Title

Benchmarking Short Text Semantic Similarity

James O'Shea, Zuhair Bandar, Keeley Crockett and David McLean

Department of Computing and Mathematics, Faculty of Science and Engineering, Manchester Metropolitan University, John Dalton Building, Manchester M1 5GD, United Kingdom

Fax: +44 161 247 1483 E-mail: {j.d.oshea, z.bandar, k.crockett, d.mclean}@mmu.ac.uk

Abstract: Short Text Semantic Similarity measurement is a new and rapidly growing field of research. "Short texts" are typically sentence length but are not required to be grammatically correct. There is great potential for applying these measures in fields such as Information Retrieval, Dialogue Management and Question Answering. A dataset of 65 sentence pairs, with similarity ratings, produced in 2006 has become adopted as a *de facto* Gold Standard benchmark. This paper discusses the adoption of the 2006 dataset, lays down a number of criteria that can be used to determine whether a dataset should be awarded a "Gold Standard" accolade and illustrates its use as a benchmark. Procedures for the generation of further Gold Standard datasets in this field are recommended.

Keywords: Short text, Sentence similarity, Semantic similarity, Benchmark, Dataset

Biographical notes: James O'Shea is a senior lecturer at Manchester Metropolitan University (MMU). He received his BSc in Chemistry from Imperial College. He worked in computer R&D at International Computers and as an independent consultant under the UK Microelectronics Applications Project until 1985. After joining MMU, he developed a research interest in AI and co-founded the Intelligent Systems Group (ISG). In addition to his work in the field of Conversational Agents (CAs), he is one of the inventors of the Silent Talker lie detector, which has attracted worldwide interest.

Zuhair Bandar is a Reader in Intelligent Systems at MMU. He received his PhD in AI and Neural Networks from Brunel University, his MSc in Electronics from the University of Kent and his BSc in Electrical Engineering from Mosul University. He is a co-founder of the ISG and his research interests include the application of AI to psychological profiling. He is the Technical Director of Convagent Ltd, an MMU spinout company which provides business rule automation with natural language interfaces using CAs.

Keeley Crockett is a Senior Lecturer at MMU. She received her PhD in Machine Learning from MMU and her BSc in Computation from the University of Manchester Institute of Science and Technology. She is a committee member of the IEEE Women into Computational Intelligence Society and a full member of the IEEE Computational Intelligence Society. Her main research interests include fuzzy decision trees, applications of fuzzy theory, and data mining. She is a knowledge engineer and founding member of Convagent Ltd.

David McLean is a Senior Lecturer at MMU. He received his PhD in Neural Networks from MMU and his BSc in Computer Science from the University of Leeds. His prime research interests lie in AI, especially neural networks and CAs. He worked as an AI researcher and developer at DERA and

Title

Thomson Marconi Sonar. He is also one of the inventors of the AI-based Silent Talker lie detector system.

1 Introduction

The original motivation for this work was the production of a dataset to evaluate STASIS (Li et al., 2006), the first measure developed specifically for Short Text Semantic Similarity (STSS). Earlier studies of semantic similarity concentrated on either single words (Resnik and Diab, 2000) or complete documents (Landauer et al., 1998).

STASIS is intended for use in Conversational Agents (CAs) (O'Shea et al., 2009, Crockett et al., 2009). Other applications that could benefit from STSS measures include the automatic processing of text and e-mail messages (Lamontagne and Lapalme, 2004), Information Retrieval (Marton and Katz, 2006), health care dialogue systems (Bickmore and Giorgino, 2006) and natural language querying of databases (Erozel et al., 2008).

There has been a surge of interest in STSS and the 2006 dataset (Li et al., 2006) has been adopted as the *de facto* Gold Standard benchmark for evaluating and comparing new developments. Consequently, publication of a thorough analysis of this dataset, acting as a validation and establishing procedures for the production of further Gold Standard STSS benchmarks, has become necessary.

The general concept of similarity has been an important research topic in psychology for decades. In 1977 the contrast model was published (Tversky, 1977), which successfully explained a number of earlier approaches to similarity in different fields.

Semantic similarity is fundamental to many experiments in fields such as NLP, linguistics and psychology (Resnik, 1999, Prior and Bentin, 2003, McNamara and Sternberg, 1991). Word similarity studies date back to the 1960s (Rubenstein and Goodenough, 1965) and in Information Retrieval (IR), document similarity has been systematically studied since the 1970s (Salton et al., 1975). IR includes topics such as question answering (Quarteroni and Manandhar, 2008), textual entailment (Jijkoun and de Rijke, 2005) and sentence fusion (Barzilay and McKeown, 2005).

Semantic similarity is held to be a widely understood concept. Miller and Charles (Miller and Charles, 1991), in a word-based study wrote " . . . subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty." This view, has been reinforced by other researchers such as Resnik (1999) who observed that similarity is treated as a property characterised by human perception and intuition. There is an implicit assumption, that not only are the participants comfortable in their understanding of the concept, but also when they perform a judgment task they do it using the same procedure or at least have a common understanding of the attribute they are measuring.

In future, complex and demanding applications where high accuracy of understanding of the user intent is needed, the stakes are high and the users may present adversarial or disruptive characteristics in interacting with systems will require the use of STSS measures. Therefore it is crucial that STSS measures are validated in a separate stage, to avoid confounding factors in final built system evaluations.

Proper evaluation requires the use of appropriate statistical methods, the creation of standard benchmark datasets and a sound understanding of the properties of such datasets. Because semantic similarity is characterized by human perception there is no "ground truth" similarity rating that can be assigned to pairs of Short texts (STs), the only way to obtain them is through carefully constructed experiments with human participants.

