# Chapter 4
# Semantic Distance Measures with Distributional Profiles of Coarse-Grained Concepts

Graeme Hirst and Saif Mohammad

**Abstract.** Although semantic distance measures are applied to words in textual tasks such as building lexical chains, semantic distance is really a property of concepts, not words. After discussing the limitations of measures based solely on lexical resources such as WordNet or solely on distributional data from text corpora, we present a hybrid measure of semantic distance based on distributional profiles of concepts that we infer from corpora. We use only a very coarse-grained inventory of concepts—each category of a published thesaurus is taken as a single concept—and yet we obtain results on basic semantic-distance tasks that are better than those of methods that use only distributional data and are generally as good as those that use fine-grained WordNet-based measures. Because the measure is based on naturally occurring text, it is able to find word pairs that stand in non-classical relationships not found in WordNet. It can be applied cross-lingually, using a thesaurus in one language to measure semantic distance between words in another. In addition, we show the use of the method in determining the degree of antonymy of word pairs.

## 4.1 Semantic Distance

Many applications in natural language processing can be cast in terms of **semantic distance between words** in one way or another. For example, word sense disambiguation can be thought of as finding the sense of the target word that is semantically closest to its context [27]. Real-word spelling errors can be detected

Graeme Hirst
Department of Computer Science, University of Toronto,
Toronto, Ontario, Canada M5S 3G4
e-mail: gh@cs.toronto.edu

Saif Mohammad
Institute for Information Technology, National Research Council Canada,
Ottawa, Ontario, Canada, K1A 0R6
e-mail: saif.mohammad@nrc-cnrc.gc.ca

A. Mehler et al. (Eds.): Modeling, Learning, and Proc. of Text-Tech. Data Struct., SCI 370, pp. 61–79.
springerlink.com 

**Table 4.1** Some NLP applications that have used semantic distance measures [18].

| |
|---|
| Cognate identification |
| Coreference resolution |
| Document clustering |
| Information extraction |
| Information retrieval |
| Multi-word expression identification |
| Paraphrasing and textual entailment |
| Question answering |
| Real-word spelling error detection |
| Relation extraction |
| Semantic similarity of texts |
| Speech recognition |
| Subjectivity determination |
| Summarization |
| Textual inference |
| Word prediction |
| Word sense disambiguation |
| Word-sense discovery |
| Word-sense dominance determination |
| Word translation |

by identifying words that are semantically distant from their context and the existence of a spelling variant that is semantically much closer [8]. Word completion and prediction algorithms may rank those candidate words higher that are semantically close to the preceding context [14]. Table 4.1 lists a number of applications of NLP identified by Mohammad [18] that have been attempted with semantic distance measures.

In particular, semantic distance measures are important in any application that involves finding **lexical chains** in a text – that is, sequences of identical or semantically close words in a text. Lexical chains arise naturally in text that is coherent and cohesive, and thus they can be good indicators of the topic structure of a text.

**Table 4.2** Intuitions of semantic distance.

| **Semantically close** | **Semantically distant** |
|---|---|
| bank–money | doctor–beer |
| apple–fruit | painting–January |
| apple–banana | money–river |
| tree–forest | apple–penguin |
| pen–paper | nurse–bottle |
| hot–cold | pen–river |
| mistake–error | clown–tramway |
| car–wheel | car–algebra |
| dog–bark | faint–porpoise |
| bread–butter | asphalt–chocolate |

Some examples of word pairs that intuitively are semantically close and semantically distant are shown in Table 4.2. People's intuitions of semantic distance are remarkably consistent. In experiments in which subjects are asked to judge the semantic distance of word pairs on a scale of 0 to 4, correlation between subjects is around .9 [29, 17].

We say that two terms are **semantically related** (or **semantically close**) if either there is a lexical semantic relation between them, such as synonymy, hyponymy, meronymy, or troponymy, or a non-classical relation [26], such as role-filler of action, causal relation, co-occurrence, or even just a strong association. We say that two semantically close terms are **semantically similar** if the relation between them is synonymy, hyponymy, or troponymy. For example, the pairs *dog–paw* and *dog–bark* are semantically related but not similar; the relationships are meronymy and typical-action, respectively. The pair *dog–golden retriever* is not only semantically related by hyponymy but is also semantically similar.

The metaphor of semantic distance implies that the measure of relatedness of two words is a continuous function with metric properties yielding a real number in the interval $[0,\infty)$, where 0 means identity and larger values imply larger distances or less relatedness. On the other hand, similarity implies a continuous function yielding a real number in $[0,1]$, where 1 means identity and 0 means maximal dissimilarity. It is necessary, therefore, to keep in mind which view is being taken at any particular time, and map between them as necessary.

Lexical ambiguity is a serious complication for these intuitive ideas of semantic distance. Relations are defined on words yet they depend on senses or concepts; two words may be related with respect to some of their senses but unrelated with respect to others. In this paper, we will take word senses and concepts to be much the same thing – we need not be concerned with the distinctions between them nor with concepts that are unlexicalized (that have no word) – and we will implicitly take a word to be, more precisely, a *lexical unit* composed of a surface string and a sense. However, in many instances, if a word is ambiguous, we know only the surface string and not its particular sense in the instance. In our exposition below, we will use the word *word* sometimes to refer to a complete lexical unit and sometimes to refer to just the surface string; the intent will be clear from the context in each case.

