond
-A man called Dave gave his fiancée a diamond ring for their engagement
-The man presented a diamond to the woman and asked her to marry him

9) Increase or diminish, if possible, the level of concern
-Global warming is what everyone is worrying about today
-The problem of global warming is a concern to every country in the world at the moment.

10) Increase or diminish, if possible, the time element
-Midday is 12 o'clock in the middle of the day
-Midday is 12 o'clock in the middle of the day

11) Increase or diminish, if possible, the temperature of the cup
-The first thing I do in a morning is make myself a cup of coffee
-The first thing I do in the morning is have a cup of coffee

12) Increase or diminish, if possible, the size of the hill
-Meet me on the hill behind the church in half an hour

-Join me on the hill at the back of the church in 30 minutes

13) Increase or diminish, if possible, the level of pleasure
-It gives me pleasure to announce the winner of this year's beauty pageant
-It's a pleasure to tell you who has won our annual beauty parade

14) Increase or diminish, if possible, the distance of the drive
-Will I have to drive far to get to the nearest petrol station?
-Is it much farther for me to drive to the next gas station?

15) Increase or diminish, if possible, the distance of somewhere
-I think I know her from somewhere, because she has a familiar face.
-You have a very familiar face; do I know you from somewhere?

16) Increase or diminish, if possible, the amount of work

-I am sorry but I can't go out as I have a heap of work to do
-I've a heap of things to finish so I can't go out I'm afraid

17) Increase or diminish, if possible, the recentness of the sofa purchase
-Get that wet dog off my brand new sofa
-Make that wet hound get off my white couch –I only recently bought it

18) Increase or diminish, if possible, the quality of wine
-Would you like to drink this wine with your meal?
-Will you drink a glass of wine while you eat?

19) Increase or diminish, if possible, the size of the tree
-Could you climb up the tree and save my cat from jumping please?
-Can you get up that tree and rescue my cat, otherwise it might jump

20) Increase or diminish, if possible, the level of hunger
-I am so hungry I could eat a whole horse plus desert
-I could have eaten another meal, I'm still starving

21) Increase or diminish, if possible, the level of the of cover up
-You shouldn't be covering what you feel
-There is no point covering up what you said, we all know.

22) Increase or diminish, if possible, the size of the supernatural entity
-The ghost appeared from nowhere and frightened the old man.
-The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.

23) Increase or diminish, if possible, the number of people invited
-I have invited a variety of people to my party so it should be interesting
-A number of invitations were given out to a variety of people inviting them down the pub

24) Increase or diminish, if possible, the level of age.
-Because I am the eldest one, I should be more responsible
-Just because of my age, people shouldn't think I'm a responsible adult, but they do.

25) Increase or diminish, if possible, the quality of the car
-That's not a very good car, on the other hand mine is great.
-This is a terrible noise level for a new car.

26) Increase or diminish, if possible, the niceness of the entity described.
-Does music help you to relax or does it distract you too much
-Does this sponge look dry to you?

27) Increase or diminish, if possible, the frequency of the problem described.
-The key doesn't seem to be working, can you give me another?
-I dislike the word quay, it confuses me, I always think of the thing for locks, there's another one.

28) Increase or diminish, if possible, the size of the entity being described

-There was a heap of rubble left by the builders outside my house this morning

-Sometimes in a crowd accidents may happen, which can cause deadly injuries.

29) Increase or diminish, if possible, the recentness of the entity being described.

-I bought a guitar today, do you like it?

-The weapon choice reflects the personality of the carrier

30) Increase or diminish, if possible, the size of the entities being discussed

-Boats come in all shapes and sizes but they all do the same thing

-Chairs can be comfy and not comfy, depending on the chair.

Appendix 4:

Semantic Sentence Similarity Questionnaire

Sentence Semantic Similarity Questionnaire

Background information –please read before you start doing the task.

Thank you for volunteering to take part in this scientific study, in the field of semantic sentence similarity

You may still withdraw before starting the task or at any point while doing it.

You are provided with a set of pairs of sentences and a recording page to write your judgements on. These pairs appear in a random order.

We want you to rate the similarity of meaning of these sentence pairs

What do we mean by similarity of meaning?

To judge similarity of meaning you should look at the two sentences and as yourself “How close do these two sentences come to meaning the same thing?”

In other words

How close do they come to making you believe the same thing?
How close do they come to making you feel the same thing? Or
How close do they come to making you do the same thing?

You will be asked to rate each of the sentence pairs based on their level of similarity of meaning. The rating scale runs from 0.0 (minimum similarity) to 10.0 (maximum similarity) please do not use values greater than 10.0

Please note that this study does not evaluate you in any way –there are no “right” or “wrong” answers, except in the sense that the right answer to each question is an accurate expression of your personal opinion.

Please return completed questionnaires, as well as any problems, comments or questions to:
David Chandran
Room E113
School of Computing, Mathematics and Digital Technology
John Dalton Building
Manchester Metropolitan University
Manchester
M1 5GD
Or Email:
d.chandran@mmu.ac.uk

DISCLAIMER: The answers to the questions about yourself will be kept for no longer than three months after the first results are published. The similarity ratings you provide will be separated from the personal data kept

permanently. This is because data can be useful in long term studies. We will never disclose your personal information to anyone outside the project. The similarity ratings will be used in publications on an international scale.

Instructions

1) Please read each of the sentence pairs

EXAMPLE SENTENCE PAIR:

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.

