International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-1 Issue-1, April 2012

A Comparison of Different Measures to Evaluate the Semantic Relatedness of Text and its Application

S.Vijay

Abstract: ***This paper presents a knowledge-based and experiment-based method for measuring the semantic similarity of texts. While there is a large body of previous work focused on finding the semantic similarity of concepts and words, the application of these word oriented methods to text similarity has not been yet explored. Five different proposed measures of similarity or semantic distance in WordNet were experimentally compared by examining their performance in a real-word spelling correction system.***

Index Terms: Dictionary-based, Information-based, Lexical-based, WordNet.

I. INTRODUCTION

Measures of text similarity have been used for a long time in applications in natural language processing and related areas. One of the earliest applications of text similarity is perhaps the vectorial model in information retrieval, where the document most relevant to an input query is determined by ranking documents in a collection in reversed order of their similarity to the given query (Salton and Lesk, 1971).

Text similarity has been also used for relevance feedback and text classification (Rocchio, 1971), word sense disambiguation (Lesk, 1986), and more recently for extractive summarization (Salton et al., 1997b), and methods for automatic evaluation of machine translation (Papineni et al., 2002) or text summarization (Lin and Hovy, 2003). The need to determine semantic relatedness between two lexically expressed concepts is a problem that pervades much in language processing. Measures of relatedness or distance are used in such applications as word sense disambiguation, determining the structure of texts, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and the automatic correction of word errors in text. It is unclear as to how to assess the relative merits of the many approaches that have been proposed for determining lexical semantic relatedness.

Revised Manuscript Received on 30 April 2013.
*** Correspondence Author**
Vijay.S, Department of Information Technology, Priyadarshini Engineering College, Vaniyambadi, India. E-mail: vijaypec2011@yahoo.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

The purpose of this paper is to compare the performance of a number of measures of semantic relatedness that have been proposed for use in applications in natural language processing and information retrieval.

II. APPROACHES TO MEASURE SEMANTIC RELATEDNESS

Seven approaches to measure semantic relatedness are:

A. Lexical Resource-based Approach

A approach to measure semantic relatedness with the use of lexical resource construe the resource, in one way or another, as a network, and then base the measure of relatedness on properties of paths.

B. Dictionary-based Approach

Kozima and Furugori (1993) turned the *Longman Dictionary of Contemporary English* (LDOCE) (Procter 1978) into a network by created a node for every headword and linking each node to the nodes for all the words used in its definition. Kozima and Ito (1997) built on this work to derive a **context-sensitive,** or **dynamic,** measure that takes into account the "associative direction" of a given word pair. For example, the context *{car, bus}* imposes the associative direction of vehicle (close words are then likely to include *taxi, railway, airplane*, etc.), whereas the context *{car, engine}* imposes the direction of components of car (*tire, seat, headlight*, etc.).

C. Roget-structured Thesauri Approach

Roget-structured thesauri, such as *Roget's Thesaurus* itself, the *Macquarie Thesaurus* (Bernard 1986), and others, group words in a structure based on **categories** within which there are several levels of finer clustering. The categories themselves are grouped into a number of broad, loosely defined classes.

D. WordNet Approach

Nouns, verbs, adjectives, and adverbs are each organized into networks of synonym sets (**synsets**) that each represent one underlying lexical concept and are interlinked with a variety of relations.

E. Computing Taxonomic Path Length Approach

A simple way to compute semantic relatedness in a taxonomy such as WordNet is to view it as a graph and identify relatedness with path length between the concepts:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved

"The shorter the path from one node to another, the more similar they are" Hirst and St-Onge (1998; St-Onge 1995) adapted Morris and Hirst's (1991) semantic distance algorithm from *Roget's Thesaurus* to WordNet. They distinguished two strengths of semantic relations in WordNet. Two words are **strongly related** if one of the following holds:
1. They have a synset in common (for example, *human* and *person*).
2. They are associated with two different synsets that are connected by the antonymy relation (for example, *precursor* and *successor*).
3. One of the words is a compound (or a phrase) that includes the other and "there is any kind of link at all between a synset associated with each word" (for example, *school* and *private school*).