Consistency of an STSS measure with human judgment is used to determine its quality. This is calculated as the product-moment correlation coefficient between the set of ratings from the algorithm and those from a

sample of human participants (Resnik, 1999), using a benchmark dataset. It is also possible to calculate bounds for expected performance of STSS measures as the average, best and worst human performances on the benchmark dataset, using leave-one-out resampling (Li et al., 2004).

The remaining sections are organized as follows: Section 2 discusses some relevant features of semantic similarity and the comparison of machine measures. Section 3 reviews the variations in prior work on word similarity and discusses the requirements for meeting the Gold Standard. Section 4 describes the design of 4 experiments to investigate the influence of two important experimental factors, section 5 analyses the results and section 6 concludes with directions for future work.

2 Models and Features of Semantic Similarity

2.1 Models of Semantic Similarity

Salton's Vector Space Model (VSM) described a method for clustering documents in a semantic space based on vectors of extracted terms (Salton et al., 1975). It also introduced a systematic method of identifying good terms, weighting terms based on their importance in a particular ontology and combining poor terms to synthesise better terms. Although the VSM is basis of modern measures where similarity is calculated as the cosine of the angle between vectors (Yeh et al., 2008), it suffers from a serious weakness. The elements of the vector are symbols and there is no knowledge of their meanings. At best, operations such as stemming or lemmatisation (Manning et al., 2008) are used to reduce inflections of a particular word to a single form. Such measures work poorly when there is little word overlap (Jeon et al., 2005).

Improvements in measuring ST similarity have come through the modeling of semantic content. LSA (Deerwester et al., 1989), developed as an IR technique, can compare two search terms in a reduced dimensionality space (Deerwester et al., 1990) and hence can compare STs. Developments in measuring word semantic similarity have fuelled other approaches (Mihalcea et al., 2006). SimFinder (Hatzivassiloglou et al., 2001) is an IR technique that uses limited NLP to extract noun phrases from short paragraphs and generate feature vectors using Wordnet and Levin verb classes. STASIS (Li et al., 2006) combines semantic information from Wordnet with corpus statistics in a short vector representation which also takes account of word order and function words. STASIS was designed for STs of 10-20 words used in applications such as CAs and e-mail processing which may not follow the grammatical rules for sentences. However, all of the STs in the 2006 dataset are valid sentences allowing comparative evaluations with measures which depend upon NLP techniques.

2.2 Relevant Features of Similarity

Empirical studies suggest that semantic similarity is more subtle than has been assumed. Some draw a distinction between "similarity" and "relatedness" (Resnik, 1999, Vigliocco et al., 2002). Resnik gives an example: cars and gasoline seem more closely related than cars and bicycles, but the latter pair is more similar. Although Resnik specifies semantic similarity as a special case of semantic relatedness, Charles (2000) has used relatedness to describe degrees of similarity in an empirical study.

Four forms of similarity are described by Klein and Murphy (2002): Taxonomic, Thematic, Goal-derived and Radial. Taxonomic similarity is the foundation of Noun similarity studies, following ISA relations through a structure such as Wordnet. Cars and gasoline are a good example of Thematic similarity (related by co-occurrence or function). Goal–derived items are connected by their significance in achieving some goal and Radial items are connected through a chain of similar items, possibly through some evolutionary process. The context in which the similarity judgment is made could result in any of the forms dominating the decision.

Semantic Distance (dissimilarity) is also measured. Distance can be converted to similarity by applying an inversion operation (Tversky, 1977) or by looking for a negative correlation with distance instead of a positive correlation with similarity (Miller and Charles, 1991).

The concept of similarity may in itself be asymmetrical, depending on the circumstances in which items are presented. According to Tversky, “A man is like a tree” and “A tree is like a man” are interpreted as having different meanings (Tversky, 1977). Gleitman et al. (1996) claim that the structural position of the noun phrases set them as figure and ground or variant and referent, leading to the asymmetry.

Most studies use similarity measures on a scale running from 0 to a specified maximum value, typically 4. However this rating scale has no capacity to represent oppositeness (antonymy) as more different than unrelatedness. Antonyms also generate high similarity values with co-occurrence measures (Miller and Charles, 1991).

Finally, it has been claimed that given two pairs of identical items to rate, participants give a higher rating to the pair with more complex features (Resnik, 1999).

If these factors influence individuals differently in making their judgments then there is scope to question the accepted concept of a common human model of similarity.

3 Requirements of a Gold Standard dataset

This section examines potentially confounding variations in experiments which could undermine the effectiveness of a Gold Standard dataset. The term "Gold Standard" originally applied to establishing the value of a currency and is associated with stability, transparency and reliability. It is also used to describe a testing method as being either the best currently available or the best possible.

Datasets described as Gold Standard have been produced for question reformulation (Shaw et al., 2008), ontology mapping (Hu et al., 2008) and spoken dialogue summarisation (Gurevych and Strube, 2004). Strategies to assure a gold standard include training the raters (Hu et al., 2008), providing a coding manual (Wiebe et al., 1999) and allowing users to rate their level of confidence in their judgements (Su and Gulla, 2006). Emphasis is placed on agreement between multiple raters (Kilgarriff, 1998) and some studies may allow the human raters to negotiate a consensus rating (Su and Gulla, 2006). Alternatively, a bias correction mechanism has been applied to raters who disagree (Wiebe et al., 1999). Another strategy uses an iterative process in which an initial set of human ratings are scored against an existing gold standard with feedback before evaluations proceed (Kilgarriff, 1998). To reduce the labour intensive nature of creating materials for datasets, it is possible to use a combination of classifiers and find data items over which the classifiers disagree, then manually classify these before adding them to the gold standard (Ngai and Yarowsky, 2000) dataset.