2) For each of the sentence pairs determine how similar in meaning you think they are to each other

3) On the recording sheet rate each of the sentence pairs based on their level of similarity of meaning. The rating scale runs from 0.0 (minimum similarity) to 10.0 (maximum similarity) please do not use values greater than 10.0

4) Please then complete the personal information form

RECORDING SHEET

Please record the levels of similarity for each sentence pair on the table below. Please rate each pair on a scale of 0 (minimum similarity) to 10 (maximum similarity). You can use the first decimal place, for example if you think that similarity is half way between 3.0 and 4.0 you can use a value like 3.5

Sentence Pair	Similarity Rating
SP 1	
SP 2	
SP 3	
SP 4	
SP 5	
SP 6	
SP 7	
SP 8	
SP 9	
SP 10	
SP 11	
SP 12	
SP 13	
SP 14	
SP 15	

Sentence Pair	Similarity Rating
SP 16	
SP 17	
SP 18	
SP 19	
SP 20	
SP 21	
SP 22	
SP 23	
SP 24	
SP 25	
SP 26	
SP 27	
SP 28	
SP 29	
SP 30	

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 utmost 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.

Personal Data Sheet

Please enter the following items of personal information

Your Name (Print):

Your highest educational qualification c (including subject):

Are you a native English speaker? (Y/N):

Please return to:

David Chandran
Room E113
School of Computing, Mathematics and Digital Technology
John Dalton Building
Manchester Metropolitan University
Manchester
M1 5GD
Or Email:
d.chandran@mmu.ac.uk

Appendix 5:

Copy of “FAST: A Fuzzy Sentence Similarity Measure”. IEEE International Conference on Fuzzy Systems (2013)

FAST: A Fuzzy Semantic Sentence Similarity Measure

David Chandran, Keeley Crockett, David Mclean, Zuhair Bandar
The Intelligent Systems Group, School of Computing, Mathematics and Digital Technology,
The Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
Telephone (+44) (0) 161 247 1497
Facsimile (+44) (0) 161 247 1483
Email: D.chandran@mmu.ac.uk

***Abstract*—A problem in the field of semantic sentence similarity is the inability of sentence similarity measures to accurately represent perception based (fuzzy) words that are commonly used in natural language. This paper presents a new sentence similarity measure that attempts to solve this problem. The new measure, Fuzzy Algorithm for Similarity Testing (FAST) 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. Through human experimentation fuzzy sets were created for six categories of words based on their levels of association with particular concepts. These fuzzy sets were then defuzzified and the results used to create new ontological relations between the words. Using these relationships allows for the creation of a new ontology based semantic text similarity algorithm that is able to show the effect of fuzzy words on computing sentence similarity as well as the effect that fuzzy words have on non-fuzzy words within a sentence. Experiments on FAST were conducted using a new fuzzy dataset, the creation of which is described in this paper. The results of the evaluation showed that there was an improved level of correlation between FAST and human test results over two existing sentence similarity measures.**

Keywords- semantic similarity measures, computing with word, ontology, FAST

I. INTRODUCTION

A sentence similarity measure is an algorithm that is able to compare two or more blocks of text and return a level of similarity between them. Early sentence similarity measures (SSMs) were based on the premise of determining similarity based on the comparison of syntax [1]. The first SSM that was able to factor in the level of semantic similarity was the seminal Latent Semantic Analysis (LSA) system [2]. Using a corpus based approach LSA was able to specifically determine the level of semantic similarity between two sets of text. This system worked through the analysis of corpus statistics, taking words in two blocks of text and referencing them from within a large corpus. Generating statistics based on the occurrences of these words in the corpus allowed the creation of semantic vectors to determine the level of similarity between the compared sets of text. A weakness that was identified with LSA however was that it was designed to deal primarily with large sets of text rather than short texts (sentences with a length less than thirty words) [2][3][4]. In [3] a text similarity measure called STASIS was proposed for representing the level of similarity between short pieces of text by determining the level of similarity between two sentences through the use of ontological relations between words using an existing word similarity measure which was created in [4]. This word similarity measure expanded on earlier work to determine relationships between concepts and entities [5][6]. It worked by looking at the distances between a pair of words in an ontology and the distance between both those words and their closest subsumer. From this, the word similarity measure was able to return a level of semantic similarity for those two words. STASIS uses this word similarity measure between all possible pairs of words (1 from each) from the two texts. The levels of similarity from these word calculations are used in conjunction with corpus statistics from the Brown corpus [7] to create a semantic vector. This semantic vector is used with a syntactic vector, which is created through the positions of words in the texts, to return an overall level of similarity for the two sentences. In [8] a dataset was created showing human similarity ratings between pairs of sentences based on the definitions of words in Rubenstein and Goodenough's dataset [9]. Both LSA and STASIS were compared against these sentence pairs and while both measures had a high correlation with human ratings, STASIS was shown to be closer to the human ratings than LSA.

A weakness however, that existed in such SSMs was their inability to accurately represent fuzzy words (words that represent subjective quantities based on an individual's perceptions) [3]. Given the wide use of fuzzy words in natural language this limits the strength of these measures in the areas where they are practically applied, for example the field of conversational agents (which involve real time conversation between humans and computers) [10] where the goal is to emulate the naturalness of human conversation. The representation of fuzzy words is an important step in improving these measures.

