Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures

Alexander Budanitsky and **Graeme Hirst**
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 3G4
{*abm, gh*}*@cs.toronto.edu*

Abstract

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. It was found that Jiang and Conrath's measure gave the best results overall. That of Hirst and St-Onge seriously over-related, that of Resnik seriously under-related, and those of Lin and of Leacock and Chodorow fell in between.

1 Introduction

The need to determine the *degree of semantic similarity*, or, more generally, *relatedness*, between two lexically expressed concepts is a problem that pervades much of computational linguistics. Measures of similarity or relatedness are used in such applications as word sense disambiguation, determining discourse structure, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and automatic correction of word errors in text.

The problem of formalizing and quantifying the intuitive notion of similarity has a long history in philosophy, psychology, and artificial intelligence, and many different perspectives have been suggested. Recent research on the topic in computational linguistics has emphasized the perspective of *semantic relatedness* of two lexemes in a lexical resource, or its inverse, *semantic distance*. It's important to note that semantic relatedness is a more general concept than similarity; similar entities are usually assumed to be related by virtue of their likeness (*bank–trust company*), but dissimilar entities may also be semantically related by lexical relationships such as meronymy (*car–wheel*) and antonymy (*hot–cold*), or just by any kind of functional relationship or frequent association (*pencil–paper, penguin–Antarctica*). Computational applications typically require relatedness rather than just similarity; for example, *money* and *river* are cues to the in-context meaning of *bank* that are just as good as *trust company*.

However, it is frequently unclear how to assess the relative and absolute merits of the many competing approaches that have been proposed. Our purpose in this paper is to compare the performance of several measures of semantic relatedness that have been proposed for use in NLP applications.

2 Evaluation methods

Three kinds of approaches to the evaluation of measures of similarity or semantic distance are prevalent in the literature. The first kind (*e.g.,*Wei (1993), Lin (1998)) is theoretical examination of a given measure for properties thought desirable, such as whether it is actually a metric (or the inverse of such), has singularities, etc. In our opinion, such analyses act at best as a coarse filter in the assessment of a measure or comparison of a set of measures.

The second approach is comparison with human judgments. Insofar as human judgments of similarity and relatedness are deemed correct by definition, this clearly gives the best assessment of the 'goodness' of a measure. Its main drawback lies in the difficulty of obtaining a large set of reliable, subject-independent judgments for comparison (see Section 4.3 below).

The third approach, which we follow in this paper, is to evaluate the measures with respect to their performance within a particular NLP application (see Section 5 below). Nonetheless, in our experiments, we also employ comparisons with human-judgment data, primarily to bootstrap our evaluation.

3 Network-based measures of semantic distance

Budanitsky (1999) presents an extensive survey and classification of measures of semantic relatedness. One category of such measures has been spurred by the advent of networks such as MeSH (*http://www.nlm.nih.gov/mesh/*) and WordNet. These vary from simple edge-counting (Rada et al., 1989) to attempts to factor in peculiarities of the network structure by considering link direction (Hirst and St-Onge, 1998), relative depth (Sussna, 1993; Leacock and Chodorow, 1998), and density (Agirre and Rigau, 1996). These analytic methods now face competition from statistical and machine learning techniques; but a number of hybrid approaches have been proposed that combine a *knowledge-rich* source, such as a thesaurus, with a *knowledge-poor* source, such as corpus statistics (Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997).

In selecting measures to analyze and compare, we focused on those that used WordNet (Fellbaum, 1998) as their knowledge source (to keep that as a constant) and permitted straightforward implementation as functions in a programming language. As a result, the five measures described below were selected.[1] 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. In the descriptions below, c_1 and c_2 are synsets.

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:

$$\mathrm{rel}_{\mathrm{HS}}(c_1, c_2) = C - \text{path length} - k \times d \ ,$$

where d is the number of changes of direction in the path, and C and k are constants; if no such path exists, $\mathrm{rel}_{\mathrm{HS}}(c_1, c_2)$ is zero and the synsets are deemed unrelated.

Leacock–Chodorow: Leacock and Chodorow (1998) also rely on the length $\mathrm{len}(c_1, c_2)$ of the shortest path between two synsets for their measure of similarity. However, they limit their attention to IS-A links and *scale* the path length by the overall depth D of the taxonomy:

$$\mathrm{sim}_{\mathrm{LC}}(c_1, c_2) = -\log \frac{\mathrm{len}(c_1, c_2)}{2D} \ . \qquad (1)$$

Resnik: Resnik's (1995) approach was, to our knowledge, the first to bring together ontology and corpus. Guided by the intuition that the similarity between a pair of concepts may be judged by "the extent to which they share information", Resnik defined the similarity between two concepts lexicalized in WordNet to be the *information content* of their lowest super-ordinate (most specific common subsumer) $lso(c_1, c_2)$:

$$\mathrm{sim}_{\mathrm{R}}(c_1, c_2) = -\log p(lso(c_1, c_2)) \ , \qquad (2)$$

where $p(c)$ is the probability of encountering an instance of a synset c in some specific corpus.