Unfortunately, strategies such as training, providing manuals and review by human experts are only applicable in domains where there is a known ground truth, which is not the case with semantic similarity. Also, high inter-rater agreement is difficult to achieve with more taxing combinations of data items.

3.1. Representative sampling

Any non-trivial benchmark data set will be a sample of a larger population of data. The two sampling issues involved here are representing the general human population and obtaining STs that represent the overall semantic space of the English Language. The sample of humans is particularly pertinent for practical applications accessible by anyone via the web. Neither of these issues was given serious consideration in the seminal word studies.

The 2006 dataset addresses the first issue by using a more representative sample of the human population than previous word similarity studies. However, because the dataset was highly novel, it was decided to use the dictionary definitions of the word pairs used by Rubenstein & Goodenough (1965) as the data sample, to benefit from the prior knowledge accumulated about the words.

3.2 Precision and Accuracy

The dataset contains judgments by human participants. Precision requires the judgments to be in close agreement with each other. Accuracy requires the derived measures to be in close agreement with the "true" similarity. Precision is affected by both the participant's internal state (mental and physical) and the measurement instrument (for example ambiguity of instructions). Accuracy depends on a common model of similarity and also on the possibility of blunders by the participant. These problems influence the design of the measurement instrument.

3.3 Measurement scale

The scale on which the similarity measurements are made determines the statistical techniques that can be applied to them (Blalock, 1979); the question is how sophisticated a measurement scale can be used?

It is a reasonable assumption that human intuitions of similarity are at least ordinal – that one pair of items can be more similar to each other than another pair. Interval scales improve on ordinal by having consistent units of measurement and ratio scales improve over interval by having an absolute zero point on the scale. Word semantic similarity has always been treated as a ratio scale attribute for both machine measures and human datasets. The ST dataset is intended for algorithms that run from an absolute zero point (unrelated in meaning) to a maximum (identical in meaning). Setting the upper bound of the scale is common in word similarity measures and transformation of the range for comparisons is permissible.

3.4 The Rubenstein & Goodenough Legacy

The most influential similarity dataset was produced in the 1960s (Rubenstein and Goodenough, 1965). This word similarity experiment has been replicated on several occasions (Miller and Charles, 1991, Charles, 2000, Resnik, 1999). These studies, coupled with that by Li et al. (2006) provide evidence to support the view that human similarity measures are at least ordinal, showing reasonably consistent ranking between individuals, groups and over time. Collectively however, these replicated experiments have a number of uncontrolled factors which prevent them from being truly comparable. An analysis of these factors and their potential confounding effects forms the basis of the experimental programme conducted in this study.

3.4.1 Method of presentation of materials

Rubenstein & Goodenough printed each of the word pairs on a separate slip of paper and the subjects were asked to sort them in order of similarity before rating them. Miller & Charles (1991) presented all of their word pairs on two sheets of paper. Charles (2000) used a questionnaire with each word pair on a separate page. Resnik used an electronic version of the Miller & Charles (1991) questionnaire.

Potential variation in the results could be introduced because the Rubenstein & Goodenough and Miller & Charles versions provide substantial exposure to the whole set of items during the rating process. This provides a grounding context not available in the Charles (2000) version. Also, the physical process of sorting before rating in the Rubenstein & Goodenough version may reduce the level of abstraction of the rating process.

3.4.2 Method of randomization of materials

Rubenstein & Goodenough shuffled the slips containing the word pairs into random order before presentation. Miller & Charles (1991) randomised the order of word pairs on the 2 sheets for each participant. Charles (2000) randomized the order of pages within his questionnaire for each participant. Resnik used two variants, a randomly selected ordering and a reversed version of that ordering.

Previously seen word pairs could influence the participant's judgment of the current pair and randomization seeks to prevent the same sequences being presented to all the participants which might bias the data. One potential variation is the amount of noise introduced to the ratings as small individual biases are spread evenly throughout the dataset. None of the procedures describe randomization of the order of words within a pair (e.g. coast-forest vs. forest-coast) despite prior work (Tversky, 1977).

3.4.3 Instructions and guidance provided to participants

Rubenstein & Goodenough instructed the participants to assign a value from 4.0 – 0.0 to each pair – the greater the similarity of meaning the higher the number. Miller & Charles (1991) instructed the participants to examine each pair closely and then to rate it on a 5-point scale from 0 to 4, where 0 represents no similarity of meaning and 4 perfect synonymy. Charles (2000) advised participants to study each pair and to rate it for semantic similarity on a 5-point scale including antonymy. Charles illustrated decreasing semantic similarity using the following list of pairs: snake-serpent, snake–eel, snake-alligator, snake-frog, snake-book and snake-bulb. Resnik used the same instructions as Miller & Charles (1991).

Potential variation arises from encouraging the use of the first decimal place (Rubenstein & Goodenough) as opposed to instructions which may encourage the use of integers only (Miller & Charles).

3.4.4 Measurement Scale Definition

Rubenstein & Goodenough's instructions focus attention on the relative similarities of items in the dataset and may encourage expanding the range of similarity judgments to fill the range 4.0 to 0.0 even if other pairs with higher or lower similarity could exist.

Miller & Charles (1991) gave absolute descriptions of the endpoints of the scale (as did Resnik) and Charles (2000) used semantic anchors for 5 points on a scale which ranged from similarity to dissimilarity. The anchors ranged from 4: "identical in meaning" to 0: "opposite in meaning." Potential variations include lack of robustness in the Rubenstein & Goodenough ratings when considered alongside new item pairs to judge. Also the use of semantic anchors in Charles (2000) could provide better interval measurement and lower noise than other methods.