This paper presents a new algorithm which aims to solve this problem, known as FAST (Fuzzy Algorithm for Similarity Testing). FAST is designed, to be able to represent the effect fuzzy words in a sentence or short text have on the overall levels of semantic sentence similarity. FAST is able to determine the levels of similarity between sets of fuzzy words through calculating the similarity between pairs of fuzzy words. This new algorithm works through taking concepts from the field of Fuzzy Logic and computing with words [11], using fuzzy sets to quantify words. The system makes use of a dataset of quantified fuzzy words (which needs to be created as no such dataset presently exists). The focus of this paper is to describe, the creation of the new semantic similarity algorithm FAST, and show through experimentation how successfully it was able to deal with the issue of fuzzy words in short texts (such as sentences). A comparative experimental study is presented that involved testing FAST against STASIS and LSA to see how well it compared in terms of correlation with human results

This paper is organized as follows; Section II provides an overview of relevant concepts in fuzzy theory and computing with words, Section III describes the creation of new datasets containing fuzzy words, Section IV describes the process of building a dataset of quantified fuzzy words. Section V describes the methodology for building FAST. Section VI discusses the creation of datasets to evaluate FAST. Experimental results and accompanying discussion are covered in Section VII and VII and finally section IX presents the conclusions

II. COMPUTING WITH WORDS AND FUZZY

In [11], Lofti Zadeh identified an issue in human-computer communication. While computers communicate with each other using crisp quantities, humans tend to communicate information to each other using perception based (fuzzy) words that are subjective. To deal with this issue Zadeh created a new framework called Computing with Words (CWW) through which these words could be communicated to computer systems. This framework can be used to solve the problem of creating an SSM that can represent human perceptions. In [11] Zadeh stated that perception based (fuzzy) words would cover a range of values, effectively being represented by a fuzzy set. He introduced the concept of granularity, discussing the fact that different concepts have different levels of association with a particular core concept. In [11],[12] he showed how entities could have different levels of association with a particular concept (for example, if we were to consider the concept of hotness, “hot” would have a higher level of association with the concept than “lukewarm”). As such, the values covered by a group of fuzzy words could theoretically be represented in terms of a concept they were associated with.

Further expansion on Zadeh’s work came from Jerry Mendel who noted that perceptions around words differed from individual to individual [13][14]. Consider the illustration (fig1) presented by Mendel [13] regarding the word “some”, with the horizontal axis representing values for the word and the y axis representing their membership functions

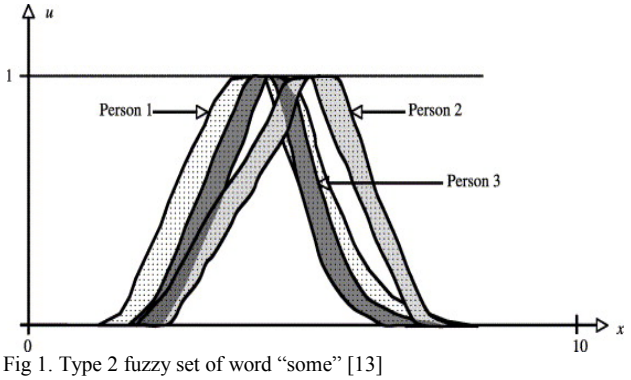

Fig 1. Type 2 fuzzy set of word “some” [13]

Mendel stated that a single word would have a value represented by a set of fuzzy sets rather than by a single one. Mendel proposed a fuzzy type-2 based solution [13], based on previous work described in [14] and [15]. A type-2 fuzzy set is a different type of fuzzy set to a type-1. A fuzzy type-2 is a set wherein all its elements are fuzzy type-1 sets. This allows it to represent the range of different perceptions about a particular word. In order to convert a fuzzy type-2 set into a fuzzy type-1 Mendel had proposed using the centroid of a type-2 set [16][17] for type reduction (projecting the type-2 set to a type-1 set). This allowed the varying perceptions that people had about a fuzzy word (the level of uncertainty) to be represented as a type 1 fuzzy set. This could then be defuzzified, to return a single value. Through this Mendel and Liu created a methodology for creating type-2 fuzzy sets for a group of fuzzy words [18]. This resulted in a codebook which contained a series of quantified fuzzy words based on their levels of association with a concept.

The concepts in Liu and Mendel’s codebook could be used to quantify a large set of fuzzy words to be used in the fuzzy sentence similarity algorithm. Crisp quantities could be calculated by creating type-2 fuzzy sets for a series of words on a given scale and then type-reducing and defuzzifying them. This could be used to create a dataset of fuzzy words. Liu and Mendel’s approach in the codebook, collected ranges of values for words for individuals and then took their centroids towards type reduction. The method presented here instead asks individuals to return single values for each word based on what value they consider to be most representative of that word. This is because, given that a much larger number of words were now being used, asking a range of individuals for a range of values could prove too onerous a task. The closeness of the results collected through the method described in Section V of this paper with Liu and Mendel’s results [18] can be determined through comparing the common words from the codebook.

III. CREATING BENCHMARK DATASETS

At present there is no suitably large dataset of quantified fuzzy words or sentences containing such words. In order to

build and test a fuzzy similarity measure both of these must be constructed. Aside from the methodology that was put forward by Liu and Mendel that allowed quantities to be derived for fuzzy words, a methodology for acquiring levels of similarities between sets of words was put forward by Rubenstein and Goodenough [9] [19]. Here, a dataset of word pair similarities was developed and used to evaluate word measures such as the one used in STASIS [3]. Furthering the work of Rubenstein and Goodenough, O'Shea designed a methodology and constructed an unbiased benchmark sentence similarity dataset [8] [3]. The results of this dataset were used in testing a number of SSM including the STASIS measure [3]. The methodology [8] detailed how sentences could be generated from groups of people. His method involved test subjects creating sentences based on prompts and guide words (to ensure that sentences relevant to desired topics were generated), pairing the sentences and then collecting similarity ratings for the pairs. The methods put forward in the O'Shea dataset can be used for the generation of sets of fuzzy words. These can be quantified for use in creating fuzzy ontologies that can be used in FAST and, later, for generating sets of fuzzy sentences to evaluate FAST. Furthermore, they can be used to ensure that the experiments are conducted in an unbiased manner.