F. Scaling the Network Approach

Despite its apparent simplicity, a widely acknowledged problem with the edge-counting approach is that it typically "relies on the notion that links in the taxonomy represent uniform distances", which is typically not true: "there is a wide variability in the 'distance' covered by a single taxonomic link, particularly when certain sub-taxonomies (e.g., biological categories) are much denser than others" (Resnik 1995).

G. Information-based and Integrated Approach

The final approach that we counters problem inherent in a general ontology by incorporating an additional, and qualitatively different, knowledge source, namely information from a corpus.

In this paper we discuss about two different approaches

- Scaling of network approach
- Information-based and Integrated Approach

These two approaches together constitute five different sub approaches for measuring the semantic relatedness. The first is claimed as a measure of semantic relatedness because it uses all relations in WordNet; the others are claimed only as measures of similarity because they use only the hyponymy relation.

Hirst–St-Onge: The idea behind Hirst and St-Onge's (1998) measure of semantic relatedness is that two lexicalized concepts are semantically close if their WordNet synsets are connected by a path that is not too long and that "does not change direction too often". The strength of the relationship is given by:

$$\text{rel}_{HS}(c1,c2) = C - \text{path length} - k \times d$$

Where d is the number of changes of direction in the path, and C and k are constants; if no such path exists, $\text{rel}_{HS}(c1,c2)$ is zero and the synsets are deemed unrelated.

The **Leacock & Chodorow** (Leacock and Chodorow, 1998) similarity is determined as:

$$Sim_{lch} = -\log \frac{length}{2*D} \quad (1)$$

Where length is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy.

The measure introduced by **Resnik** (Resnik, 1995) returns the information content (IC) of the LCS of two concepts:

$$Sim_{res} = IC(LCS) \quad (2)$$

Where IC is defined as:

$$IC(c) = -\log p(c)$$

and $P(c)$ is the probability of encountering an instance of concept c in a large corpus.

Jiang & Conrath (Jiang and Conrath, 1997), which returns a score determined by:

$$Sim_{jnc} = \frac{1}{IC(concept_1) + IC(concept_2) - 2*IC(LCS)} \quad (3)$$

The next measure we use in our experiments is the metric introduced by **Lin** (Lin, 1998), which builds on Resnik's measure of similarity, and adds a normalization factor consisting of the information content of the two input concepts:

$$Sim_{lin} = \frac{2*IC(LCS)}{IC(concept_1) + IC(concept_2)} \quad (4)$$

Table 1: The coefficients of correlation between human ratings of similarity (by Miller and Charles and by Rubenstein and Goodenough) and the five computational measures.

Similarity measure	M&C	R&G
Hirst and St-Onge (rel_{HS})	.744	.786
Leacock and Chodorow (sim_{LC})	.816	.838
Resnik (sim_R)	.774	.779
Jiang and Conrath ($dist_{JC}$)	.850	.781
Lin (sim_L)	.829	.819

III. Evaluation Methods

Three kinds of evaluation methods are prevalent in this literature.

The first kind (Wei 1993; Lin 1998b) is a (chiefly) theoretical examination of a proposed measure for those mathematical properties thought desirable, such as whether it is a metric (or the inverse of a metric), whether it has singularities, whether its parameter-projections are smooth functions, and so on. In our opinion, such analyses act at best as a coarse filter in the comparison of a set of measures and an even coarser one in the assessment of a single measure.

The second kind of evaluation is comparison with human judgments. Insofar as human judgments of similarity and relatedness are deemed to be correct by definition, this clearly gives the best assessment of the "goodness" of a measure.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved

Its main drawback lies in the difficulty of obtaining a large set of reliable, subject-independent judgments for comparison—designing
a psycholinguistic experiment, validating its results, and so on.

The third approach is to evaluate the measures with respect to their performance in the framework of a particular application. If some particular NLP system requires a measure of semantic relatedness, we can compare different measures by seeing which one the system is most effective with, while holding all other aspects of the system constant.