3.4.5 Sampling the population for Participants

Rubenstein & Goodenough used two groups of college undergraduates for a total of 51 participants. Miller & Charles (1991) used 38 students; all specified as Native English speakers. Charles (2000) used two groups of undergraduates (50 participants on the 65 pair data set and 58 on the 30 pair subset). All were Native English Speakers and all received credit for courses taken in psychology. Resnik used 10 participants who were all computer scientists at graduate student or postdoctoral level. The lack of specificity in the published protocols prevents us from knowing:

- if Rubenstein & Goodenough and Resnik used only Native English speakers,
- the academic background of students used by Miller & Charles and Resnik,
- the academic level of the students used by Miller & Charles (1991),
- the gender or age composition of the groups.

These could be confounding factors in comparing experiments along with group size, although experience suggests that only Resnik's sample of 10 is likely to be a problem.

The most important potential issue is the high homogeneity of participants within the groups of particular experiments and their distinct differences from the general population which could reduce their value as representative samples.

4. Design of Experiments

The primary aim of this study was to investigate the influence of two factors, Order and Anchor on the outcomes of studies of STSS. The two different states of Order investigate the difference between the sorting approach of Rubenstein & Goodenough and the questionnaire form used in the replications. The two different states of Anchor investigate the difference between Charles (2000) in using semantic anchors to describe the major similarity scale intervals and the other experiments which did not.

Given the possibility of an interaction between the two factors, a 2-level, 2-factor ANOVA analysis was conducted. Four experiments were required to collect data for each permutation of the factors. Experiment 1 simply involved extracting the data for the first 18 participants from the experiment in Li et al. (2006); experiments 2.1,2.2 and 2.3 involved the collection of new data.

4.1 Control Variables

Randomization, instructions and population sample were treated as control variables.

4.1.1 Randomization.

In each of the 4 experiments each participant received the 65 sentence pairs in an individually randomized order. Also, sets of materials were produced in pairs (A and B), one of which had each pair of sentences in a particular order and the other of which had the reverse order (so A would have the sentence defining *coast* before the sentence defining *forest* and B would have *forest* before *coast*). Individual sentence pairs were then transferred between A and B at random to produce two new sets of materials containing mixtures of the A and B orders.

4.1.2 Common Instructions and variations

Some variation in instructions was required as two experiments involved sorting cards and two involved questionnaires. In all experiments the task was described as "rate the similarity of meaning." Participants were asked to do this by "writing a number between 0.0 (minimum similarity) and 4.0 (maximum similarity) on the form, please do not use values greater than 4.0. You can use the first decimal place (e.g. 2.2) to show finer degrees of similarity."

For the questionnaire forms participants were instructed to work through from start to end without going back to revise earlier judgments. For the card sort participants were asked to

"…start by reading through the cards in the order you got them in, thinking about the similarity of the meanings of the two sentences on each card. Now please sort the cards in a rough order of the similarity of meaning of the sentence pairs" before recording their judgments.

4.1.3 Population and Sampling

The aspiration was to represent the general population. However, because participants would be performing the task without supervision, it was decided to restrict the sample to people with graduate-level education. They were also restricted to Native English speakers, in common with many other verbal experiments. Overall gender balance was achieved across the study with 51% males, 41% females and 8%

withholding information. The balance was generally achieved within the experimental group samples although one was moderately biased towards males. All participants volunteered without compensation.

4.1.4 Age

The overall average age was 41.9 (SD = 10.1) years (5 participants withheld their age). The range of average ages for the experimental groups was 39.7 (SD = 8.7) – 43.6 (SD = 12.8) years. Thus the age distribution was more diverse than in prior word studies.

4.1.5 Education

All but one participant (who had an equivalent professional qualification) had a bachelor's degree. The overall breakdown by academic background was 50% Science & Engineering, 39% Arts & Humanities, 8% mixed Arts/Sciences (e.g. Architecture) and 3% withheld information. The overall breakdown of qualifications was 56% bachelor's degrees, 21% PhDs, 15% Masters, 4% PGCE with the remainder being either professional (graduate equivalent) qualifications or withheld.

In 3 groups the balance was approximately two-thirds bachelor level and one-third postgraduate degree. In the remaining group the balance was approximately reversed. In 3 groups there was an almost perfect balance between arts and sciences backgrounds, the remaining group had was moderately biased towards sciences.

4.2 Experimental factors.

Four experiments were undertaken to examine the influence of varying two factors, Order and Anchor, on human ratings.

4.2.1 Order

The order factor had two levels, *quest* and *card*. The *card* variant used a deck of 65 stiff cards, with one sentence pair printed on each. The *quest* variant presented the sentence pairs on a questionnaire, with one pair on each sheet. Judgments were recorded on the sheet before moving on. In all experiments the question pairs were identified by code strings for ease of transcription, values were allocated so as not imply any pre-conceived similarity values.

4.2.2 Anchor

The anchor factor had two levels, w*ith* and *without* semantic anchors. The *with* variant described the scale points with 5 semantic anchors taken from Charles (but not used as a specific set in his study). The *without* variant simply presented the sentence pairs without any accompanying semantic anchors. The semantic anchors are shown in table 1.

Table 1 Semantic anchors adopted from Charles (2000)

Scale Point	Semantic Anchor
0.0	The sentences are unrelated in meaning.
1.0	The sentences are vaguely similar in meaning.