Jiang–Conrath: Jiang and Conrath's (1997) approach also uses the notion of information content, but in the form of the conditional probability of encountering an instance of a child-synset given an instance of a parent-synset. Thus the information content of the two nodes, as well as that of their most specific subsumer, plays a part. Notice that this formula measures semantic distance, the inverse of similarity.

$$\mathrm{dist}_{\mathrm{JC}}(c_1, c_2) = \qquad (3)$$
$$2\log(p(lso(c_1, c_2))) - (\log(p(c_1)) + \log(p(c_2))) \ .$$

Lin: Lin's (1998) similarity measure follows from his theory of similarity between arbitrary objects. It uses the same elements as $\mathrm{dist}_{\mathrm{JC}}$, but in a different fashion:

$$\mathrm{sim}_{\mathrm{L}}(c_1, c_2) = \frac{2 \times \log p(lso(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \ . \qquad (4)$$

Similarity measure	M&C	R&G
Hirst and St-Onge ($\mathrm{rel}_{\mathrm{HS}}$)	.744	.786
Leacock and Chodorow ($\mathrm{sim}_{\mathrm{LC}}$)	.816	.838
Resnik ($\mathrm{sim}_{\mathrm{R}}$)	.774	.779
Jiang and Conrath ($\mathrm{dist}_{\mathrm{JC}}$)	.850	.781
Lin ($\mathrm{sim}_{\mathrm{L}}$)	.829	.819

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.

4 Comparison with human ratings of similarity

4.1 Data

Rubenstein and Goodenough (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" and ignoring any other observed semantic relationships (such as in the pair *journey–car*). Miller and Charles (1991) subsequently 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.

4.2 Results

For each of our five implemented measures, we obtained similarity or semantic relatedness scores for the human-rated pairs mentioned above.

We follow Resnik (1995) in summarizing the comparison results by means of coefficient of correlation with the reported human ratings for each computational measure; see Table 1.[2]

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 $\mathrm{rel}_{\mathrm{HS}}$, $\mathrm{sim}_{\mathrm{LC}}$, and $\mathrm{sim}_{\mathrm{R}}$ but deteriorates for $\mathrm{dist}_{\mathrm{JC}}$ and $\mathrm{sim}_{\mathrm{L}}$.

[1] See Budanitsky (1999) for the selection rationale.

[2] Resnik (1995), Jiang and Conrath (1997), and Lin (1998) report the coefficients of correlation between their measures and the Miller–Charles ratings to be 0.7911, 0.8282, and 0.8339, respectively, which slightly differ from the corresponding figures in Table 1. These discrepancies can be explained by minor differences in implementation, different versions of WordNet (Resnik), and differences in the corpora used to obtain the frequency data (Jiang and Conrath, Lin; see Budanitsky (1999)).

4.3 Discussion

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.

5 An application-based evaluation of measures of relatedness

5.1 Malapropism detection as a testbed

We now turn to a different approach to the evaluation of similarity and relatedness measures that tries to overcome the problems described in the previous section. The idea is that naturally occurring coherent texts, by their nature, contain many instances of related pairs of words (Halliday and Hasan, 1976; Morris and Hirst, 1991; Hoey, 1991). That is, they implicitly contain human judgments of relatedness that we could use in the evaluation of our relatedness measures. But, of course, we don't know in practice just which pairs of words in a text are and aren't related. We can get around this problem, however, by deliberately perturbing the coherence of the text and looking at the ability of different relatedness measures to detect and correct the perturbations. Specifically, we will look at the detection and correction of real-word spelling errors in open-class words, that is, *malapropisms*.

Our malapropism corrector (Budanitsky and Hirst, in preparation) is based on the idea behind that of Hirst and St-Onge (1998): 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 variation[3] 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.

5.2 Method

Following Hirst and St-Onge (1998), we took 500 articles from the *Wall Street Journal* corpus and, 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. We then tried to detect and correct the malapropisms by the algorithm above, using in turn each of the five measures of semantic relatedness. For each, we used four different *search scopes* (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 entire article (scope = MAX).

Each of the measures that we 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. We 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.

5.3 Results

Malapropism detection was 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

[3] The *spelling variations* of a word *w* are those words in the lexicon derived from *w* by a single insertion, deletion, or transposition.

[4] Although it shares underlying assumptions, our algorithm differs from that of Hirst and St-Onge in its mechanisms. In particular, Hirst and St-Onge's algorithm was based on lexical chains (Morris and Hirst, 1991), whereas our algorithm regards regions of text as bags of words.

[5] In fact, there is a semantic bias in human typing errors (Fromkin, 1980), but not in our malapropism generator.