IV. Comparison With Knowledge Based Ratings Of Similarity

A. Data

The relationship between similarity of context and similarity of measuring, Rubenstein and Good enough (1965) obtained "synonymy judgments" of 51 human subjects on 65 pairs of words. The pairs ranged from "highly synonymous" (*gem–jewel*) to "semantically unrelated" (*noon–string*).Subjects were asked to rate them on the scale of 0.0 to 4.0 according to their "similarity of meaning" (such as in the pair *journey–car*). Miller and Charles (1991) extracted 30 pairs from the original 65, taking 10 from the "high level (between 3 and 4...), 10 from the intermediate level (between 1 and 3), and 10 from the low level (0 to 1) of semantic similarity", and then obtained similarity judgments from 38 subjects.

B. Results

For each of our five implemented measures, we obtain similarity or scores for the human rated pairs. We follow Resnik (1995) in summarizing the comparison results by means of coefficient of correlation with the reported human ratings for each computational measure. While the difference between the values of the highest and lowest correlation coefficients in the second column of Table 1 is of the order of 0.1, all of the coefficients compare quite favorably with Resnik's estimate of 0.88 as the upper bound on performance of a computational measure. Furthermore, the difference halves as we consider the larger Rubenstein–Goodenough dataset. In fact, the measures are divided in their reaction to increasing the size of the dataset: the correlation improves for rel_{HS}, sim_{LC}, and sim_{R} but deteriorates for $dist_{JC}$ and sim_{L}.

C. Discussion and Limitations

While comparison with human judgments is the ideal way to evaluate a measure of similarity or semantic relatedness, in practice the tiny amount of data available (and only for similarity, not relatedness) is quite inadequate. But constructing a large-enough set of pairs and obtaining human judgments on them would be a very large task.

But even more importantly, there are serious methodological problems with this whole approach. It was implicit in the Rubenstein–Goodenough and Miller–Charles experiments that subjects were to use the dominant sense of the target words. But what we are really interested in is the relationship between the concepts for which the words are merely surrogates; the human judgments that we need are of the relatedness of word-senses, not words. So the experimental situation would need to set up contexts that bias the sense selection for each target word and yet don't bias the subject's judgment of their *a priori* relationship, an almost self-contradictory situation.

V. An Application-Based Evaluation Of Measures Of Relatedness

A different approach to the evaluation of similarity and relatedness measures that tries to overcome the problems of comparison to human judgments that were described in the previous section. Here, we compare the measures through the performance of an application that uses them: the detection and correction of real-word spelling errors in open-class words, i.e., **malapropisms**.

A. Malapropism detection as a Testbed

Our malapropism (Budanitsky and Hirst, in preparation) is based on the idea behind that of Hirst and St-Onge (1998): semantic anomalies that can be removed by small changes to spelling words are (crudely) disambiguated where possible by accepting senses that are semantically related to possible senses of other nearby words. If all senses of any open-class, non–stop-list word that occurs only once in the text are found to be semantically unrelated to accepted senses of all other nearby words, but some sense of a spelling variation3 of that word would be related (or is identical to another token in the context), then it is hypothesized that the original word is an error and the variation is what the writer intended; the user is warned of this possibility.4 For example, if no nearby word in a text is related to diary but one or more are related to dairy, we suggest to the user that it is the latter that was intended. The exact window size implied by "nearby" is a parameter to the algorithm.

This method makes the following assumptions:

- A real-word spelling error is unlikely to be semantically related to the text.5
- Frequently, the writer's intended word will be semantically related to nearby words.
- It is unlikely that an intended word that is semantically unrelated to all those nearby will have a spelling variation that *is* related.

While the performance of the malapropism corrector is inherently limited by these assumptions, we can nonetheless evaluate measures of semantic relatedness by comparing their effect on its performance, as its limitations affect all measures equally.

B. Method

Hirst and St-Onge (1998), 500 articles from the *Wall Street Journal* corpus and, has taken after removing proper nouns and stop-list words from consideration, replaced one word in every 200 with a spelling variation, choosing always WordNet nouns with at least one spelling variation. This gave us a corpus with 107,233 such words, 1408 of which were malapropisms. Next detect and correct the malapropisms by the algorithm , using five measures of semantic relatedness. Four different *search scopes were used* (window sizes): just the paragraph containing the target word (scope = 1); that paragraph plus one or two adjacent paragraphs on each side (scope = 3 and 5); and the

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved

entire article (scope = MAX).