2.0	The sentences are very much alike in meaning.
3.0	The sentences are strongly related in meaning.
4.0	The sentences are identical in meaning.

The structure of the experiments making up the study is shown in table 2.

Table 2 Experiments comprising the study

Experiments	With Anchors	Without Anchors
Questionnaire	Expt 1 Questionnaires with Semantic anchors	Expt 2.1 Questionnaires Without Semantic Anchors
Card Sort	Expt 2.3 Card Sort with Semantic Anchors	Expt 2.2 Card Sort Without Semantic Anchors

4.3 ANOVA Analysis of Order and Anchor

Statistical tests (Bartlett's, Levine's and probability plot) were conducted to confirm the validity of using ANOVA with the data. The 2-level, 2-factor design was implemented as a 1-factor, 4-level General Linear Model in Minitab, combining all 65 sentence pairs in a blocked design to analyse the results. A total of 72 participants was distributed into 4 groups of 18, one for each combination of factors. The results were Order F = 7.49, P= 0.007 and Anchor F= 63.67, P= 0.000 indicating both Order and Anchor factors were significant. However, Order.Anchor was reported as F= 3.22, P= 0.074 indicating that the interaction between Order and Anchor was not significant.

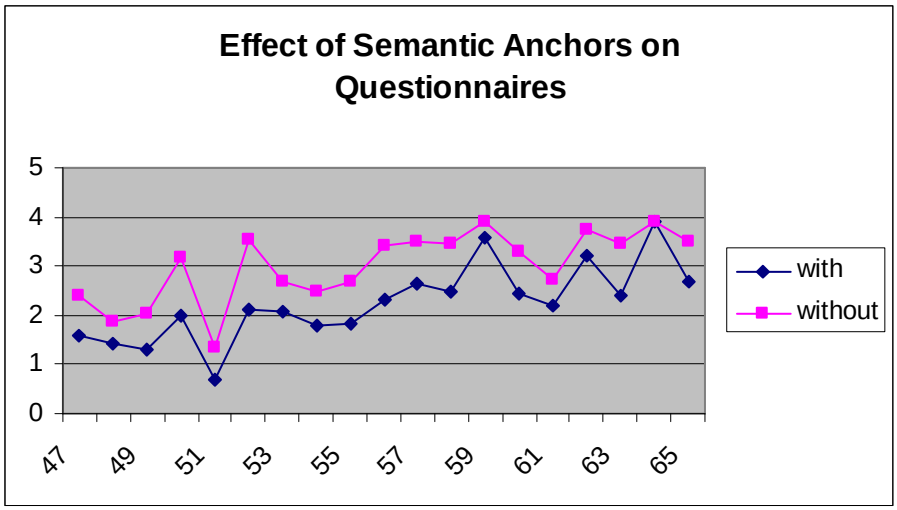
5. Analysis of results

5.1 Effect of varying Order and Anchor.

Having established that Order and Anchor are significant factors in the variation of semantic similarity ratings, the question remains "How do they affect the ratings?" This is best illustrated by line graphs of the medium to high similarity items concentrated in sentence pairs 47-65. Each graph keeps one of the factors constant and plots the trend for both versions of the other factor.

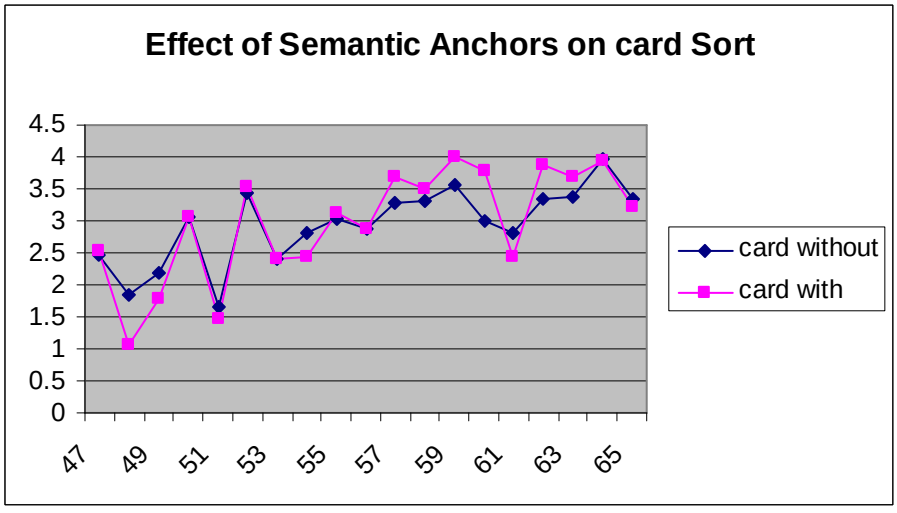
Figure 1 shows the effect of the presence or absence of semantic anchors on the questionnaire form and figure 2 shows their effect on the card sort form.

Figure 1 The effect of Semantic Anchors on ratings collected using Questionnaires



For the questionnaire form, the two plots are almost identical, except that *without semantic anchors* is displaced upwards to give higher semantic similarity ratings. There is a minor disagreement at sentence pair 49.