IV. Building a Dataset of Quantified Fuzzy Words

A. Overview of creating a New Dataset

Prior to any work on creating a new sentence similarity measure the issue of word similarity between fuzzy words, for the same concept had to be resolved. For example, to calculate the similarity between "This is a big tree." and "That is a small house" we must first calculate the relationship between the words "big" and "small". To do this, the words need to be scaled against each other. A methodology was developed to quantify sets of fuzzy words on particular scales. The methodology involved creating a set of categories to contain fuzzy words, populating those categories with fuzzy words and then quantifying the fuzzy words against each other based on their level of association with a particular category. In this section the steps of this methodology are presented.

B. Collecting and Categorising a Set of Words

Six categories of fuzzy words were created. When Zadeh first described Computing with Words in [11], he talked in detail about 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 category due to the large overlap between their members in terms of fuzzy words. Four other broad categories were also selected. They were Goodness, Frequency, Temperature and Completeness (which were selected for the number of fuzzy words they could cover). This represented a substantial increase over the single category that was presented in Liu and Mendel's codebook. Once the categories had been determined, the next phase was the population of the categories with fuzzy words.

Collecting the category words involved asking a group of twenty native English speakers to return a questionnaire that asked them to write down as many words as they could think of from the different categories. For example on the category of "temperature", they were asked to write down all the adjectives that they could think of that related to levels of temperature. To ensure that that there was a wide range of words with different values across the categories O' Shea's concept of guide words was used [8]. Guide words (words that could act as prompts) were used at different points of a scale of size across each category.

C. Quantifying a Set of Fuzzy Words.

After the category words had been collected they needed to be quantified. This involved collecting values for each category word from a group of human subjects. These values are then used to make a fuzzy set for that category word. Through the defuzzification of the set a single crisp value can be returned. It was decided that the questionnaire, for the quantification portion of the experiment, would ask respondents to rate words in each category on a scale of 0 to 10. The words would be rated based on their levels of association with the highest point in that category. For example, in the temperature scale words would be rated based on their level of association with the maximum possible temperature conceivable. These points are taken as the highest membership functions for those words, this differs from Liu and Mendel's where the centroids were taken instead. A total of 20 questionnaires were distributed to native English speakers, two were filled in erroneously leaving 18 completed questionnaires.

The union of the 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. This was done through taking the mean of these ratings. The usefulness of this crisp value is then determined by looking at the standard deviation of the members of the set. If a low level of standard deviation exists, the implication is that there is a tendency towards that value with the highest membership function. If on the other hand, the standard deviation is high, the implication would be that there is no such tendency and taking the centroid of a range would have been better and that other defuzzification methods would need to be considered. The results were crisp defuzzified values for each word in each of the categories. Through a review of the standard deviations of the values of the words, it was observed that in most of cases the standard deviation was less than 2.00. From comparing common words contained in both the size category and Mendel's codebook there was a very high correlation (0.99) between the results collected using this method and the means of the centroids from Mendel's method .There was also a very small average standard deviation of 0.51. The quantities attributed to the words were used to determine the relationships between these words and create fuzzy ontologies with them.

V. A METHODOLOGY FOR BUILDING FAST

A. Overview

This section describes the new algorithm called FAST. The fundamental building block of FAST is the STASIS algorithm which is an existing and well recognized SSM [4]. A brief overview of STASIS will be provided in section *B*. The first step of FAST used the words that had been quantified (as described in section IV) to create fuzzy ontologies (described in section VI) for each category. These ontologies would be used to determine the relations between words within the same category. Relationships between words from different categories are not defined (for example the relationship between "cold" and "good" is not known). Adapting the STASIS formula to these relationships delivers a similarity for pairs of fuzzy words. Furthermore the effect that fuzzy words have on non fuzzy words can be determined through the relationships between fuzzy words (a separate algorithm has been created that determines what these associated words are which is discussed in Section *E*). This section will describe the methodology of FAST and describe its main components: the creation of fuzzy ontologies and a fuzzy word similarity measure; development of an algorithm that determines the association of non-fuzzy words with fuzzy words the effect of fuzzy words on non-fuzzy words' levels of similarity.

B. Overview Of STASIS

The STASIS algorithm is adapted to deal with non-fuzzy words, corpus statistics and syntactic similarity [4]. STASIS [4] takes two sets of text as input. Every pair of words in the texts is referenced in the WordNet ontology [26]. Their path length, l, (the length of the shortest path between them) and their depth h, (the subsumer depth) are then retrieved. The level of similarity between the words (w_1 and w_2) is determined with equation (1):

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

The parameters α and β, based on calculations done in [3] and [4] take on the values of 0.2 and 0.6 respectively.
These similarity values are taken along with word frequency information and information on word positions from a short joint word set vector (represented as r in the following equation) to determine the total level of similarity between the two sentences (T_1 and T_2). Overall similarity is calculated using equation (2), with δ being defined as the total sum of all possible values and s1 and s2 referring to pairs semantic similarity vectors which were determined in (1).