Each of the measures tested returns a numerical relatedness or similarity value, not the boolean *related–unrelated* judgment required by the algorithm, and the values from the different measures are incommensurate. Therefore set the threshold of relatedness of each measure at the value at which it separated the higher level of Rubenstein–Goodenough pairs from the lower level.

C. Results

Malapropism detection as viewed as a retrieval task and evaluated in terms of precision, recall, and *F*-measure. Observe that semantic relatedness is used at two different places in the algorithm—to judge whether an original word of the text is related to any nearby word and to judge whether a spelling variation is related and success in malapropism detection requires success at both stages. For the first stage, we say that a word is *suspected* of being a malapropism (and the word is a *suspect*) if it is judged to be unrelated to other words nearby; the word is a *true suspect* if it is indeed a malapropism. At the second stage, we say that an *alarm* is raised when a spelling variation of a suspect is judged to be related to a nearby word; and if an alarm word is a malapropism, we say that the alarm is a *true alarm* and that the malapropism has been *detected*. Then we can define precision (*P*), recall (*R*), and *F*-measure (*F*) for suspicion (*S*), involving only the first stage, as follows:

$$ps = \frac{number\ \text{of true suspects}}{number\ \text{of suspects}} \qquad (5)$$

$$R_s = \frac{number\ \text{of true suspects}}{number\ \text{of malapropisms in text}} \qquad (6)$$

$$Fs = \frac{2 \times Ps \times Rs}{Ps + Rs} \qquad (7)$$

and for detection (*D*), involving both stages, analogously (replacing *suspects* with *alarms*).

D. Suspicion

The results for suspicion just identifying words that have no semantically related word nearby. The chance of finding some word that is judged to be related to the target word will increase with the size of the scope of the search (with a large enough scope, *e.g.*, a complete book, we would probably find a relative for just about any word). So recall to decrease as scope increases, because some relationships will be found even for malapropisms (*i.e.*, there will be more false negatives). But the precision will increase with scope, as it becomes more likely that (genuine) relationships will be found for malapropisms (*i.e.*, there will be fewer false positives), and this factor will outweigh the decrease in the overall number of suspects found.

Compute suspicion precision, recall, and *F* for each of the 5 X 4 combinations of measure and scope. The values of precision range from 3.3% to 11%, with a mean of 6.2%, increasing with scope, as expected, for all measures except Hirst–St-Onge. The values of recall range from just under 6%to more than 72%, with a mean of 39.7%, decreasing with scope, as expected. *F* ranges from 5% to 14%, with a mean of just under 10%. Even though the lower ends of these ranges appear unimpressive, they are still significantly ($p < .001$) better than chance, for which all measures are 1.29%. The value for precision is inherently limited by the likelihood that, especially for small search scopes, there will be words other than our deliberate malapropisms that are genuinely unrelated to all others in the scope.

Table 4: Precision (*PS*), recall (*RS*), and *F*-measure (*FS*) for malapropism suspicion with five measures of semantic relatedness, varying the scope of the search for related words to 1, 3, or 5 paragraphs or the complete news article (MAX).

Measure	Scope	P_S	R_S	F_S
Hirst–St-Onge	1	.056	.298	.091
	3	.067	.159	.089
	5	.069	.114	.079
	MAX	.051	.059	.049
Jiang–Conrath	1	.064	.536	.112
	3	.086	.383	.135
	5	.097	.326	.141
	MAX	.111	.233	.137
Leacock–Chodorow	1	.042	.702	.079
	3	.052	.535	.094
	5	.058	.463	.101
	MAX	.073	.356	.115
Lin	1	.047	.579	.086
	3	.062	.421	.105
	5	.067	.350	.110
	MAX	.078	.253	.110
Resnik	1	.033	.727	.063
	3	.038	.589	.070
	5	.039	.490	.072
	MAX	.043	.366	.075

Because it combines recall and precision, we focus on the results for F_S by measure and scope (see Figure 1) to determine whether the performance of the five measures was significantly different and whether scope made a significant difference.