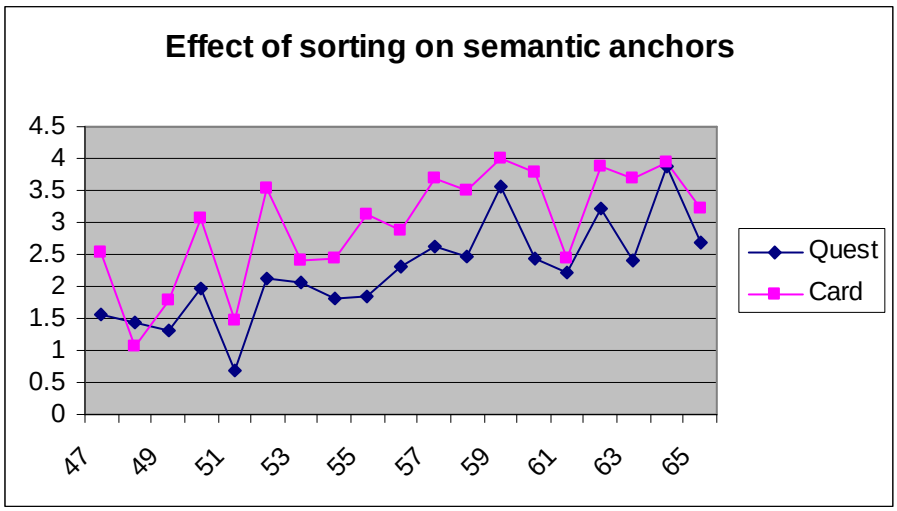
Figure 2 The effect of Semantic Anchors on ratings collected using Card Sorting



For card sorting, a number of items have identical ratings. There is more disagreement than the questionnaire form, but there are still regions where the general trends are in agreement. The main disagreements are at sentence pairs 54 and 61.

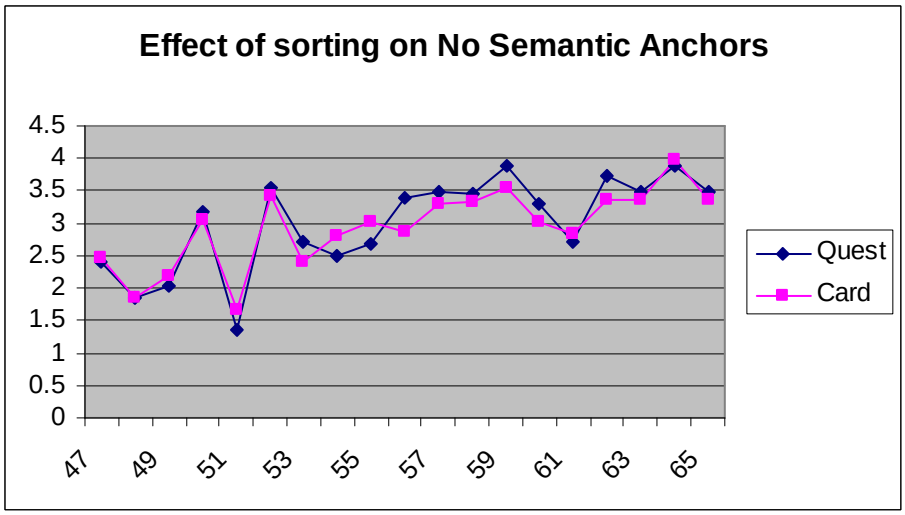
Figure 3 shows the effect of using questionnaire or card sorting on the *with semantic anchors* form and figure 4 using questionnaire or card sorting on the *without semantic anchors form*.

Figure 3 The effect of Questionnaires vs. Card Sorting on ratings collected using Semantic Anchors



For the *with semantic anchors* versions, the overall trends are similar, with the card sorting version displaced upwards to give higher semantic similarity ratings. The noticeable disagreements are at sentence pairs 48 and 56.

Figure 4 The effect of Questionnaires vs. Card Sorting on ratings collected without Semantic Anchors



For the *without semantic anchors* versions there are a number of points which are effectively identical and areas with similar trends. The main disagreements are at sentence pairs 54 and 56.

5.2 Correlation between experiments

In the following tables, Q means Questionnaires, C means Card Sort, W means With Semantic Anchors and O means withOut Semantic Anchors, e.g. QW means a combination of Questionnaires with Semantic Anchors. As might be expected from the graphs, there is a strong correlation between the ratings obtained in each of the experiments. These are shown in table 3:

Table 3 Correlations between the ratings obtained from the different experiments

Experiment	1 QW	2.1 QO	2.2 CO	2.3 CW
1 QW	X			
2.1 QO	0.959	X		
2.2 CO	0.944	0.981	X	
2.3 CW	0.927	0.981	0.958	X

Using 4 groups of 18 participants, all p-values were less than 0.01. In the original experiment with n=32, measures of inter-rater agreement (r) were calculated using the leave-one-out method of cross-validation. The average human achieved a correlation of 0.825 and the best human achieved a correlation of 0.921 (with the average ratings for the rest of the group). So the different combinations agree with each other better than the best agreement within the original human group.

5.3 Consistency of judgment within experiments

Consistency (related to accuracy) can be investigated by calculating inter-rater agreement using product-moment correlation coefficients (r) as in (Resnik, 1999). The higher the value of r, the better is the performance of the measure. The results are shown in table 4.

Table 4 Inter-rater agreement within experiments

Experiment	Mean r	Best participant r	Worst participant r
1 QW	0.855	0.926	0.633
2.1 QO	0.899	0.96	0.75
2.2 CO	0.885	0.919	0.799
2.3 CW	0.938	0.976	0.83

The combination of *Card Sorting with Semantic Anchors* scores best on all three criteria, the correlations of the best and worst humans with the rest of the participants and the average of the correlations of all the human participants. The individual correlation coefficients are significant at the 0.01 level.

5.4 The effect of the factors on noise

Noise (related to precision) can be obtained by calculating the standard deviations of the human ratings for each sentence pair within each experiment, then taking their mean, as shown in table 5.