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:

$$P_S = \frac{\text{number of true suspects}}{\text{number of suspects}}, \tag{5}$$

$$R_S = \frac{\text{number of true suspects}}{\text{number of malapropisms in text}}, \tag{6}$$

$$F_S = \frac{2 \times P_S \times R_S}{P_S + R_S}, \tag{7}$$

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

5.3.1 Suspicion

We look first at the results for suspicion—just identifying words that have no semantically related word nearby. Obviously, 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 we expect recall to decrease as scope increases, because some relationships will be found even for malapropisms (*i.e.*, there will be more false negatives). But we expect that precision will increase with scope, as it becomes more likely that (genuine) relationships will be found for non-malapropisms (*i.e.*, there will be fewer false positives), and this factor will outweigh the decrease in the overall number of suspects found.

We computed suspicion precision, recall, and F for each of the 5×4 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%. (See Budanitsky and Hirst (in preparation) for details.) 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%. Moreover, 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.

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.[6]

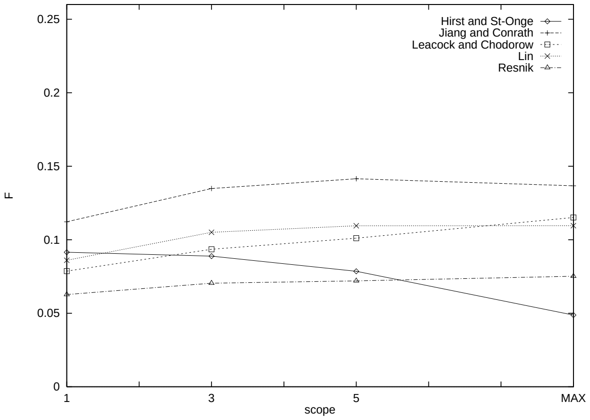

Figure 1: Suspicion F-measure (F_S), by measure and scope.

Scope differences: For Jiang–Conrath and Resnik, the analysis confirms only that the methods perform significantly better with scope 5 than scope 1; for Lin, that scope 3 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 outperforms 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.

5.3.2 Detection

We now turn to the results for malapropism 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

[6] All the comparisons presented, except those with the baseline, were performed with the Bonferroni multiple-comparison technique (Agresti and Finlay, 1997), with an *overall* significance level of .05.

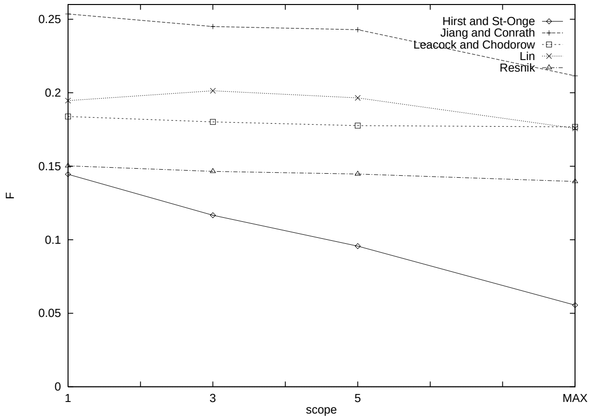

Figure 2: Detection F-measure (F_D), by measure and scope.

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. 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 F_D, though not the highest precision or recall), despite the limitations described in Section 5.1.[7]

Scope differences: Our analysis of scope differences in F (see Figure 2) 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.

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.

5.4 Interpretation of the results

In our interpretation, we focus largely on the results for suspicion; those for detection, though somewhat opaque on their own, both add to the pool of relatedness judgments on which we draw and corroborate what we observe for suspicion.

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 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.[9]

6 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

[7] In conventional interactive spelling correction, it is generally assumed that very high recall is imperative but precision of 25% or even less is acceptable—that is, the user may reject more than 3 out of 4 of the system's suggestions. It must be accepted that very high recall is presently unachievable in real-word spelling correction, but it is unclear just what a typical user would consider to be acceptable performance.

[8] It is debatable whether this metonymic sense of *house* should appear in WordNet at all, though given that it does, its relationship to *crowd* follows, and, as it happens, this sense was the correct one in the context for this particular case; see Section 6 for discussion.

[9] By comparison, for the other measures, the number of suspects drops only to around a third or a quarter from scope $= 1$ to scope = MAX, and the number of articles with no suspect stays at 1 for both Leacock–Chodorow and Resnik and increases only from 1 to 4 for Lin and from 1 to 12 for Jiang–Conrath.

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 WordNet, 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 WordNet 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.

Acknowledgments

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, Stephen Green, Jay Jiang, Keith Knight, Claudia Leacock, Dekang Lin, Radford Neal, Manabu Okumura, Philip Resnik, David St-Onge, and Michael Sussna.

References

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.

Alan Agresti and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences* (third edition). Prentice-Hall.

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. *http://www.cs.toronto.edu/compling/Publications/Abstracts/Theses/Budanitsky-thabs.html*

Alexander Budanitsky and Graeme Hirst. In preparation. Semantic relatedness between lexicalized concepts.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Victoria A. Fromkin. 1980. *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand.* Academic Press.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

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.

Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press.

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.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum 1998, pp. 265–283.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.

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.

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.

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.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10): 627–633.

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.

Mei Wei. 1993. An analysis of word relatedness correlation measures. Master's thesis, University of Western Ontario.