## 4.2 Measures of Semantic Distance

There are many ways that semantic distance can be computed in NLP applications. The first class of methods is **resource-based measures** that use the lexicographers' judgments that are implicit in thesauri, dictionaries, or wordnets. In a thesaurus, for example, the semantic distance between two words can be defined as the length of the path between them through the thesaurus's category structure and/or cross-references and index [25, 10]. In a dictionary or wordnet it can be the number of words that occur in the definitions of both target words and possibly, in the case of a wordnet, their neighbours [1]. In a wordnet, it can be the length of the path from one word-sense (synset) to the other (possibly with scaling factors to account for change

**Table 4.3** Correlations of several resource-based measures of semantic relatedness with data on human judgments from experiments by Miller and Charles [17] (*M&C*) and by Rubenstein and Goodenough [29] (*R&G*). Based on a table by Budanitsky and Hirst [4], with additional data.

| Measure | M&C | R&G |
|---|---|---|
| Hirst and St-Onge [9] | .744 | .786 |
| Jiang and Conrath [11] | .850[a] | .781[a] |
| Leacock and Chodorow [13] | .816 | .838 |
| Lin [15] | .829 | .819 |
| Resnik [28] | .774 | .779 |
| *Roget*-as-tree [10] | .878 | .818 |
| Gloss overlaps [1] | .67 | .60 |
| Latent semantic analysis [2] | .73 | .64 |

[a] Absolute value of correlation coefficient.

in grainedness with depth [9, 13, 31]); or it can be the amount of information shared by both nodes [11, 28, 16]. Most of these methods are reasonably successful in that they correlate well with the human judgments observed in experiments [4]; see Table 4.3. However, they also have serious limitations:

- Each measure is only as good as the resource it depends on. And most wordnet measures use only the noun portion of the wordnet and only the hyponymy relation.
- The measures typically do not work across parts of speech; that is, one can compare nouns only to other nouns, verbs only to other verbs, and so on.
- Non-similarity relationships are not well covered.
- High-quality resources are not available for many languages.
- The role of context is not accounted for.

An alternative to resource-based measures that overcomes these limitations is to use a **distributional measure** as a proxy for 'real' semantics. These methods look only at surface strings of words without regard to their sense. In this class of methods, e.g., [15, 6, 30], we say that two words are semantically related or similar if they tend to co-occur with similar word contexts – that is, if they have similar distributions among other words. The distance between two words is thus defined as the distance between the distributions of the contexts in which they occur. For example, if our target word is *credit* and we see the phrase *a rise in credit and the money supply* in the corpus, we will add 1 to our count of occurrences of *credit* in contexts of *rise*, of *money*, and of *supply*, building a **distributional profile of the word**. Later, we might observe that *debit* tends to occur with many of the same context words, and hence has a distributional profile similar to that of *credit*. Within this idea, there are many definitions of context (e.g., a window of *n* tokens or a syntactic argument relationship), many definitions of "tend to co-occur" (e.g., conditional probability or pointwise mutual information), and many measures of distributional similarity

(e.g., $\alpha$-skew divergence, cosine, Jensen-Shannon divergence, Lin's similarity measure). To define a specific measure, a choice must be made for each of these parameters. See Mohammad and Hirst [19] for a detailed survey of these methods.

These methods overcome some of the limitations of the resource-based approaches. Being corpus-based, they reflect true language usage for which a corpus is available and they are not limited to any particular part of speech or lexical relationship. Moreover, by their very definition they take into account at least a local view of context.

But these methods have limitations too, the most serious of which is that they don't actually work. Their performance is mediocre to awful; Weeds [30] experimented with a number of measures and found their correlation with human data to be between .26 and .62; one of the poorer measures that she experimented with returned this list as the ten words most similar to *hope*: *hem, dissatisfaction, dismay, skepticism, concern, outrage, break, warrior, optimism, readiness.* Moreover:

- The measures are based only on the occurrence of the surface forms of words, not meanings; hence ambiguity is a confound. For example, *credit* has both financial and non-financial senses (*... credited with the invention of the sextant*), but contexts of the different meanings will be conflated in the word's distributional profile. This leads both to attenuation of the measures in the case of true relatedness and to spuriously higher measures between unrelated words.
- They rely on inter-substitutability, which is far too strict a criterion for similarity, let alone relatedness.
- They require enormous corpora to gather sufficient data. Weeds [30] found that the 100M-token British National Corpus was adequate for gathering data for only 2000 word-types. Yet their use in tasks such as real-word spelling correction requires distributional data for a very large vocabulary. This is especially a problem for applications in specific domains and in low-resource languages.

## 4.3 A Hybrid Method for Semantic Distance Measures

We propose a solution to the limitations of these two classes of methods of measuring semantic distance: a hybrid method that uses both distributional information and a lexicographic resource [21, 18]. Our goal is to gain the performance of resource-based methods and the breadth of distributional methods. The central ideas are these:

- In the lexicographical component of the method, concepts are defined by the category structure of a **Roget-style thesaurus**.
- In order to avoid data sparseness, the concepts are very **coarse-grained**.
- The distributional component of the method is based on concepts, not surface strings. We create **distributional profiles of concepts**.

A Roget-style thesaurus classifies all lexical units into approximately 1000 **categories**, with names such as CLOTHING, CLEANNESS, and DESIRE. Each category

is divided into paragraphs that classify lexical units more finely.[1] We take these thesaurus categories as the coarse-grained concepts of our method. That is, for our semantic distance measure, there are only around 1000 concepts (word-senses) in the world; each lexical unit is just a pairing of the surface string with the thesaurus category in which it appears.

In the distributional component of the method, we look at the distribution of these concepts in word contexts. For example, when we see in the corpus *a rise in credit and the money supply*, and given that *credit* appears in category 729 FINANCE in the thesaurus, it's now the count for category 729 that we increment for the context words *rise, money*, and *supply*. To implement this idea, just as for the word-distribution methods, we must choose a definition of context, a measure of strength of association, and a measure of distributional similarity. Given these distributional profiles of concepts, we then define the distance between two concepts as the distance between the distributions of the contexts in which they occur.

But what if a word is ambiguous – appears in more than one thesaurus category? An inability to cope with lexical ambiguity, after all, was one of the limitations of the distributional method that we described earlier. We resolve the ambiguity by bootstrapping as follows. On the initial pass, we count a word for all its categories. This gives a noisy result, but, unlike the word-distribution case and as a consequence of the coarse-grainedness of the concepts, the signal shows through because there are many words in each category. On the second pass, we disambiguate each word by taking the greatest strength of association from the first pass. (We found that additional passes don't increase accuracy.) We define the distance between two lexical units as the distance between their closest senses.

Thus the method is still primarily distributional at heart; its use of lexicographic information is solely for mapping words to the coarse-grained set of concepts. Therefore, we cannot expect it to have the fine performance of measures that are based on rich lexical resources. Nonetheless, the distributional component will give it the breadth that is presently lacking in measures based on those resources.

## 4.4 Evaluation in Monolingual Applications

We carried out several task-oriented monolingual evaluations of our hybrid method. Our corpus was the British National Corpus, our online thesaurus was the *Macquarie Thesaurus* [3], and context was defined to be a $\pm$5-word window, We defined four different versions of the method by choosing four combinations of measures of strength of association and distributional similarity that are frequently used in the literature on the simple word-distance measures described in section 4.2 above:

- Conditional probability ($cp$) with
  - $\alpha$-skew divergence ($ASD_{cp}$);

[1] We do not use other characteristics of Roget-style thesauri, such as the hierarchical structure of the category system, the index, the cross-references, and the further subdivision of paragraphs.

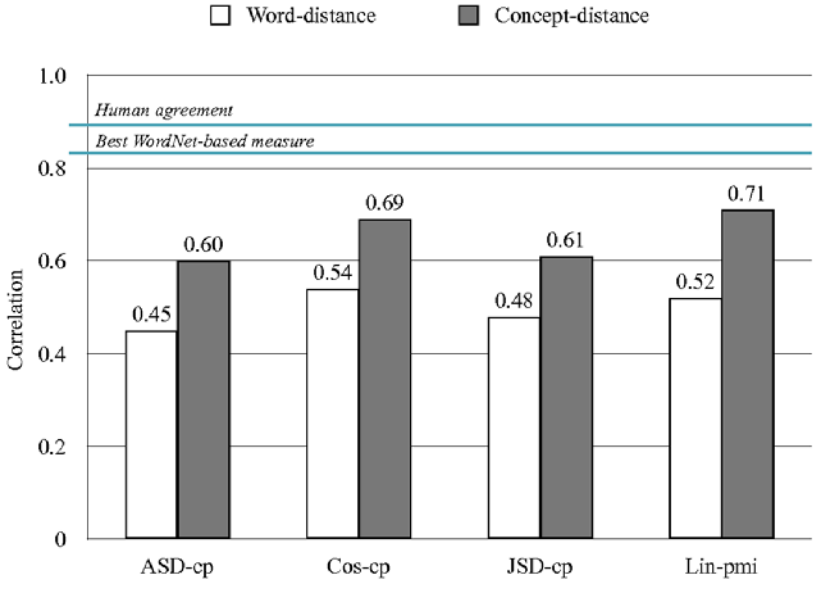


**Fig. 4.1** Performance of four distributional concept-distance measures (grey bars) compared with the corresponding word-distance measures (white bars) on the task of ranking word-pairs by semantic distance (correlation with human judgments).

  - Jensen-Shannon divergence ($JSD_{cp}$);
  - Cosine similarity ($Cos_{cp}$).
- Pointwise mutual information (*pmi*) with Lin's [16] distributional similarity[2] ($Lin_{pmi}$).

We then compared these four distributional concept-distance measures with distributional word-distance measures using the same four choices.

Our first evaluation was simply to compare the measures' ranking of word-pair distances with human norms [21]. The results are shown in Figure 4.1. In each case, using concepts instead of words improved the results markedly. Nonetheless, as we would expect, the performance is not at the level of the best WordNet-based measures (shown in Table 4.3).

Our second evaluation was to use the measure in correcting real-word spelling errors. Hirst and Budanitsky [8] presented a semantic-distance method for finding and correcting real-word spelling errors in a text, and used it to compare six WordNet-based semantic distance measures. We tried our four measures in the method, along with the corresponding four word-distance versions, with the results shown in Figure 4.2 [21]. The *y*-axis shows the **correction ratio** for each method, which is a statistic that takes into account both the number of errors corrected and the number

[2] Lin's distributional similarity measure [16] should not be confounded with his WordNet-based semantic distance measure [15], which was mentioned in section 4.2 above.

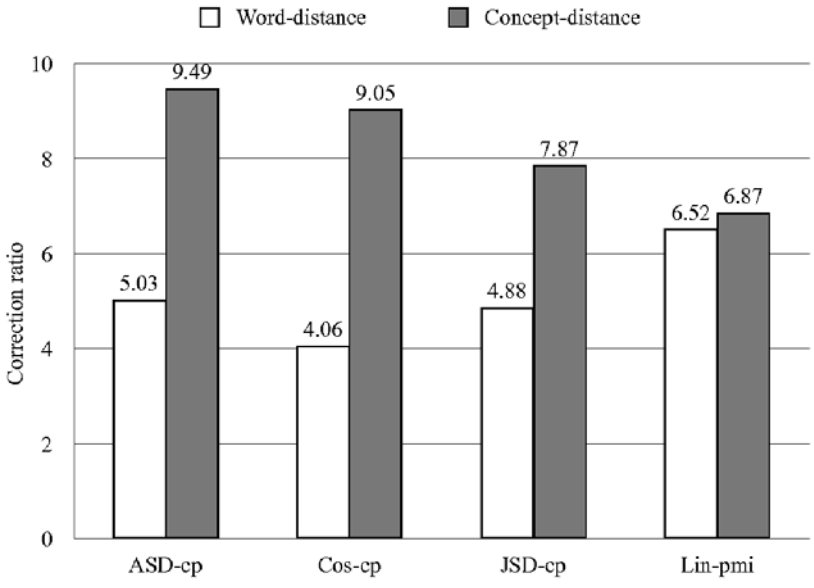


**Fig. 4.2** Performance of four distributional concept-distance measures (grey bars) compared with the corresponding word-distance measures (white bars) on the task of real-word spelling-error correction.

of non-errors flagged as errors (false positives). Again, concept-distance measures give better results than word-distance measures, and except for $Lin_{pmi}$, the difference is quite large. In fact, here the performance of the two best concept-distance measures exceeded that of all but one of the WordNet-based measures as well – though the WordNet-based measure that did better, that of Jiang and Conrath [11], did *much* better, with a score of 12.91 [21]; the second-best WordNet-based measure scored 8.48.

It should be noted that the Rubenstein and Goodenough word-pairs used in the ranking task and the real-word spelling errors in the correction task are all nouns. We expect that the WordNet-based measures will perform less well when other parts of speech are involved, as those hierarchies of WordNet are not as extensively developed. Further, the various hierarchies are not well connected, nor is it clear how to use these interconnections across parts of speech for calculating semantic distance. On the other hand, our hybrid measures do not rely on any hierarchies (even if they exist in the thesaurus) but on sets of words that unambiguously represent each sense. And because our measures are tied closely to the corpus from which co-occurrence counts are made, we expect the use of domain-specific corpora to give even better results.

Our other two monolingual evaluations involved word senses. In the task of determining which sense of a word is dominant in a text, we achieved near upper bound results [20]. And using the measures in word sense disambiguation with an unsupervised naive Bayes classifier, we achieved respectable results in SemEval 2007 [23].

## 4.5 Extension to Cross-Lingual Applications

### *4.5.1 Method*

It is not necessary in our method that the corpus of text used to determine the distributional profiles of concepts be in the same language as the thesaurus used to define the concepts. In particular, the thesaurus may be in English ($E$) while the corpus is in a lower-resource language $L$ that has no Roget-style thesaurus. All that is necessary to make this work is a **bilingual dictionary** from $L$ to $E$ that can map the words of the corpus from $L$ to their thesaurus concepts in $E$. Of course, there will be ambiguity in the translation that creates spurious candidate senses, but this is background noise, as before, that can be eliminated by bootstrapping [22].

Figure 4.3 illustrates the method with two examples in which German plays the role of the low-resource language (see section 4.5.2 below). The first example, *Stern*, is mapped by the bilingual dictionary to *star*, which has additional senses in English; the second example, *Bank*, is ambiguous in German and is mapped to two different English words, *bank* and *bench*, in its different senses (Figure 4.3*(a)*). Concepts, that is thesaurus categories, are obtained for each of the English words (Figure 4.3*(b)*), at which point the English words themselves can be ignored (Figure 4.3*(c)*); observe that some of the concepts are spurious, relative to the original German words, being artifacts of the intermediate English (Figure 4.3*(d)*). However, on the next iteration in the bootstrapping process, these spurious concepts can be identified and removed (Figure 4.3*(e)*) because of their relatively low strength of association with the original German words.

### *4.5.2 Evaluation*

We evaluated the method on two tasks, with German playing the role of the low-resource language $L$. Of course, German is not really a low-resource language, but the logic of the evaluation requires that the test language $L$ actually have sufficient resources that our method can be compared with resource-based monolingual methods in $L$. The two tasks were ranking German word pairs for relatedness and solving "Word Power" problems (which require finding the word semantically closest to the target word from a choice of four alternatives) from the German edition of *Reader's Digest*. Our aim was not to perform better than the monolingual method but merely to obtain results that are not markedly poorer; after all, the cross-lingual method is inherently noisy, and is intended for situations only when the resources for monolingual methods are not available at all.

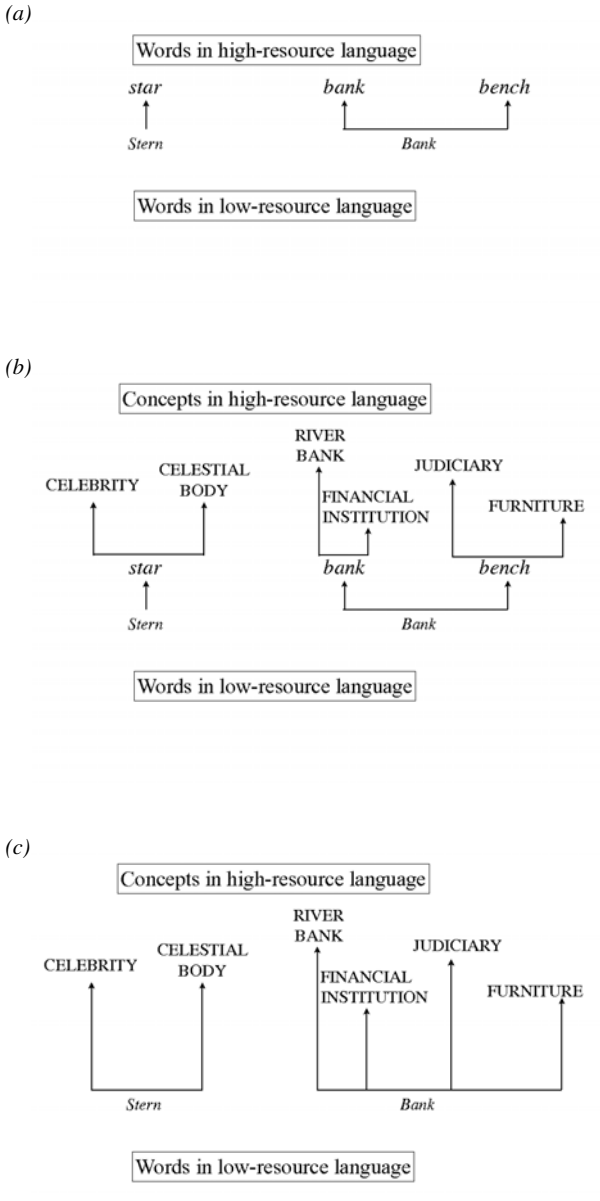


**Fig. 4.3** Cross-lingual examples (German to English) demonstrating how bootstrapping removes the artifacts of lexical ambiguity. *(a)* The bilingual dictionary maps the words from German to English. *(b)* The English words are then mapped to thesaurus concepts. *(c)* The English words can now be ignored. [*Figure continues on next page.*]

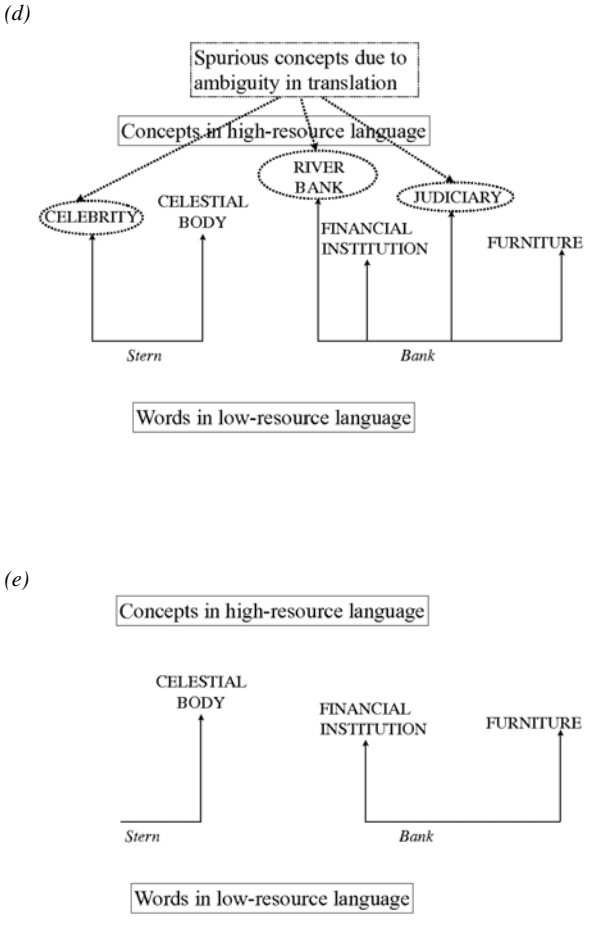


**Fig. 4.3 (cont.)** *(d)* Some of the concepts are spurious artifacts of the intermediate English. *(e)* In the bootstrapping process, these spurious concepts can be identified and removed.

As resources for the cross-lingual measure, we used the German newspaper corpus *taz* and the German–English bilingual lexicon BEOLINGUS. As before, the English thesaurus was the *Macquarie Thesaurus*. We tried the same four versions of the method that we used in the evaluations of section 4.4 above. Our benchmark for comparison as a monolingually based semantic distance measure in the same tasks was WordNet-style measures (see section 4.2 above) with GermaNet as the resource; in addition to the measures of Jiang and Conrath, Lin, and Resnik (see Table 4.3) we also used two pseudo-gloss-based measures proposed explicitly for GermaNet by Gurevych [7].

We found the cross-lingual method to be not just the equal of the GermaNet-based monolingual methods but better in both tests. Figure 4.4 illustrates the results. The upper histogram shows Spearman rank correlations with human rankings of the best of our cross-lingual measures (which was $Lin_{pmi}$) and the best of the GermaNet measures (which was Jiang and Conrath's); the former achieves a notably better result. The lower series of histograms shows results on the "Word Power" problems for the best methods of each type; for the GermaNet methods, this was one of Gurevych's, and for the cross-lingual method this was $JSD_{cp}$ and $Lin_{pmi}$ equally, with $Cos_{cp}$ only a tiny amount behind. Although the cross-lingual measures have a lower precision than the best monolingual measure, they have higher recall and overall a slightly better $F$-score. The higher recall implies that the bilingual dictionary had a better coverage of the vocabulary of the "Word Power" problems than GermaNet did.

In addition to these tests, we tried the cross-lingual method out in a Chinese–English setting in the SemEval 2007 task of choosing the best English translation for an ambiguous Chinese word in context, and we achieved good results with an unsupervised naive Bayes classifier [23].

## 4.6 Antonymy and Word Opposition

In this section, we show that our method for semantic distance can be extended to solve the related problem of finding words that are antonyms or, more generally, pairs of words whose meanings are contrasting or opposed to one another [24]. Thus we want to go beyond the conventional kinds of antonymy (*wet–dry, open–closed, life–death*), which are already well-recorded in lexical resources such as WordNet, to a more-general notion of contrast in meaning (*closed–accessible, flinch–advance, cogent–unconvincing*) which is largely unrecorded. This has application in tasks such as detecting contradictions and differences in opinion, and detecting paraphrases in which one alternative is negated (*caught–not evaded*).

We base our approach on two hypotheses:

- **The co-occurrence hypothesis** (Charles and Miller [5]): Antonyms co-occur more often than chance.
- **The distributional hypothesis** (after Justeson and Katz [12]): Antonyms tend to occur in similar contexts.

By comparing 1000 randomly chosen antonym pairs from WordNet with a control set of 1000 randomly chosen (non-antonymous) word pairs, we showed [24] that both of these hypotheses are correct: antonym pairs have a higher strength of co-occurrence (by pointwise mutual information) than random word pairs ($p < .01$) and are distributionally more similar (by Lin's measure [16]) than random pairs ($p < .01$). The same is true, of course, of semantically similar and semantically related words. So these two hypotheses alone are not sufficient to identify contrasting word pairs.

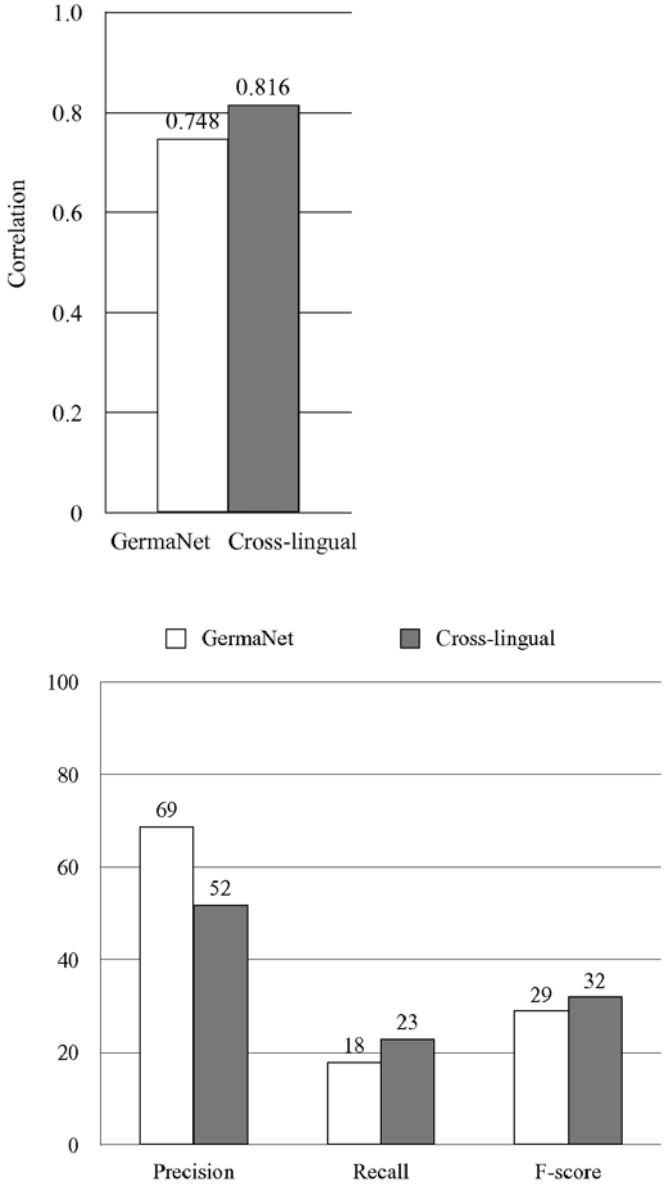


**Fig. 4.4** Performance of the cross-lingual method (grey bars) compared with monolingual GermaNet-based method (white bars) on ranking word-pairs by distance (correlation with human judgments) *(top)* and on *Reader's Digest* "Word Power" problems (precision, recall, and $F$-measure) *(bottom)*.

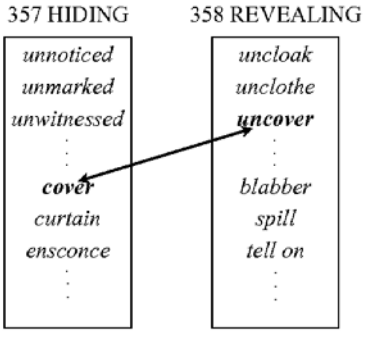


**Fig. 4.5** Two thesaurus categories are assumed to be contrasting if each contains one member of a pair from the seed set of antonyms.

The central ideas of our method are these: First, we identify **contrasting category pairs** in the thesaurus using the structure of the thesaurus and a set of seed antonym pairs. We then determine the *degree of antonymy* between a pair of words, one from each of a pair of contrasting categories, using the two hypotheses mentioned above and, again, the structure of the thesaurus.

### *4.6.1 Contrasting Categories*

We have two heuristics for recognizing contrasting categories. First, thesaurus lexicographers often explicitly place contrasting categories adjacent to each other; for example, the LOVE category may follow the HATE category. So we assume that all adjacent category pairs are contrasting. This is obviously untrue in general; for example, the other category adjacent to HATE may be INDIFFERENCE. Second, we manually create a list of 16 affixes that tend to generate antonyms, such as X–*anti*X (*clockwise–anticlockwise*), X*less*–X*ful* (*harmless–harmful*), and *im*X–*ex*X (*implicit–explicit*) and we use this list to generate a seed set of about 2600 pairs of likely antonyms.[3] We then assume that a pair of thesaurus categories containing a word pair in the seed set is contrasting (see Figure 4.5); this is our second heuristic. Additionally, we also use antonym pairs from WordNet to find contrasting categories where possible; WordNet contains 10,800 antonym pairs for which both words were in our thesaurus.

[3] The affix list obviously overgenerates (*part–depart; tone–intone; sect–insect; coy–decoy*), but this has little effect on the results.

| 'hardened in feelings' | 'resistant to persuasion' | 'persistent' |
|---|---|---|
| **obdurate:** | **obdurate:** | **obdurate:** |
| a. *meager* | a. *yielding** | a. *commensurate* |
| b. *unsusceptible* | b. *motivated* | b. *transitory** |
| c. *right* | c. *moribund* | c. *complaisant* |
| d. *tender** | d. *azure* | d. *similar* |
| e. *intelligent* | e. *hard* | e. *uncommunicative* |

**Fig. 4.6** GRE-style multiple-choice closest-opposite questions using the same prompt in different senses. The correct answer is marked with an asterisk.

### *4.6.2 Degree of Antonymy*

We can now determine the **degree of antonymy** between two thesaurus categories, and from that between two lexical units (a word and its thesaurus category), and from that between two words:

- **Categories:** Following the distributional hypothesis for antonyms, we stipulate that the degree of antonymy between two *contrasting* categories is proportional to the semantic closeness of the two categories as measured by our hybrid semantic-distance measure (section 4.3 above).
- **Lexical units:** We assign four discrete levels of antonymy. If the units do not occur in contrasting categories, then they have ZERO antonymy. Otherwise, if each occurs in its respective category in the same paragraph as one of the seeds that is the basis for the contrast between the categories, then antonymy is HIGH. Otherwise, following the co-occurrence hypothesis, the antonymy is MEDIUM or LOW depending on the strength of co-occurrence between the categories.
- **Words:** We take the degree of antonymy of two words to be that of their most antonymous pair of senses.

### *4.6.3 Evaluation*

We evaluated the method on 950 GRE-style multiple-choice closest-opposite questions. Each question contains a prompt word and five alternatives from which the closest opposite to the prompt must be chosen. Typically the alternatives will include as distractors both another close opposite and a near-synonym of the prompt. An ambiguous word may appear in more than one question in different senses; Figure 4.6 shows three questions all using the prompt *obdurate* in different senses. (Of course, the system is not informed of the intended sense.)

The results are shown in Figure 4.7. The baselines for our evaluation are simple random choice from the five alternatives, and looking for the answer in WordNet but choosing at random if none of the alternatives are listed as an antonym of the prompt. In fact, the answer is so rarely found in WordNet that it scarcely improves on random choice.

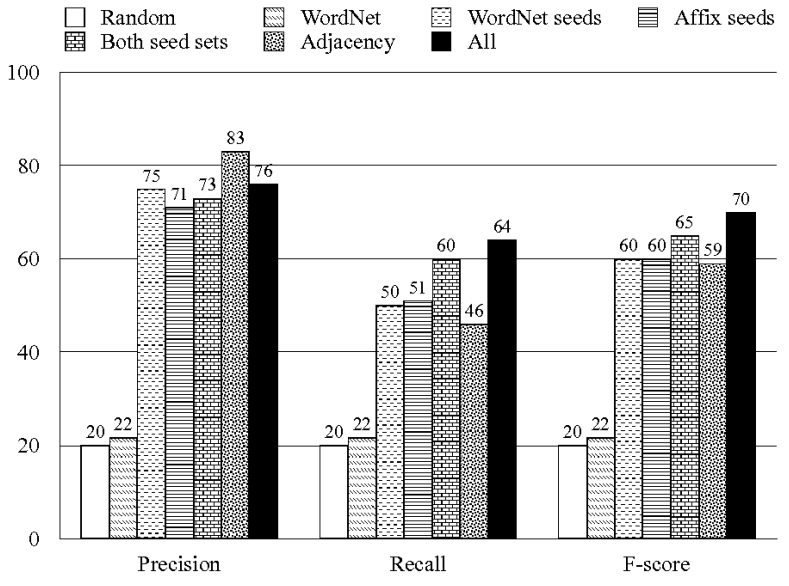


**Fig. 4.7** Results of evaluation of method for determining degree of antonymy. In each group, the bars show, from left to right: a random-choice baseline; random choice except using WordNet antonyms where possible; the method using only WordNet-generated seed-pairs; the method using only affix-generated seed-pairs; the method using both seed sets; the method using only the category-adjacency heuristic; and the method using all heuristics.

We also tried the heuristics individually as well as in combination. The relatively small set of affix-generated seed-pairs performed almost as well by itself as the larger set of WordNet-generated seed-pairs; but the two together performed better than either alone. The simple adjacency heuristic achieved better precision than this combination; however, its recall was much lower. The highest $F$-score was achieved by a combination of all three heuristics.

## 4.7 Conclusion

There have been many prior proposals for measuring semantic distance: measures based on lexicographical resources and measures based on word distributions in word contexts. Both kinds have significant limitations. By proposing a hybrid measure based on distributions of coarse-grained concepts (thesaurus categories) in word contexts, we avoid the limitations of purely corpus-based and WordNet-based measures. Its performance is competitive with WordNet-based measures (and better than corpus-based measures), it operates across parts of speech, and it

offers the possibility of cross-lingual use for resource-poor languages. In addition we have shown how it can be used in a method for determining the degree of antonymy between words.

**Acknowledgements.** This paper is based on work that was first reported in the following papers: [21], [22], [23], and [24]. Parts of the work were carried out in collaboration with Bonnie Dorr and Philip Resnik, University of Maryland, and Iryna Gurevych and Torsten Zesch, Technische Universität Darmstadt. The research was supported by the Natural Sciences and Engineering Research Council of Canada (Hirst, Mohammad), the University of Toronto (Mohammad), the U.S. National Science Foundation (Mohammad, Dorr), the Human Language Technology Center of Excellence (Dorr), die Deutsche Forschungsgemeinschaft (Gurevych, Zesch), and the U.S. Office of Naval Research and the U.S. Department of Defense (Resnik). We are grateful to Alex Budanitsky, Diana McCarthy, Michael Demko, Suzanne Stevenson, Frank Rudzicz, Afsaneh Fazly, Afra Alishahi, Siddharth Patwardhan, Xinglong Wang, and Vivian Tsang for helpful discussions.

## References

[1] Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 805–810 (2003)

[2] Beigman Klebanov, B.: Semantic relatedness: Computational investigation of human data. In: Proceedings of the 3rd Midwest Computational Linguistics Colloquium, Urbana-Champaign, USA (2006)

[3] Bernard, J. (ed.): The Macquarie Thesaurus. Macquarie Library, Sydney, Australia (1986)

[4] Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance. Computational Linguistics 32(1), 13–47 (2006)

[5] Charles, W.G., Miller, G.A.: Contexts of antonymous adjectives. Applied Psychology 10, 357–375 (1989)

[6] Dagan, I.: Contextual word similarity. In: Dale, R., Moisl, H., Somers, H. (eds.) Handbook of Natural Language Processing, pp. 459–475. Marcel Dekker Inc., New York (2000)

[7] Gurevych, I.: Using the structure of a conceptual network in computing semantic relatedness. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing, Jeju Island, Republic of Korea, pp. 767–778 (2005)

[8] Hirst, G., Budanitsky, A.: Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering 11, 87–111 (2005)

[9] Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, ch. 13, pp. 305–332. The MIT Press, Cambridge (1998)

[10] Jarmasz, M., Szpakowicz, S.: Roget's Thesaurus and semantic similarity. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003), pp. 212–219 (2003)

[11] Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics (ROCLING X), Taiwan, pp. 19–33 (1997)

[12] Justeson, J.S., Katz, S.M.: Cooccurrences of antonymous adjectives and their contexts. Computational Linguistics 17, 1–19 (1991)
[13] Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, ch. 11, pp. 265–283. The MIT Press, Cambridge (1998)
[14] Li, J., Hirst, G.: Semantic knowledge in a word completion task. In: Proceedings, 7th International ACM SIGACCESS Conference on Computers and Accessibility, Baltimore, MD (2005)
[15] Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 36th annual meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING- ACL 1998), pp. 768–774 (1998)
[16] Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, pp. 296–304 (1998)
[17] Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
[18] Mohammad, S.: Measuring semantic distance using distributional profiles of concepts. PhD thesis, Department of Computer Science, University of Toronto (2008)
[19] Mohammad, S., Hirst, G.: Distributional measures as proxies for semantic relatedness (2005), `http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf`
[20] Mohammad, S., Hirst, G.: Determining word sense dominance using a thesaurus. In: Proceedings of the 11th conference of the European chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, pp. 121–128 (2006)
[21] Mohammad, S., Hirst, G.: Distributional measures of concept-distance: A task-oriented evaluation. In: Proceedings, 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Sydney, Australia (2006)
[22] Mohammad, S., Gurevych, I., Hirst, G., Zesch, T.: Cross-lingual distributional profiles of concepts for measuring semantic distance. In: 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), Prague (2007)
[23] Mohammad, S., Hirst, G., Resnik, P.: TOR, TORMD: Distributional profiles of concepts for unsupervised word sense disambiguation. In: SemEval-2007: 4th International Workshop on Semantic Evaluations, Prague (2007)
[24] Mohammad, S., Dorr, B., Hirst, G.: Computing word-pair antonymy. In: 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Waikiki, Hawaii (2008)
[25] Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17(1), 21–48 (1991)
[26] Morris, J., Hirst, G.: Non-classical lexical semantic relations. In: Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Boston, MA (2004): reprinted in: Hanks, P.(editor), Lexicology: Critical Concepts in Linguistics, Routledge (2007)
[27] Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pp. 241–257 (2003)

[28] Resnik, P.: Using information content to evaluate semantic similarity. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, pp. 448–453 (1995)
[29] Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Communications of the ACM 8(10), 627–633 (1965)
[30] Weeds, J.E.: Measures and applications of lexical distributional similarity. PhD thesis, University of Sussex (2003)
[31] Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp. 133–138 (1994)