Table 5 Noise within experiments

Experiment	Mean of SDs
1 QW	0.589
2.1 QO	0.723
2.2 CO	0.615
2.3 CW	0.36

The lowest noise value was obtained from the combination of *Card Sorting with Semantic Anchors*. Conducting 2-sample t-tests indicated that there was a strongly significant difference between the noise levels for 2.1 vs. 2.3 and that the difference for 2.2 vs. 2.3 was just on the boundary of significance (p=0.05), the other combinations did not achieve significance.

5.5 Effect of order of presentation of sentences within a pair

The data also provides an opportunity to investigate the influence of asymmetry on similarity judgements (Tversky, 1977, Gleitman et al., 1996).

This was investigated using the ratings from the full set of 32 participants originally collected for experiment 1. A sample of 10 sentence pairs (2, 5, 44, 49, 52, 53, 57, 59, 64, 65) spanning the similarity range was selected and the A and B versions were separated out. 2-sample t-tests were conducted to determine whether the means for the two orders of presentation differed significantly. Due to some non-returns of questionnaires there was a small variation in the numbers of A and B versions of questionnaires therefore the Mann-Whitney test was also conducted for a robust second opinion. The results are shown in table 6.

Table 6 Effect of order of presentation of sentences within a pair

Sentence Pair	Human Ratings (0.0-4.0)	2-sample t-test p	Mann-Whitney test p
2	0.02	*	*
5	0.02	*	*
44	0.97	0.191	0.1925
49	1.17	0.144	0.2748
52	1.94	0.626	0.6318
53	1.93	0.694	0.7312
57	2.51	0.898	0.9388
59	3.45	0.126	0.2017

Title

64	3.82	0.929	0.1084
65	2.61	0.366	0.5154

The tests could not be conducted for sentence pairs 2 and 5 because in each case one of the two columns contained only zeros. As virtually all of the data was zeros for these pairs we can assume no effect of ordering. In all of the other cases, the p-values are substantially greater than the commonly accepted upper limit of 0.05. For both the tests this is considered as providing no evidence for a difference in similarity rating based on order of presentation of sentences within a pair.

6 Applications, conclusions and future work

6.1 Conclusions of experimental evaluation

The high inter-rater agreement between the sets of semantic similarity ratings provides evidence to support the view that a robust underlying model of semantic similarity is being accessed by the participants regardless of the variation in experimental factors. This robustness provides strong evidence to support the validity of the 2006 dataset in representing human similarity judgments and supports the use of the data set as a Gold Standard benchmark for STSS measures.

The reduction in the noise level achieved by using card sorting compared with questionnaires is significant or on the significance borderline suggesting that this technique improves the precision of human ratings.

The combination of *Card Sorting with Semantic Anchors* has the highest inter-rater agreement indicating highest consistency of judgement (significant at p=0.01) and the lowest noise suggesting the highest accuracy and precision.

The examination of asymmetry provided evidence that the order of sentences within a pair does not influence the mean of the ratings obtained (all calculated results significant p>0.05).

6.1 Application to recently developed measures

As an illustration of the value of the data set, this section draws together self-reported results from 6 studies which have used the 2006 dataset. Most of the approaches follow Li et al in using a corpus distance measure in combination with some other approach. Kennedy and Szpakowitz (2008) use Roget's Thesaurus with a cosine measure, Islam and Inkpen (2008) combine a variant of Pointwise Mutual Information (using the British national Corpus) with LCS string matching and Feng et al. (2008) use Wordnet with the Needleman-Wunsch algorithm to measure indirect relevance. O'Shea et al. (2008) performed an independent test of LSA using the web portal (Laham, 1998). A comparison of the performance of these algorithms is shown in table 7.

Table 7 Performance of STSS measures

Authors	Year	Measure	r
Li et al.	2006	STASIS Wordnet plus word position information in short vectors	0.816

Kennedy & Szpakowitz	2008	Roget's thesaurus (weighted) plus cosine measure	0.873
Kennedy & Szpakowicz	2008	Wordnet (weighted) plus cosine measure	0.851
Feng et al.	2008	Wordnet and Brown Corpus-based measure incorporating direct and indirect relevance information	0.756
Islam & Inkpen	2008	String matching (LCS) plus SOC-PMI	0.853
O'Shea et al.	2008	LSA	0.838

These results identify a clear improvement in the performance of STSS measures published in 2008.

6.3 Future work

More benchmark datasets will be produced. These are required to extend the range of speech acts inherited from the Rubenstein & Goodenough dataset, and include a richer sample of the English lexicon. Future data sets will use a combination of *Card Sorting with Semantic Anchors* to obtain the ratings, without randomising the order of sentences within a pair. Sampling will continue to be extended to provide better coverage of the general population than student-only groups.

References

Barzilay, R. & Mckeown, K. (2005) Sentence Fusion for Multidocument News Summarization. *Computational Linguistics, 2005.* , 31, 297-328.

Bickmore, T. & Giorgino, T. (2006) Health dialog systems for patients and consumers. *J Biomed Inform. 39(5)*, 39, 556-571.

Blalock, H. M. (1979) *Social Statistics* McGraw-Hill Inc.

Charles, W. G. (2000) Contextual Correlates of Meaning. *Applied Psycholinguistics* 21, 505-524.

Crockett, K., Bandar, Z., O'shea, J. & Mclean, D. (2009) Bullying and Debt: Developing Novel Applications of Dialogue Systems. *Knowledge and Reasoning in Practical Dialogue Systems (IJCAI).* Pasadena, CA, IJCAI.