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

C. FAST

This section provides an overview of the FAST algorithm. The pseudo code for FAST can be found in Figure 2.

```
Tag every word in the two sentences
Pair every combination of tagged words
For every word pair:
If A and B are both fuzzy words:
    If A and B are in the same category:
            Reference Subsumer Depth from Fuzzy ontology
            Reference Length between words from Fuzzy ontology (described in section X)
            Using these values, calculate Level of Similarity with formula (1), the STASIS word similarity formula
            Return Level of similarity (on a scale of 0 to 10)
    Else:
        Apply STASIS word similarity measure [3]
    End If
  Return Level of similarity
Else
    Apply STASIS word similarity measure.
    Determine presence of fuzzy words associated with the non-fuzzy words (described in section E).
    If Associated Fuzzy Words are Present:
      Calculate Subsumer Depth and length modifications using the process (described in section E).
      Recalculate Word Similarity
      Return Level of Similarity
    Else:
      Return level of similarity.
    End If
End If
Apply Corpus statistics [4]
Next
Determine Syntactic similarity [4]
Determine Total similarity using formula (2)
```

Fig 2. Pseudo code for FAST measure

This pseudo code describes how the algorithm deals with pairs of words (in terms of calculating their semantic similarity). Once the similarities for all the words are calculated, FAST uses the same method as STASIS to determine overall similarity. For every pair of words, the FAST algorithm determines if they are fuzzy or not (based on their presence in any of the categories). If they are fuzzy but do not belong to the same category the WordNet based method that STASIS uses determines their level of similarity. If they are present in the same category, then the algorithm calculates their level of similarity based on their subsumer depth and distance from each other in that category's ontology using the formula presented in [4]. Once the similarities for all the pairs of words are calculated (given the corpus statistics and syntactic similarity are calculated separately, with the methods discussed in [3]) the total level of similarity can be determined using equation (2). The creation of these fuzzy ontologies is described in section *D*. The similarity between non fuzzy words is

calculated with the existing STASIS word similarity measure. If these words have associated fuzzy words, then their level of similarity is amended using the ontological relations between fuzzy words (as is discussed in Section *E)*

D. Creating a Fuzzy Ontology

A fuzzy ontology structure was created from the fuzzy words that had been quantified (as described in Section IV). This allows the similarity between fuzzy words to be determined using their depths and lengths in the same manner as STASIS' word similarity component (described in part B of this section) This ontology structure would fill a role akin to the WordNet ontology [20] in terms of being used to provide distances between fuzzy words as well as distances to a common subsumer. This ontology allows similarity values to be generated for any pair of fuzzy words (from the same category) in the same way that WordNet generates numbers for non-fuzzy words.

To create a fuzzy ontology, WordNet is used as a template. This is because of WordNet's wide use, particularly for sentence similarity. Relationships in WordNet are determined through entities belonging to others branching away from a central point. This structure is replicated in creating a fuzzy ontology, with nodes containing sets of words branching away from a central node based on differences in quantity from that node. The first step was to divide each category (as identified in Section IV) into nodes that were related to each other through subsumer relations. The division of categories in this manner, allows for sets of words (from the categories) to be stored within these nodes. This allows for relations between these words to 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 (i.e. with size the nodes were "very small", "small", "average", "large" and "very large"). It should be noted that the names of nodes are not necessarily words contained in those nodes. For example "very large" and "very small" are not contained within the size ontology. At this point the issue is then, how the set of fuzzy words are classified into the correct node. To classify words to the correct node they need to be re-scaled to reflect them moving away from a central point (representing the top subsumer node). In this research the words were, based on their quantities, re-scaled on a -1 to 1 scale with the midpoint representing a value of 0. This allowed all the words to allocated to a node (for example for the size category, all words with values of -1 to -0.6 were put in the very small node, vales between -0.6 to -0.2 put in the small node etc.).

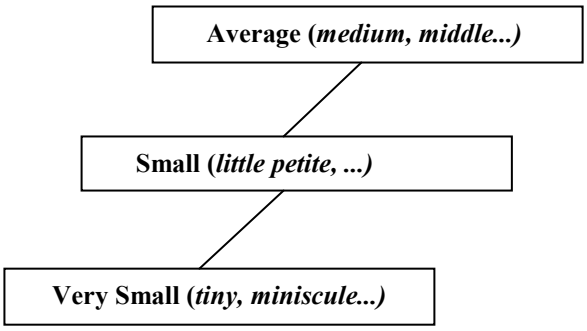

Fig 3. Portion of ontology for the category "size"

E. Determining the Affect of Fuzzy Words on Non Fuzzy Words

In addition to word relations there is another area wherein fuzzy words affect overall sentence similarity which is discussed in this section, the effect of fuzzy words on non fuzzy words. Most fuzzy words in a sentence have associated non fuzzy words whose meanings they can affect. For example consider the following two words:

"Mountain" and "Hill"

There is a given semantic vector between them. Now consider addition of two fuzzy words creating the phrases:

"Small Mountain" and "Big Hill"

This addition causes the vector to change. This is because the addition of the two fuzzy words has altered the level of semantic similarity between the non-fuzzy ones. This section describes how FAST deals with this problem.

The first stage in representing the effect of fuzzy words on non-fuzzy words was in determining which pairs of words were associated. An algorithm was implemented for this purpose. For each sentence, the algorithm tagged each of the words according to type (e.g. noun, verb, adjective, etc.) Given that a vast majority of fuzzy words that could affect other words are either adjectives or adverbs, the system was designed to find associated words to these word types. The algorithm found the associated word based on locations within the sentence through running a series of rules. After the implementation of the system it was tested for its accuracy. This was done through taking three random articles from a newspaper and running all the sentences from them into the system. This showed that the algorithm was able to correctly identify the associated fuzzy words with non-fuzzy words in 80% of the sentences where fuzzy words were present. This therefore allowed for this algorithm to be used in the SSM to determine which non fuzzy words were associated to any fuzzy words in a sentence

To represent the impact of a fuzzy word on a non-fuzzy word the quantified fuzzy words were scaled on a 1 to -1 scale, allowing a parent node with subsequent nodes moving in either direction from it. Each of the fuzzy words has an effect on the non-fuzzy word's distance and subsumer depth

from other words. When two fuzzy words with associated words from the same category are compared to each other the difference between the fuzzy words on the scale is added to both the distance between the words and the subsumer depth to represent the pull the fuzzy words have on those two distances.

VI. CREATING AN EVALUATION DATASET

A problem with evaluating FAST was that none of the existing datasets of human ratings for sentence similarity were equipped to properly test it as they didn't contain enough sentences that contained fuzzy words. Therefore a new fuzzy sentence similarity dataset had to be created. The dataset also needed to evaluate whether or not increasing the number of fuzzy words affected the accuracy of FAST against human ratings. Therefore two sentence similarity datasets were needed one with a set of sentence pairs where each sentence in each pair contained a fuzzy word from a particular category and one with a set of sentence pairs where each sentence contained two fuzzy words from either the same or different categories.

The initial step in building the single fuzzy word per sentence dataset was to generate a list of sentence pairs. The method for generating these sentence pairs was through adding a fuzzy component to sentence pairs from an existing dataset. The dataset that was selected was the benchmark O'Shea dataset [21], an extension of the original O'Shea dataset [8][3]. There were a series of steps involved in adding fuzzy components to the sentences. Firstly30 sentence pairs were selected from the dataset. Of the 30 pairs, 20 were selected with a high level of similarity, 5 were selected with a medium level of similarity and 5 were selected with a low level of similarity. This was done to prevent the results from being clustered around a small group of values and instead return a large range. Each of the sentence pairs was split up and the sentences from them randomly divided among three experts, with backgrounds related to the English Language. Each of the experts was then instructed to add a fuzzy word to each sentence enhancing or detracting from a particular attribute within the sentence. Through this method three versions of each sentence were collected. From these, two random sentences per corresponding sentence pair were paired together, thus creating a set of sentence pairs with a fuzzy component.

The next stage was to collect human test data for the sentence pairs. Towards this end a similar methodology was applied as was used in the O'Shea dataset [8, 21] in terms of collecting human ratings. The main difference in methods was that in the case of this dataset respondents were asked to rate sentences based on how similar they were to each other on a 0 to 10 scale (as was done in the Mendel codebook [13]). The reason for this was to bring the human results in line with the scale used by STASIS and FAST. 18 people in total were surveyed through the use of questionnaires and their responses were taken to form the first part of the dataset.

The same method could not be used for generating the dataset of sentences with two fuzzy words due to the inherent complexity. Therefore it was decided to generate a new set of sentence pairs through taking sentences from the Gutenberg Corpus [22] and adapting them. The following steps were taken to generate the new sentences:

1. 25 sentences were extracted from the corpus.
2. For each of those sentences, a new sentence was generated through substituting the fuzzy words with random fuzzy words from the same category (In the case of 5 sentences the substitution was from the same node to ensure some high similarity pairs existed). This was an automated process.
3. To ensure the presence of low similarity pairs an additional 10 random sentences with two fuzzy words were paired.

Given the increased number of fuzzy words per sentence a larger number of human responses would be required than for the earlier component as the increased number of fuzzy words increased the potential level of variance between individuals. Through this a total of 26 responses were collected from test subjects.

With the dataset now having been built, the implemented FAST SSM was in a position to be tested and compared against two established short text semantic similarity measures: STASIS [3] and LSA [2]. This was done through running each of the sentence pairs in the datasets through each measure and comparing the results with the human ratings.

VII. RESULTS AND DISCUSSION

Table I and II show the results collected from tests with sentences containing one and two fuzzy words respectively. Each table shows, for every sentence pair (SP) the average human rating, the average standard deviation (SD) for the human results for each word, the result from LSA [2], the result from STASIS and the result from FAST. The testing was done in keeping with the earlier test with STASIS and LSA in [23]

Using Pearson's correlation, the levels of correlation between FAST, LSA and STASIS against the human ratings was calculated. Pearson's correlation was the method originally used to benchmark both the word similarity measure and STASIS (when it was first tested against LSA) [4][21].

The results in both tables show that fuzzy words have an impact on sentence similarity. From the original tests of LSA [2], [23] it can be observed that there is a substantially lower correlation between LSA and the human dataset than there is between STASIS, FAST and the same dataset (with LSA only showing a correlation of 0.64. The difference in correlations between FAST and STASIS is smaller, with both of the measures showing high correlations with the human test data of 0.74 and 0.71 respectively.) There is an

improvement shown by the FAST measure over the STASIS measure of 0.5.

TABLE I. RESULTS FOR SENTENCE PAIRS WITH 1 FUZZY WORD

SP	Average Human Rating	SD	LSA	STASIS	FAST
SP 1	3.833333333	2.020725942	**0.48**	**0.7502233**	**0.719345**
SP 2	0	0	**0.01**	**0.4681352**	**0.468135**
SP 3	7.3	1.994993734	**0.26**	**0.670747**	**0.670747**
SP 4	7.952380952	1.850032175	**0.84**	**0.7466893**	**0.744221**
SP 5	1.280952381	2.429736415	**0.02**	**0.5553658**	**0.555366**
SP 6	8.719047619	1.00180789	**0.95**	**0.6267825**	**0.626782**
SP 7	7.095238095	1.736512652	**0.63**	**0.8544305**	**0.848375**
SP 8	6.719047619	1.761992919	**0.81**	**0.7799703**	**0.774533**
SP 9	0.952380952	1.799616361	**0.49**	**0.6156106**	**0.675977**
SP 10	8.247619048	1.008275284	**0.46**	**0.707534**	**0.824531**
SP 11	4.957142857	1.489486968	**0.49**	**0.4133481**	**0.406414**
SP 12	0.528571429	0.978336781	**0.32**	**0.4879589**	**0.476782**
SP 13	3.285714286	2.570075041	**0.05**	**0.5651501**	**0.604964**
SP 14	6.371428571	1.826784842	**0.93**	**0.9240607**	**0.890983**
SP 15	9.138095238	0.891894719	**1**	**0.9999256**	**0.999926**
SP 16	6.780952381	1.809590851	**0.7**	**0.8441184**	**0.844118**
SP 17	3.228571429	2.385821212	**0.59**	**0.3209244**	**0.320302**
SP 18	2.10952381	1.994969865	**0.61**	**0.4967571**	**0.501914**
SP 19	6.757142857	2.212141819	**0.79**	**0.7792921**	**0.769594**
SP 20	8.985714286	0.783763813	**0.36**	**0.8237529**	**0.835592**
SP 21	3.547619048	3.240002939	**0.28**	**0.5446825**	**0.544683**
SP 22	8.852380952	1.45004105	**0.42**	**0.8823861**	**0.901932**
SP 23	7.042857143	1.622828219	**0.8**	**0.8586637**	**0.865622**
SP 24	3.833333333	2.296156208	**0.39**	**0.7073969**	**0.708897**
SP 25	8.857142857	0.963624112	**0.72**	**0.7419708**	**0.76871**
SP 26	7.583333333	1.834893276	**0.96**	**0.8666659**	**0.918686**
SP 27	8.919047619	1.075927064	**0.71**	**0.7077952**	**0.794916**
SP 28	6.914285714	2.015511279	**0.88**	**0.8618352**	**0.861835**
SP 29	1.295238095	2.211442107	**0.16**	**0.3848023**	**0.384802**
SP 30	6.623809524	2.398312899	**0.48**	**0.53408**	**0.574442**

TABLE II. RESULTS FOR SENTENCE PAIRS WITH 2 FUZZY WORDS

SP	Average Human Rating	SD	LSA	STASIS	FAST
SP 1	5.623076923	2.934117611	**0.66**	**0.867506344**	**0.90438071**
SP 2	1.715384615	2.059066256	**0.72**	**0.401883406**	**0.588147648**
SP 3	3.769230769	2.270113518	**0.82**	**0.725867977**	**0.944205512**
SP 4	0.75	1.620678887	**-0.01**	**0.236970266**	**0.210048607**
SP 5	3.707692308	2.748370146	**0.84**	**0.878031338**	**0.901081239**
SP 6	8.35	1.906462693	**0.99**	**0.996767012**	**0.996767012**
SP 7	5.676923077	2.615998118	**0.98**	**0.89789871**	**0.937227926**
SP 8	3.842307692	2.814984629	**0.9**	**0.946419311**	**0.978161396**
SP 9	4.873076923	2.593616424	**0.73**	**0.794149494**	**0.821506237**
SP 10	6.865384615	2.156096901	**0.92**	**0.898909566**	**0.969422172**
SP 11	1.223076923	2.372561096	**0.08**	**0.544778109**	**0.577446938**
SP 12	7.126923077	2.366188106	**0.72**	**0.499725184**	**0.996164928**
SP 13	5.284615385	2.619876685	**0.16**	**0.859676943**	**0.96962976**
SP 14	5.938461538	2.144402373	**0.59**	**0.842630421**	**0.966586013**
SP 15	7.380769231	1.948541861	**0.18**	**0.921057892**	**0.942793257**
SP 16	3.238461538	2.843529767	**0.71**	**0.667407844**	**0.759756521**
SP 17	4.311538462	2.880184289	**0.86**	**0.811781296**	**0.964534331**
SP 18	1.446153846	2.390687059	**0.06**	**0.337421118**	**0.362451298**
SP 19	7.792307692	2.608896024	**1**	**0.974805896**	**0.974805896**
SP 20	7.815384615	1.973969059	**0.93**	**0.734179502**	**0.791638477**
SP 21	2.111538462	3.369786572	**0.06**	**0.625114234**	**0.625114234**
SP 22	6.25	2.718860055	**0.78**	**0.946154934**	**0.992869046**
SP 23	8.161538462	1.910618104	**0.97**	**0.9988911**	**0.996478098**
SP 24	7.215384615	2.429599524	**0.93**	**0.843101818**	**0.844179256**
SP 25	7.484615385	1.915973342	**0.92**	**0.853863238**	**0.853863238**
SP 26	6.330769231	2.481897537	**0.68**	**0.746150767**	**0.858831406**
SP 27	3.842307692	2.564398265	**0.92**	**0.95620636**	**0.967062298**
SP 28	1.269230769	1.870351674	**0.07**	**0.439591166**	**0.438047509**
SP 29	6.069230769	2.655826686	**0.47**	**0.714128898**	**0.912815456**
SP 30	6.488461538	2.614930504	**0.79**	**0.747826177**	**0.965312492**