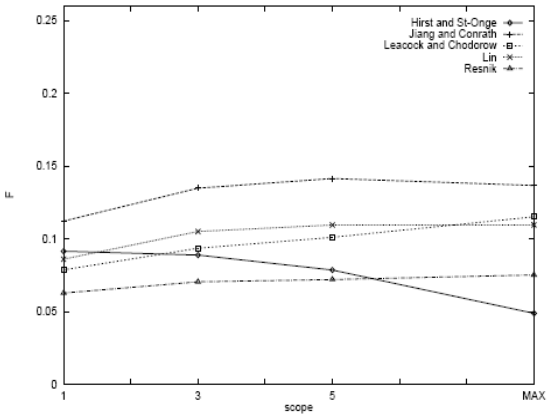

Figure 1: Suspicion *F*-measure (F_S), by measure and scope.

Scope differences: The analysis confirms only that the methods perform significantly better with scope 5 than scope 1; for Lin, that scope 3

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved

is significantly better than scope 1; for Leacock– Chodorow, that 3 is significantly better than 1 and MAX better than 3; and for Hirst–St-Onge, that MAX is worse than 3. From the standpoint of simple detection of unrelatedness (suspicion in malapropism detection), these data point to overall optimality of scopes 3 or 5.

Measure differences: Jiang–Conrath significantly out performs the others for all scopes (except for Leacock–Chodorow and Lin for scope MAX, where it does better but not significantly so), followed by Lin and Leacock–Chodorow (whose performances are not significantly different),in turn followed by Resnik. Hirst–St-Onge, with its irregular behavior, performs close to Lin and Leacock–Chodorow for scopes 1 and 3 but falls behind as the scope size increases, finishing worst for scope MAX. Thus the Jiang–Conrath measure with scope 5 is optimal for the suspicion phase.

E. Detection

During the detection phase, the suspects are winnowed by checking the spelling variations of each for relatedness to their context. Since (true) alarms can only result from (true) suspects, recall cannot increase from that for suspicion (*cf* equation 6). However, if a given measure of semantic relatedness is any good, we expect the proportion of false alarms to reduce more considerably—far fewer false suspects will become alarms than true suspects thus resulting in higher precision for detection than for suspicion (*cf* equation 5).

We computed detection precision, recall, and *F* for each measure–scope combination by the same method as for suspicion. The values of recall range from 5.9% to over 60%. While these values are, as expected, lower (by 1–16 percentage points) than those for suspicion recall, the decline is statistically significant for only 3 out of the 20 combinations. The values of precision range from 6.7% to just under 25%, increasing, as expected, from suspicion precision; each combination improves by from 1 to 14 percentage points; the improvement is statistically significant for 18 out of the 20 combinations.

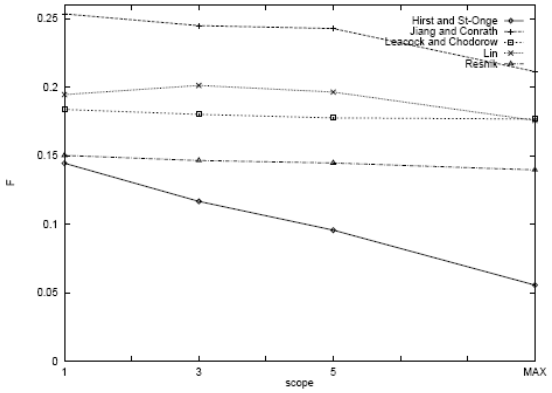

Figure 2: Detection *F*-measure (*FD*), by measure and scope.

Furthermore, the increase in precision outweighs the decline in recall, and *F*, which ranges from 6% to 25%, increases by 7.6% on average; the increase is significant for 17 out of the 20 combinations. Again, even the lower ends of the *P*, *R*, and *F* ranges are significantly ($p < .001$) better than chance (which again is 1.29%for all measures), and the best results are quite impressive (*e.g.*, 18% precision, 50%recall for Jiang–Conrath at scope=1, which had the highest *FD*, though not the highest precision or recall), despite the limitations described in Section 5.1.

Scope differences: Our analysis of scope differences in *F* shows a somewhat different picture for detection from that for suspicion: there are significant differences between scopes only for the Hirst–St-Onge measure. The *F* graphs of the other four methods thus are not significantly different from being flat, and we can choose 1 as the optimal scope.