Deerwester, S., Dumais, S., Furnas, G. W., Harshman, R., Landauer, T., Lochbaum, K. & Streeter, L. (1989) Computer information retrieval using Latent Semantic Structure. IN OFFICE, U. S. P. (Ed.) United States of America, Bell Communications Research Inc.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science,* 41, 391-407.

Erozel, G., Cicekli, N. K. & Cicekli, I. (2008) Natural language querying for video databases. *Information Sciences* 178 2534–2552.

Feng, J., Zhou, Y. & Martin, T. (2008) Sentence Similarity based on Relevance. *IPMU 2008.* Torremolinos.

Gleitman, L. R., Gleitman, H., Miller, C. & Ostrin, R. (1996) Similar, and similar concepts. *Cognition,* 58, 321-376.

Gurevych, I. & Strube, M. (2004) Semantic Similarity Applied to Spoken Dialogue Summarization. *20th International Conference on Computational Linguistics.* Geneva, Switzerland.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y. & Mckeown, K. R. (2001) SimFinder: A Flexible Clustering Tool for Summarization. *Workshop on Automatic Summarization, Annual Meeting of the North American Association for Computational Linguistics (NAACL-01).* Pittsburgh, Pennsylvania.

Hu, W., Qu, Y. & Cheng, G. (2008) Matching large ontologies: A divide-and-conquer approach. *Data & Knowledge Engineering,* 67, 140-160.

Islam, A. & Inkpen, D. (2008) Semantic Text Similarity using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data,* 2.

Jeon, J., W.B., C. & Lee, J. (2005) Finding Similar Questions in Large Question and Answer Archives. *The ACM Fourteenth Conference on Information and Knowledge Management (CIKM 2005).*

Jijkoun, V. & De Rijke, M. (2005) Recognizing Textual Entailment Using Lexical Similarity. *The PASCAL RTE Challenge.*

Kennedy, A. & Szpakowitz, S. (2008) Evaluating Roget's Thesauri. *ACL-08 HLT.* Columbus, Ohio.

Kilgarriff, A. (1998) Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs

Computer Speech and Language, Special Issue on Evaluation., 12.

Klein, D. & Murphy, G. (2002) Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language,* 47, 548-570.

Laham, D. (1998) Latent Semantic Analysis @ CU Boulder. Boulder Colorado.

Lamontagne, L. & Lapalme, G. (2004) Textual Reuse for Email Response. *Lecture Notes in Computer Science* 3155, 234-246.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998) An Introduction to Latent Semantic Analysis. *Discourse Processes,* 25 259-284.

Li, Y., Bandar, Z., Mclean, D. & O'shea, J. (2004) A method for measuring sentence similarity and its application to conversational agents. IN BARR, V. & MARKOV, Z. (Eds.) *The 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004),.* Miami Beach, FL. , AAAI Press.

Li, Y., Bandar, Z., Mclean, D. & O'shea, J. (2006) Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering,* 18, 1138-1150.

Manning, C. D., Raghavan, P. & Schütze, H. (Eds.) (2008) *Introduction to Information Retrieval*, Cambridge University Press.

Marton, G. & Katz, B. (2006) Using Semantic Overlap Scoring in Answering TREC Relationship Questions. *LREC 2006.* Genoa, Italy.

McNamara, T. P. & Sternberg, R. J. (1991) Processing Verbal Relations *Intelligence,* 15 193-221.

Mihalcea, R., Corley, C. & Strapparava, C. (2006) Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *Proceedings of the American Association for Artificial Intelligence (AAAI 2006).* Boston.

Miller, G. A. & Charles, W. G. (1991) Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6 1-28.

Ngai, G. & Yarowsky, D. (2000) Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking. *38th Annual Meeting of the Association for Computational Linguistics* Hong Kong, China.

O'shea, J. D., Bandar, Z., Crockett, K. & Mclean, D. (2008) A Comparative Study of Two Short Text Semantic Similarity Measures. *Lecture Notes in Artificial Intelligence.*

O'shea, K., Bandar, Z. & Crockett, K. (2009) Towards a New Generation of Conversational Agents Based on Sentence Similarity *Lecture Notes Electrical Engineering*, 39, 505-514.

Prior, A. & Bentin, S. (2003) Incidental formation of episodic associations: The importance of sentential context. . *Memory and Cognition*, 31 306-316

Quarteroni, S. & Manandhar, S. (2008) Designing an Interactive Open-Domain Question Answering System. *Natural Language Engineering* 1, 1-23.

Resnik, P. (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research,* Vol. 11, 95-130.

Resnik, P. & Diab, M. (2000) Measuring Verb Similarity. *Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000).* Philadelphia.

Rubenstein, H. & Goodenough, J. (1965) Contextual Correlates of Synonymy. *Communications of the ACM,* 8, 627-633.

Salton, G., Wong, A. & Yang, C. S. (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM,* 18, 613–620.

Shaw, R., Solway, B., Gaizauskas, R. & Greenwood, M. A. (2008) Evaluation of Automatically Reformulated Questions in Question Series. *Coling 2008* Manchester.

Su, X. & Gulla, J. A. (2006) An information retrieval approach to ontology mapping. *Data & Knowledge Engineering*, 58, 47–69.

Tversky, A. (1977) Features of Similarity. *Psychological Review*, 84 327-352.

Vigliocco, G., Vinson, D., Lewis, W. & Garrett, M. (2002) Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology,* 48, 422-488

Wiebe, J., Bruce, R. & O'hara, T. (1999) Development and use of a gold standard data set for subjectivity classifications. . *The 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).*

Yeh, J.-Y., Ke, H.-R. & Yang, W.-P. (2008) iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. *Expert Systems with Applications*, 35, 1451–1462.