A Comparative Study of Two Short Text Semantic Similarity Measures

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

Department of Computing and Mathematics, Manchester Metropolitan University,
Chester St., Manchester M1 5GD, United Kingdom
{j.d.oshea, z.Bandar, k.crockett, d.mclean}@mmu.ac.uk

Abstract. This paper describes a comparative study of STASIS and LSA. These measures of semantic similarity can be applied to short texts for use in Conversational Agents (CAs). CAs are computer programs that interact with humans through natural language dialogue. Business organizations have spent large sums of money in recent years developing them for online customer self-service, but achievements have been limited to simple FAQ systems. We believe this is due to the labour-intensive process of scripting, which could be reduced radically by the use of short-text semantic similarity measures. "Short texts" are typically 10-20 words long but are not required to be grammatically correct sentences, for example spoken utterances and text messages. We also present a benchmark data set of 65 sentence pairs with human-derived similarity ratings. This data set is the first of its kind, specifically developed to evaluate such measures and we believe it will be valuable to future researchers.

Keywords: Natural Language, Semantic Similarity, Dialogue Management, User Modeling, Benchmark, Sentence.

1 Introduction

A Conversational Agent (CA) is a computer program that interacts with a human user by means of natural language dialogue. The motivation for our work is the development of a new generation of Conversational Agents (CAs) with improved techniques for dialogue management. These techniques involve highly sophisticated algorithms for the measurement of Short Text Semantic Similarity (STSS) [1, 2]. A Short Text (ST) in typical human dialogue would be a sentence in the range of 10-20 words, bearing in mind that user utterances include other forms that fail to conform to the grammatical rules of sentences. Other applications that can benefit from STSS measures are the automatic processing of text and e-mail messages [3] and natural language interfaces to databases [4]. Academic studies include health care dialogue systems [5], real estate sales [6], phone call routing [7] and intelligent tutoring [8]. CAs will be increasingly important in the future as these applications are delivered remotely via the internet.

One of the most important applications of CAs is online customer self-service, providing the user with the kind of services that would come from a knowledgeable or

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 172–181, 2008.

experienced human. In 2005 there were at least 10 major companies operating in this area, including IBM and strategic partners of Microsoft [9]. At least 28 patents have been registered concerning Conversational Agents and closely related technologies. With so much investment in R&D, where are the tangible results? Commercial CAs are basic question answering systems, incapable of genuine mixed-initiative or extended dialogue. It is now recognized that there are genuine obstacles to the transfer of CAs from the research environment to the real world [5].

Pattern matching has been identified as one of the most common and capable methods for developing dialogues that seem to be coherent and intelligent to users [5]. Patterns are grouped in rules which in turn are contained in a script file [10]. When a script is executed user utterances are compared to the patterns and the closest match results in the relevant rule firing. This generates a response to the user and passes information to other programs making up the agent for relevant action. Creating scripts is a highly skilled craft [11], requiring the anticipation of user utterances, generation of permutations of the utterances and generalization of patterns through the replacement of selected terms by wild cards. Modifications to rules containing the patterns can impact on the performance of other rules and modern pattern matching systems contain many parameters that further modify their behaviour. The main disadvantage of pattern matching systems is the labour-intensive (and therefore costly) nature of their development.

State-based systems, popular in healthcare [5], provide an alternative form of dialogue management; undergoing state transitions triggered by the content of user utterances. In simple systems tight constraints are placed on the utterances that the users can produce. This can be done with forced choice questions (e.g. yes or no answers) or the detection of a very restricted set of highly salient speech fragments. More flexible dialogue is possible, but is not trusted when high accuracy of understanding of the user intent is required [5]. Furthermore chains of NLP processes can incur a high computational overhead creating scalability problems for real-world deployment.

We propose a completely new method for CAs, which has the ability to reduce greatly the amount of effort and skill required in generating scripts. The new scripts will be composed of rules containing a few prototype sentences. The similarity measure is used to compute a match between the user utterance and the sentences, generating a firing strength for each rule. Because the match is on the meaning of the statement rather than words, scripting will largely be reduced to identifying the set of appropriate prototype statements.

Studies of semantic similarity to date have concentrated on one of two levels of detail, either single words [12] (in particular nouns) or complete documents [2].

Because STSS is a novel approach, there are no established methods for evaluating such measures. We expect future CAs to be used in applications where high accuracy of understanding the user intent is required, where the stakes are high and where the users may present adversarial or disruptive characteristics in the conversation. Therefore it is crucial that STSS measures are validated before being incorporated into systems, to do otherwise would be building on sand. 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 STs, the only way to obtain them is through carefully constructed experiments with human participants.

This paper uses the first of a group of benchmark data sets that we are creating. It describes the process of selecting a set of sentence pairs, obtaining human ratings based on the best practice from word similarity studies and comparing two machine measures, STASIS and LSA.

The rest of this paper is organized as follows: Section 2 discusses some relevant features of semantic similarity and the two machine measures; Section 3 describes the experiments which capture the human similarity ratings and section 4 describes the comparative study. Section 5 outlines directions for future work.

2 Measures and Models of Semantic Similarity

Semantic similarity is fundamental to many experiments in fields such as natural language processing, linguistics and psychology [13],[14],[15] and is held to be a widely understood concept. Miller and Charles[16], 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, dating back to the 1960s, has been reinforced by other researchers such as Resnik [13] 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.

2.1 Relevant Features of Similarity

Empirical studies suggest that semantic similarity is a little more subtle than has been assumed. Some draw a distinction between "similarity" and "relatedness" [13], [17]. 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 has used relatedness to describe degrees of similarity in an empirical study [18].

Four forms of similarity are described by Klein and Murphy [19]: 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.

In some studies Semantic Distance (difference) is measured. Distance can be thought of as dissimilarity - the counterpart of semantic similarity. So if a study measures distance, it is taken as having measured similarity, by applying an inversion operation [20] or by looking for a negative correlation with distance instead of a positive correlation with similarity [16].

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 [20]. Gleitman et al [21] 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 having no similarity at all. Antonyms also generate high similarity values with co-occurrence measures [16].

The final interesting property of similarity is that given two pairs of identical items to rate, experimental participants appear to give higher rating to the pair with more complex features [13].

2.2 Models of Similarity, STASIS and LSA

A full description of STASIS is given in [1]. It calculates word similarity (equation 1) from a structured lexical database taking account of path length (*l*) and depth (*h*)

$$s(w_1, w_2) = e^{-\alpha l} . \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{1}$$

This is combined with word position information from a short joint word set vector (the *r* terms in equation 2) and word frequency information from a large corpus to give the overall similarity:

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

The parameters α,β and δ (which adjusts the relative contributions of semantic and word order) are all chosen empirically.

A full description of LSA is given in [22]. A large rectangular matrix is created from a set of terms and documents, then decomposed into a representation as the product of 3 other matrices (equation 3).

$$X = T_0 S_0 D_0' \tag{3}$$

S_0 is a diagonal matrix. Its members are sorted by size, small elements are set to zero leaving the *k* largest unchanged. Zero rows and columns in S_0 are deleted, with corresponding reductions in the size of T_0 and D_0, described by equation 4 – this reduction makes LSA computationally efficient.

$$X \approx Xhihat = TSD' \tag{4}$$

LSA has the facility to compare search terms, which we have used to compare the similarity of two sentences. Each term has a corresponding row vector in *Xhihat* and the dot product between the rows is a measure of their similarity:

$$XhihatXhihat' = TS^2T' \tag{5}$$

The important parameter k is chosen empirically, selecting a value which gives good information retrieval performance.

2.3 Desirable Properties of a Benchmark Dataset

2.3.1 Precision and Accuracy

The data set consists of 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.

2.3.2 Measurement Scale

The scale on which the similarity measures are made determines the statistical techniques that can be applied to them later [23]. Human similarity measures are at least ordinal, showing reasonably consistent ranking between individuals [1], groups [24] and over time [13]. 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. Absolute scales are used where there is only one way of making the measurement: counting occurrences. Word semantic similarity has always been treated as a ratio scale attribute for both machine measures and human data sets. Our sentence data set 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 Production of the Data Set

3.1 Experimental Design

Trial 1 collected ratings from 32 graduate Native English speakers to create the initial benchmark data set. The sample had a good balance of Arts/Humanities vs. Science/Engineering backgrounds. We also conducted three smaller-scale trails to investigate the importance of the randomization and semantic anchor factors as a basis for future work.

3.2 Materials

We followed the word-based procedure used in [24]. We took the first definition for 48 nouns from the Collins Cobuild dictionary [25] which uses sentence definitions derived from a large corpus of natural English. These were combined to make 65 sentence pairs in the same combinations as in [24]. Table 1 contains some examples, including two that required minor modifications to make usable sentences, *bird* and *smile*.

Table 1. Example sentence pairs derived from Rubenstein and Goodenough

Sentence pair	Cobuild Dictionary Definitions
1.cord: smile	Cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.
42.bird: crane	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A crane is a large machine that moves heavy things by lifting them in the air.
56.coast: shore	The coast is an area of land that is next to the sea. The shores or shore of a sea, lake or wide river is the land along the edge of it.
62.cemetery: graveyard	A cemetery is a place where dead people's bodies or their ashes are buried. A graveyard is an area of land, sometimes near a church, where dead people are buried.

3.3 Experimental Procedures

Each of the 65 sentence pairs was printed on a separate sheet. We randomized both the order of presentation of sentences within a pair and the order of sentence pairs within the questionnaire to minimize asymmetry and ordering effects. Participants were instructed to work through the questionnaire in a single pass. Following [24], the participants were presented with a pair of sentences and asked to "rate how similar they are in meaning." The rating scale ran from 0 (minimum similarity) to 4.0 (maximum similarity). We also included the statement "You can use the first decimal place, for example if you think the similarity is half way between 3.0 and 4.0 you can use a value like 3.5." to emphasize the linearity of the judgment. We used the Semantic Anchors in table 2, developed by Charles [18] to establish interval scale properties. Note however, that anchor 3.0 was tested but not used by Charles.

Table 2. Semantic anchors adopted from Charles

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 full data set can be downloaded from [26].

4 Application of the Data Set

Trial 2 compared the ratings produced by STASIS and LSA with those from the human raters in the benchmark data set.

4.1 Materials and Procedure

We used a subset of the 65 sentence pairs described in section 3. This subset was the same 30 sentence pairs used in [1]. The data set contains a large number of low-similarity sentence pairs (46 pairs in the range 0.0 to 1.0), so we sampled across the low end of the range at approximately equal intervals to counter this bias. STASIS ratings were obtained directly, by running the sentence pairs through the algorithm, LSA ratings were obtained by submitting the sentence pairs through the LSA portal [27].

Table 3. Human, STASIS and LSA similarity measures for 30 sentence pairs

Sentence Pair	Human	STASIS	LSA
1.cord:smile	0.01	0.329	0.51
5.autograph:shore	0.005	0.287	0.53
9.asylum:fruit	0.005	0.209	0.505
13.boy:rooster	0.108	0.53	0.535
17.coast:forest	0.063	0.356	0.575
21.boy:sage	0.043	0.512	0.53
25.forest:graveyard	0.065	0.546	0.595
29.bird:woodland	0.013	0.335	0.505
33.hill:woodland	0.145	0.59	0.81
37.magician:oracle	0.13	0.438	0.58
41.oracle:sage	0.283	0.428	0.575
47.furnace:stove	0.348	0.721	0.715
48.magician:wizard	0.355	0.641	0.615
49.hill:mound	0.293	0.739	0.54
50.cord:string	0.47	0.685	0.675
51.glass:tumbler	0.138	0.649	0.725
52.grin:smile	0.485	0.493	0.695
53.serf:slave	0.483	0.394	0.83
54.journey:voyage	0.36	0.517	0.61
55.autograph:signature	0.405	0.55	0.7
56.coast:shore	0.588	0.759	0.78
57.forest:woodland	0.628	0.7	0.75
58.implement:tool	0.59	0.753	0.83
59.cock:rooster	0.863	1	0.985
60.boy:lad	0.58	0.663	0.83
61.cushion:pillow	0.523	0.662	0.63
62.cemetery:graveyard	0.773	0.729	0.74
63.automobile:car	0.558	0.639	0.87
64.midday:noon	0.955	0.998	1
65.gem: jewel	0.653	0.831	0.86

4.2 Results and Discussion

The human similarity measures from trial 1 are shown with the corresponding machine measures in Table 3. STASIS produces results in the range 0 to +1. LSA produced some negative results implying that it covers a true Cosine range of -1 to +1. All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 4 illustrates the agreement of both of the machine measures with human perception by calculating the product-moment correlation coefficient between the machine rating and the average rating from the human participants over the data set. We also used leave-one-out resampling to calculate the product-moment correlation coefficient for each of the human raters with the rest of the participants to establish a normative value with upper and lower bounds for performance.

Table 4. Product-moment correlation coefficients with mean human similarity ratings

	Correlation *r*	Comment
STASIS	0.816	With average of all 32 participants, significant at 0.01 level
LSA	0.838	With average of all 32 participants, significant at 0.01 level
Average Participant	0.825	Mean of individuals with group (n=32, leave-one-out resampling). Standard Deviation 0.072
Worst participant	0.594	Worst participant with group (n=32, leave-one-out resampling).
Best participant	0.921	Best participant with group (n=32, leave-one-out resampling).

Both measures have performed well with this particular data set. The normative value from the human participants (r = 0.825) sets a realistic level of expectation for the machine measures; both are close to it and LSA slightly exceeds this level. Upper and lower bounds for the expected performance are established by the performance of the best (r = 0.921) and worst (r = 0.594) performing human participants and both perform markedly better than the worst human. LSA has performed markedly better than might have been expected, despite the fact that it makes no use of word order, syntax or morphology. Plotting graphs for the subset of sentence pairs used in this test reveals STASIS is more faithful to the human ratings at low and high similarities but is brought down by its negative correlation in the similarity range 0.2 to 0.4 (r= -0.335), where LSA has a positive, if low correlation (r= 0.344). STASIS' performance in dealing with the extremes of the similarity range (particularly at the high end of the range) could be more useful in a practical CA application.

The current data set is limited to a particular type of speech act – statements (and in particular definitions). However, one of the basic tasks of a CA is to put a question to the user which can be answered with one of a number of statements. For example "What is the nature of your debt problem?" could be answered with "I am still waiting for my student loan.", "My direct debit has not gone through." Etc. This benchmark

can be considered to be a form of *pons asinorum* - a measure that does not perform well will be of little use in a CA.

Overall, we consider that humans would find similarity judgements made using these algorithms to be reasonable and consistent with human judgement. However, it might be argued that if we could improve the quality of human data, we could set a more demanding target for the machine measures.

5 Future Work

We are in the early stages of a completely new study designed to give a more comprehensive set of sentence pair ratings. The current data set is limited to a particular type of speech act – statements (and in particular definitions). We will extend this to include a fuller range of relevant acts such as questions and instructions. We will draw on a variety of techniques ranging from conventional grammar to neuropsychology to construct a dataset with better representation of the semantic space and also include linguistic properties such as polysemy, homophony, affect and word frequency. Three subsidiary trials (n=18) have provided evidence of a data collection technique, card sorting combined with semantic anchors, which should reduce the noise in the data substantially. We have also gained some insights into improving the wording of the instructions to participants.

References

1. Li, Y., et al.: Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering 18(8), 1138–1150 (2006)
2. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)
3. Lapalme, G., Lamontagne, L.: Textual Reuse for Email Response. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 242–256. Springer, Heidelberg (2004)
4. Glass, J., et al.: A Framework for Developing Conversational User Interfaces. In: Fourth International Conference on Computer-Aided Design of User Interfaces, Funchal, Isle of Madeira, Portugal (2004)
5. Bickmore, T., Giorgino, T.: Health dialog systems for patients and consumers. J. Biomed. Inform. 39(5), 556–571 (2006)
6. Cassell, J., et al.: Embodied Conversational Agents (2000)
7. Gorin, A.L., Riccardi, G., Wright, J.H.: How I help you? Speech Communication 23, 113–127 (1997)
8. Graesser, A.C., et al.: AutoTutor: An Intelligent Tutoring System With Mixed Initiative Dialogue. IEEE Transactions on Education 48(4), 612–618 (2005)
9. McGeary, Z., et al.: Online Self-service: The Slow Road to Search Effectiveness, in Customer Relationship Management (2005)
10. Sammut, C.: Managing Context in a Conversational Agent. Electronic Transactions in Artificial Intelligence Volume, 191–201 (2001)
11. Michie, D.: Return of the Imitation Game. Electronic Transactions in Artificial Intelligence Volume, 205–220 (2001)

12. Resnik, P., Diab, M.: Measuring Verb Similarity. In: Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI 2000), Philadelphia (2000)
13. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
14. Prior, A., Bentin, S.: Incidental formation of episodic associations: The importance of sentential context. Memory and Cognition 31, 306–316 (2003)
15. McNamara, T.P., Sternberg, R.J.: Processing Verbal Relations. Intelligence 15, 193–221 (1991)
16. Miller, G.A., Charles, W.G.: Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
17. Viggliocho, G., et al.: Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. Cognition 85, B1–B69 (2002)
18. Charles, W.G.: Contextual Correlates of Meaning. Applied Psycholinguistics 21, 505–524 (2000)
19. Klein, D., Murphy, G.: Paper has been my ruin: conceptual relations of polysemous senses. Journal of Memory and Language 47(4), 548–570 (2002)
20. Tversky, A.: Features of Similarity. Psychological Review 84(4), 327–352 (1977)
21. Gleitman, L.R., et al.: Similar, and similar concepts. Cognition 58, 321–376 (1996)
22. Deerwester, S., et al.: Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
23. Blalock, H.M.: Social Statistics. McGraw-Hill Inc., New York (1979)
24. Rubenstein, H., Goodenough, J.: Contextual Correlates of Synonymy. Communications of the ACM 8(10), 627–633 (1965)
25. Sinclair, J.: Collins Cobuild English Dictionary for Advanced Learners, 3rd edn. Harper Collins, New York (2001)
26. O'Shea, J.D.: http://www.docm.mmu.ac.uk/STAFF/J.Oshea/
27. Laham, D.: (October 1998) (cited 30/09/2007), http://lsa.colorado.edu/