Table 5: Precision (*PD*), recall (*RD*), and *F*-measure (*FD*) for malapropism detection with five measures of semantic relatedness, varying the scope of the search for related words to 1, 3, or 5 paragraphs or the complete news article (MAX).

Measure	Scope	P_D	R_D	F_D
Hirst–St-Onge	1	.105	.286	.145
	3	.107	.159	.117
	5	.101	.114	.096
	MAX	.067	.059	.056
Jiang–Conrath	1	.184	.498	.254
	3	.205	.372	.245
	5	.219	.322	.243
	MAX	.247	.231	.211
Leacock–Chodorow	1	.111	.609	.184
	3	.115	.499	.180
	5	.118	.440	.178
	MAX	.132	.338	.177
Lin	1	.125	.514	.195
	3	.145	.398	.201
	5	.150	.335	.197
	MAX	.168	.242	.176
Resnik	1	.088	.562	.150
	3	.087	.512	.146
	5	.088	.454	.145
	MAX	.093	.344	.140

Measure differences: The relative position of each measure's *F* graph for detection is identical to that for suspicion, except for Hirst–St-Onge, which slides further down. Statistical testing confirms this, with Jiang–Conrath leading, followed by Lin and Leacock–Chodorow together, Resnik, and then Hirst–St-Onge. Thus Jiang and Conrath's method with scope = 1 proves to be the optimal parameter combination for our malapropism detector.

F. Interpretation of the results

In the interpretation, focus is based largely on the results for suspicion; those for detection, though somewhat opaque on their own, both add to the pool of relatedness judgments.

The Resnik measure's comparatively poor precision and good recall suggest that the measure simply marks too many words as potential malapropisms—it 'underrelates', being far too conservative in its judgments of relatedness. For example, it was the only measure that flagged *crowd* as a suspect in a context in which all the other measures found it to be related to *house*: **crowd** IS-A **gathering / assemblage** SUBSUMES **house / household / family / menage**.8 Indeed, for every scope, Resnik's measure generates more suspects than any other measure—*e.g.*, an average of 62.5 per article for scope = 1, compared to

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved

the average of 37 for the other measures. The Leacock–Chodorow measure's superior precision and comparable recall (the former difference is statistically significant, the latter is not), which result in a statistically-significantly better *F*-value, indicate its better discerning ability.

The same comparison can be made between the Lin and Jiang–Conrath measures (the latter being best overall; see above). The Lin and Leacock–Chodorow measures, in turn, have statistically indistinguishable values of *F* and hence similar ratios of errors to true positives.

Finally, the steady downward slope that distinguishes the *F*-graph of Hirst–St-Onge from those of the other four measures in Figure 1 evidently reflects the corresponding difference in precision behavior, which is a result of the measure's 'over-relating'—it is far too promiscuous in its judgments of relatedness. For example, it was the only measure that considered *cation* (a malapropism for *nation*) to be related to *group*: **cation** IS-A **ion** IS-A **atom** PART-OF **molecule** HAS-A **group / radical** ('two or more atoms bound together as a single unit and forming part of a molecule'). Because of its promiscuity, the Hirst–St-Onge measure's mean number of suspects for scope = 1 is 15.07, well below the average, and moreover it drops to one-ninth of that, 1.75, at scope = MAX; the number of articles without a single suspect grows from 1 to 93.

VI. Conclusion

We have shown that there are considerable differences in the performance of five proposed measures of semantic relatedness. Jiang and Conrath's measure was shown to be best overall. It remains unclear, however, just why it performed so much better than Lin's measure, which is but a different arithmetic combination of the same terms.