The results from Table II show the similarity measures being measured against a second set of human ratings, using sentences that contained two fuzzy words. From this there is a clear difference in the levels of correlations with sentences with one fuzzy word. The results showed that LSA had a correlation of 0.63, which is almost identical to its result from the previous set. This shows that as the number of fuzzy words increases, LSA does not become less reliable. Both STASIS and FAST continued to show strong correlations with the sentences however the differences in correlations increased, with FAST showing an even greater level of correlation, with FAST having a correlation of 0.77 and STASIS having a correlation of 0.71.

The results present a picture of the effect of fuzzy words on the overall nature of text similarity and the importance of a new SSM specifically factoring these words in.

VIII. CONCLUSION

This paper has presented a novel sentence similarity algorithm known as FAST which has shown an improvement over existing algorithms STASIS and LSA which do not take into consideration of fuzzy words when computing semantic sentence similarity. 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 have an effect on the performance of SSM. This is demonstrated through the improvement that FAST had over STASIS and LSA.

The second contribution of this paper is that through the quantification of fuzzy words a collection of six categories is now created that can be utilized in future work that deals with the relationships between words in these categories. Therefore it is beneficial in terms of future work in the area of fuzzy similarity. Through use of the approach outlined in section III further words can be added to the categories (and the ontology structures expanded) increasing the number of accurately represented relations.

The third contribution is the creation of the evaluation dataset. This dataset could be used as a benchmark for the any future similarity measures and a new methodology is presented that can be used to create new fuzzy datasets.

Future work can involve implementing this measure in systems based around human-computer dialogue such as interactive conversational agents. Furthermore, the FAST algorithm can be easily expanded to include more fuzzy categories. In addition when fuzzy words were classified within the ontological structures, the class boundaries were crisp. Determining a method of fuzzifying these boundaries may lead to a system that is more representative of natural language.

REFERENCES

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

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

[3] Li, Y. Mclean, D. Bandar, Z. O'Shea, J. Crockett, K. "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 8, pp.1138-1150, 2006.

[4] Li, Y, Bandar, Z. McLean, D. "An approach for measuring semantic similarity between words using multiple information sources". IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp.871-882, 2003.

[5] Rada, R. Mili, H. Bicknell, E. Blettner, M. "Development and application of a metric on semantic nets", IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, pp.17-30, 1989.

[6] Resnik, Philip. "Using information content to evaluate semantic similarity in taxonomy". arXiv preprint cmp-lg/9511007, 1995.

[7] Marcus, M. Marcinkiewicz, M. Santorini, B. "Building a large annotated corpus of English: The Penn Treebank" Computational linguistics vol. 19, no.2, pp.313-330, 1993.

[8] 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: 28/05/13.

[9] Rubenstein, H, Goodenough, J. "Contextual correlates of synonymy", Communications of the ACM vol. 8, no. 10, pp.627-633, 1965.

[10] 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 no. 3, pp.499-513, 2012.

[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, no.3, pp. 307–324, 2002.

[12] Zadeh, L. "Outline of a new approach to the analysis of complex systems and decision processes". IEEE Transaction on Systems, Man and Cybernetics, vol. 1, pp.28-44, 1973.

[13] Mendel, J. "Computing with words and its relationships with fuzzistics", Information Sciences vol. 177, no. 4, pp.988-1006, 2007.

[14] Mendel, J, "Advances in type-2 fuzzy sets and systems, Information Sciences" vol. 177, no.1, pp.84-110, 2007.

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

[16] Karnik, N. Mendel, J. "Centroid of a type-2 fuzzy set", Information Sciences vol. 132 no.1, pp.195-220, 2001.

[17] Mendel, J., Hongwei, W. "New results about the centroid of an interval type-2 fuzzy set, including the centroid of a fuzzy granule", Information Sciences, vol. 177, no.2, pp.360-377, 2007.

[18] Liu, F., Mendal, J. "Encoding words into interval type-2 fuzzy sets using an interval approach", IEEE Transactions on Fuzzy Systems, vol. 16, no. 6, pp.1503, 2008

[19] Miller, G, Walter, G., "Contextual correlates of semantic similarity, Language and cognitive processes", vol. 6, no. 1, pp.1-28, 1991.

[20] Miller, G, "Word Net: a lexical database for English, Communications", ACM vol. 38, no.11, pp.39-41, 1995

[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, no. 2, pp.103-120, 2010.

[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, pp.172-181, 2008.

[23] Hart, M. The history and philosophy of Project Gutenberg. Available: http://www.gutenberg.org. Date Accessed: 28/5/2013.

Appendix 6:

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

IEEE International Conference on Fuzzy Systems (2013)

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.

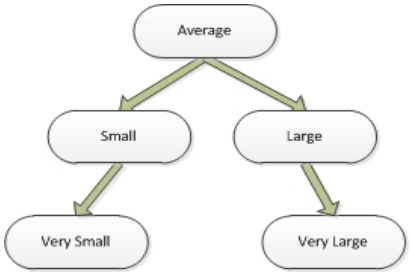

Fig. 1. Size Category Ontology

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.