All the measures that we looked at, except for that of Hirst and St-Onge, were, strictly speaking, similarity measures, considering only the hyponymy hierarchy of WordNet, rather than measures of more-general semantic relatedness. Yet the Hirst–St-Onge measure gave by far the worst performance largely because it ventured beyond hyponymy into other lexical relations in Word-Net, and in practice this hurt more often than it helped. Nonetheless, it remains a strong intuition that hyponymy is only one part of semantic relatedness; meronymy, such as *wheel–car*, is most definitely an indicator of semantic relatedness, and, *a fortiori*, semantic relatedness can arise from little more than common or stereotypical associations or statistical co-occurrence in real life (for example, *penguin–Antarctica; birthday–candle; sleep–pajamas*). Perhaps, then, the problem with the Hirst–St-Onge measure lies more in its tendency to wander too far than in its use of all WordNet relationships, and a more-constrained version might perform much better. More than the other methods, it is vulnerable to the promiscuity of Word-Net itself WordNet's tendency to give obscure senses equal prominence to more-frequent senses, which limits our crude and greedy approach to disambiguation and this bends our assumption that, despite the limitations of the malapropism detection method, our comparison of the measures occurs on a "level playing field".

Because all of the measures except Hirst–St-Onge returned a similarity value rather than a *yes–no* relatedness judgement, our comparison of the measures was constrained by the need to find, for each measure, a point in its range to serve as the threshold of relatedness. Our use of the relatedness bands of the human-judgment norms was, we feel, an elegant solution to this problem, but the accuracy of the calibration of the threshold is inherently limited by the fact that the data covers just a few dozen pairs of words. More data is needed for more accurate calibration.

Our use of malapropism detection as a testbed has proved to be an effective way of comparing the measures of semantic distance. (In particular, the results with the Jiang–Conrath measure show that the method approaches practical usability; for more discussion of this, see Budanitsky and Hirst (in preparation).) By examining the ability of the measures to find deliberate malapropisms introduced into text presumed to be otherwise coherent, we have been able to show their relative strengths and weaknesses.

Acknowledgment

This research was supported financially by the Natural Sciences and Research Council of Canada, the Ontario Graduate Scholarship Program, and the University of Toronto. For discussions, comments, and advice, we are grateful to Mark Chignell, StephenGreen, Jay Jiang,Keith Knight, Claudia Leacock, Dekang Lin, Radford Neal, Manabu Okumura, Philip Resnik, David St-Onge, and Michael Sussna.

References

1. Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 16–22, Copenhagen.
2. Alan Agresti and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences* (third edition). Prentice-Hall.
3. Alexander Budanitsky. 1999. *Lexical Semantic Relatedness and its Application in Natural Language Processing*, technical report CSRG-390, Department of Computer Science, University of Toronto, August 1999.
4. Alexander Budanitsky and Graeme Hirst. In preparation. Semantic relatedness between lexicalized concepts.
5. Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.
6. Victoria A. Fromkin. 1980. *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand.* Academic Press.
7. M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.
8. Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum 1998, pp. 305–332.
9. [9] Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press.
10. [10] Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.
11. Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum 1998, pp. 265–283.
12. Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*,Madison, WI.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved

13. George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.
14. Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1): 21–48.
15. Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems,Man, and Cybernetics*, 19(1): 17–30.
16. Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal.
17. Herbert Rubenstein and John B. Goodenough. 1965.Contextual correlates of synonymy. *Communications of the ACM*, 8(10): 627–633.
18. Michael Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)*, pages 67–74, Arlington, VA.
19. Agresti, Alan and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences*. Prentice Hall, Upper Saddle River, NJ, 3rd edition.
20. Banerjee, Satanjeev and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, pages 805–810, August.
21. Barsalou, LawrenceW. 1983. Ad hoc categories. *Memory and Cognition*, 11:211–227.
22. Barsalou, LawrenceW. 1989. Intra-concept similarity and its implications for interconcept similarity. In Stella Vosniadou andAndrew Ortony, editors, *Similarity and Analogical Reasoning*. Cambridge University Press, pages 76–121.

AUTHOR BIOGRAPHY

I am **Vijay. S** s/o Mr.K.P.A.Sundarajan born on 23.11.1979. I have completed my M.Sc. degree in Computer Science from University of Madras in year 2004 and M.Tech degree in Information Technology from Sathyabama University, Chennai in year 2009. I am presently working as Assistant Professor in Department of Information Technology in Priyadarshini Engineering College, Vaniyambadi. I Published my paper in International Journal of Soft Computing and Engineering. My research interest includes Cloud Computing, Data Mining and Data Warehousing and Network Security. I have completed my 5 years experience till date in teaching and has guided several projects and dissertations in B.Tech Courses.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved