Evaluating WordNet-based Measures of Lexical Semantic Relatedness

Alexander Budanitsky*
University of Toronto

Graeme Hirst*
University of Toronto

The quantification of lexical semantic relatedness has many applications in NLP, and many different measures have been proposed. We evaluate five of these measures, all of which use WordNet as their central resource, by comparing their performance in detecting and correcting real-word spelling errors. An information-content–based measure proposed by Jiang and Conrath is found superior to those proposed by Hirst and St-Onge, Leacock and Chodorow, Lin, and Resnik. In addition, we explain why distributional similarity is not an adequate proxy for lexical semantic relatedness.

1 Introduction

The need to determine *semantic relatedness* or its inverse, *semantic distance*, between two lexically expressed concepts is a problem that pervades much of natural 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's important to note that semantic relatedness is a more general concept than *similarity*; similar entities are semantically related by virtue of their similarity (*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, rain–flood*). 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*.

* Department of Computer Science, Toronto, Ontario, Canada M5S 3G4; {abm, gh}@cs.toronto.edu

However, it is frequently unclear how to assess the relative merits of the many competing approaches that have been proposed for determining lexical semantic relatedness. Given a measure of relatedness, how can we tell whether it is a good one or a poor one? Given two measures, how can we tell whether one is better than the other, and under what conditions it is better? And what is it that makes some measures better than others? Our purpose in 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.

1.1 Terminology and notation

In the literature related to this topic, at least three different terms are used by different authors or sometimes interchangeably by the same authors: *semantic relatedness*, *similarity*, and *semantic distance*.

Resnik (1995) attempts to demonstrate the distinction between the first two by way of example. "*Cars* and *gasoline*", he writes, "would seem to be more closely related than, say, *cars* and *bicycles*, but the latter pair are certainly more similar." *Similarity* is thus a special case of *semantic relatedness*, and we adopt this perspective in this paper. Among other relationships that the notion of *relatedness* encompasses are the various kinds of meronymy, antonymy, functional association, and other "non-classical relations" (Morris and Hirst, 2004).

The term *semantic distance* may cause even more confusion, as it can be used when talking about either just similarity *or* relatedness in general. Two concepts are "close" to one another if their similarity *or* their relatedness is high, and otherwise they are "distant". Most of the time, these two uses are consistent with one another, but not always; antonymous concepts are dissimilar and hence distant in one sense, and yet are strongly related semantically and hence close in the other sense. We would thus have very much preferred to be able to adhere to the view of semantic distance as the inverse of semantic *relatedness*, not merely of *similarity*, in the present paper. Unfortunately, because of the sheer number of methods measuring *similarity*, as well as those measuring distance as the "opposite" of *similarity*, this would have made for an awkward presentation. Therefore, we have to ask the reader to rely on context when interpreting what exactly the expressions *semantic distance, semantically distant,* and *semantically close* mean in each particular case.

Various approaches presented below speak of *concepts* and *words*. As a means of acknowledging the polysemy of language, in this paper the term *concept* will refer to a particular sense of a given *word*. We want to be very clear that, throughout this paper, when we say that two words are "similar", this is a short way of saying that they denote similar concepts; we are *not* talking about similarity of distributional or co-occurrence behavior of the words, for which the term *word similarity* has also been used (Dagan, 2000; Dagan et al., 1999). While similarity of denotation might be inferred from similarity of distributional or co-occurrence behavior (Dagan, 2000; Weeds, 2003), the two are distinct ideas. We return to the relationship between them in Section 6.2.

When we refer to hierarchies and networks of concepts, we will use both the terms *link* and *edge* to refer to the relationships between nodes; we prefer the former term when our view emphasizes the taxonomic aspect or the meaning of the network, and the latter when our view emphasizes algorithmic or graph-theoretic aspects. In running text, examples of concepts are typeset in sans-serif font, whereas examples of words are given in italics; in formulas, concepts and words will usually be denoted by c and w, with various subscripts. For the sake of uniformity of presentation, we have taken the liberty of altering the original notation accordingly in some other authors' formulas.

2 Lexical resource–based approaches to measuring semantic relatedness

All approaches to measuring semantic relatedness that use a lexical resource construe the resource, in one way or another, as a network or directed graph, and then base the measure of relatedness on properties of paths in this graph.

2.1 Dictionary-based approaches

Kozima and Furugori (1993) turned the *Longman Dictionary of Contemporary English* (LDOCE) (Procter, 1978) into a network by creating a node for every headword and linking each node to the nodes for all the words used in its definition. The 2851-word controlled defining vocabulary of LDOCE thus becomes the densest part of the network: the remaining nodes, which represent the headwords outside of the defining vocabulary, can be pictured as being situated at the fringe of the network, as they are linked only to defining-vocabulary nodes and not to each other. In this network, the similarity function $\mathrm{sim}_{\mathrm{KF}}$ between words of the defining vocabulary is computed by means of spreading activation on this network. The function is extended to the rest of LDOCE by representing each word as a list $W = \{w_1, \ldots, w_r\}$ of the words in its definition; thus, for instance,

$$\begin{aligned}
&\mathrm{sim}_{\mathrm{KF}}(\textit{linguistics}, \textit{stylistics}) \\
&\quad = \mathrm{sim}_{\mathrm{KF}}(\{\textit{the, study, of, language, in, general, and, of, particular,} \\
&\qquad\quad \textit{languages, and, their, structure, and, grammar, and, history}\}, \\
&\qquad\quad \{\textit{the, study, of, style, in, written, or, spoken, language}\})\ .
\end{aligned}$$

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.).

2.2 Approaches based on Roget-structured thesauri

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. However, while the classes and categories are named, the finer divisions are not; the words are clustered without attempting to explicitly indicate how and why they are related. The user's main access is through the *index*, which contains category numbers along with *labels* representative of those categories for each word. Polysemes are implicitly disambiguated, to a certain extent, by the other words in their cluster and in their index entry. Closely related concepts might or might not be physically close in the thesaurus: "Physical closeness has some importance ... but words in the index of the thesaurus often have widely scattered categories, and each category often points to a widely scattered selection of categories" (Morris and Hirst, 1991). Methods of semantic distance that are based on Roget-structured thesauri therefore rely not only on the category structure but also on the index and on the *pointers* within categories that cross-reference other categories. In part as a consequence of this, typically no numerical value for semantic distance can be obtained: rather, algorithms using the thesaurus compute a distance implicitly and return a boolean value of 'close' or 'not close'.

Working with an abridged version of *Roget's Thesaurus*, Morris and Hirst (1991) identified five types of semantic relations between words. In their approach, two words were deemed to be related to one another, or semantically close, if their base forms satisfy any one of the following conditions:

1. they have a category in common in their index entries;

2. one has a category in its index entry that contains a pointer to a category of the other;
3. one is either a label in the other's index entry or is in a category of the other;
4. they are both contained in the same subcategory;
5. they both have categories in their index entries that point to a common category.

These relations account for such pairings as *wife* and *married*, *car* and *driving*, *blind* and *see*, *reality* and *theoretically*, *brutal* and *terrified*. (However, different editions of *Roget's Thesaurus* yield somewhat different sets of relations.) Of the five types of relations, perhaps the most intuitively plausible ones — the first two in the list above — were found to validate over 90% of the intuitive lexical relationships that the authors used as a benchmark in their experiments.

Jarmasz and Szpakowicz (2003) also implemented a similarity measure with *Roget's Thesaurus*; but because this measure is based strictly on hierachy rather than the index structure, we discuss it in Section 2.4 below.

2.3 Approaches using WordNet and other semantic networks

Most of the methods discussed in the remainder of Section 2 use WordNet (Fellbaum, 1998), a broad coverage lexical network of English words. 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. (A polysemous word will appear in one synset for each of its senses.) In the first versions of WordNet (those numbered 1.*x*), the networks for the four different parts of speech were not linked to one another.[1] The noun network of WordNet was the first to be richly developed, and most of the researchers whose work we will discuss below therefore limited themselves to this network.[2]

The backbone of the noun network is the subsumption hierarchy (*hyponymy/hypernymy*), which accounts for close to 80% of the relations. At the top of the hierarchy are 11 abstract concepts, termed *unique beginners*, such as entity ('something having concrete existence; living or nonliving') and psychological feature ('a feature of the mental life of a living organism'). The *maximum depth* of the noun hierarchy is 16 nodes. The nine types of relations defined on the noun subnetwork, in addition to the synonymy relation that is implicit in each node are: hyponymy (IS-A) relation, and its inverse, hypernymy; six meronymic (PART-OF) relations — COMPONENT-OF, MEMBER-OF and SUBSTANCE-OF and their inverses; and antonymy, the COMPLEMENT-OF relation.

In discussing WordNet, we use the following definitions and notation:

- The *length* of the shortest path in WordNet from synset c_i to synset c_j (measured in edges or nodes) is denoted by $\mathrm{len}(c_i, c_j)$. We stipulate a global root *root* above the 11 unique beginners to ensure the existence of a path between any two nodes.

- The *depth* of a node is the length of the path to it from the global root, *i.e.*, $depth(c_i) = \mathrm{len}(root, c_i)$.

- We write $lso(c_1, c_2)$ for the *lowest super-ordinate* (or *most specific common subsumer*) of c_1 and c_2.

[1] We began this work with WordNet 1.5, and stayed with this version despite newer releases in order to maintain strict comparability. Our experiments were complete before WordNet 2.0 was released.

[2] It seems to have been tacitly assumed by these researchers that results would generalize to the network hierarchies of other parts of speech. Nonetheless, Resnik and Diab (2000) caution that the properties of verbs and nouns might be different enough that they should be treated as separate problems, and recent research by Banerjee and Pedersen (2003) supports this assumption: they found that in word-sense disambiguation task, their gloss-overlap measure of semantic relatedness (see section 6.1 below) performed far worse on verbs (and slightly worse on adjectives) than it did on nouns.

- Given any formula $\mathrm{rel}(c_1, c_2)$ for semantic relatedness between two concepts c_1 and c_2, the relatedness $\mathrm{rel}(w_1, w_2)$ between two words w_1 and w_2 can be calculated as

$$\mathrm{rel}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\mathrm{rel}(c_1, c_2)] \,, \tag{1}$$

where $s(w_i)$ is "the set of concepts in the taxonomy that are senses of word w_i" (Resnik, 1995). That is, the relatedness of two words is equal to that of the most-related pair of concepts that they denote.

2.4 Computing taxonomic path length

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: "The shorter the path from one node to another, the more similar they are" (Resnik, 1995). This approach was taken, for example, by Rada and colleagues (Rada et al., 1989; Rada and Bicknell, 1989), not on WordNet but on MeSH (Medical Subject Headings), a semantic hierarchy of terms used for indexing articles in the bibliographic retrieval system Medline. The network's 15,000 terms form a nine-level hierarchy based on the BROADER-THAN relationship. The principal assumption of Rada and colleagues was that "the number of edges between terms in the MeSH hierarchy is a measure of conceptual distance between terms". Despite the simplicity of this distance function, the authors were able to obtain surprisingly good results in their information retrieval task. In part, their success can be explained by the observation of Lee et al. (1993) that while "in the context of … semantic networks, shortest path lengths between two concepts are not sufficient to represent conceptual distance between those concepts … when the paths are restricted to IS-A links, the shortest path length does measure conceptual distance." Another component of their success is certainly the specificity of the domain, which ensures relative homogeneity of the hierarchy. Notwithstanding these qualifications, Jarmasz and Szpakowicz (2003) also achieved good results with *Roget's Thesaurus* by ignoring the index and treating the thesaurus as a simple hierarchy of clusters. They computed semantic similarity between two words as the length of the shortest path between them. The words were not explicitly disambiguated.

Hirst and St-Onge (1998; St-Onge 1995) adapted Morris and Hirst's (1991) semantic distance algorithm from *Roget's Thesaurus* to WordNet.[3] 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*).

Two words are said to be in a **medium-strong**, or **regular**, relation if there exists an *allowable path* connecting a synset associated with each word (for example, *carrot* and *apple*). A path is *allowable* if it contains no more than five links and conforms to one of eight patterns, the intuition behind which is that "the longer the path and the more changes of direction, the lower the weight". The details of the patterns are outside of the scope of this paper; all we need to know for the purposes of subsequent discussion is that an allowable path may include more than one link and that the directions of links on the same path may vary (among *upward* (*hypernymy* and *meronymy*), *downward*

[3] The original ideas and definitions of Hirst and St-Onge (1998) (including those for the direction of links — see below) were intended to apply to all parts of speech and the entire range of relations featured in the WordNet ontology (which include *cause, pertinence, also see*, etc.). Like other researchers, however, they had to resort to the noun subnetwork only. In what follows, therefore, we will use appropriately restricted versions of their notions.

(*hyponymy* and *holonymy*) and *horizontal* (*antonymy*)). Hirst and St-Onge's approach may thus be summarized by the following formula for two WordNet concepts $c_1 \neq c_2$:

$$\text{rel}_{\text{HS}}(c_1, c_2) = C - \text{len}(c_1, c_2) - k \times \text{turns}(c_1, c_2) \tag{2}$$

where C and k are constants (in practice, they used $C = 8$ and $k = 1$), and $\text{turns}(c_1, c_2)$ is the number of times the path between c_1 and c_2 changes direction.

2.5 Scaling the network

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). For instance, in WordNet, the link rabbit ears IS-A television antenna covers an intuitively narrow distance, whereas white elephant IS-A possession covers an intuitively wide one. The approaches discussed below are attempts undertaken by various researchers to overcome this problem.

2.5.1 Sussna's depth-relative scaling Sussna's (1993; 1997) approach to scaling is based on his observation that sibling-concepts deep in a taxonomy appear to be more closely related to one another than those higher up. His method construes each edge in the WordNet noun network as consisting of two directed edges representing inverse relations. Each relation r has a weight or a range $[\text{min}_r; \text{max}_r]$ of weights associated with it: for example, *hypernymy, hyponymy, holonymy,* and *meronymy* have weights between $\text{min}_r = 1$ and $\text{max}_r = 2$.[4] The weight of each edge of type r from some node c_1 is reduced

[4] Sussna's experiments proved the precise details of the weighting scheme to be material only in fine-tuning the performance.

by a factor that depends on the number of edges, edges_r, of the same type leaving c_1:

$$\text{wt}(c_1 \rightarrow_r) = \max_r - \frac{\max_r - \min_r}{\text{edges}_r(c_1)} \; . \tag{3}$$

The distance between two adjacent nodes c_1 and c_2 is then the average of the weights on each direction of the edge, scaled by the depth of the nodes:

$$\text{dist}_S(c_1, c_2) = \frac{\text{wt}(c_1 \rightarrow_r) + \text{wt}(c_2 \rightarrow_{r'})}{2 \times \max\{\text{depth}(c_1), \text{depth}(c_2)\}} \; , \tag{4}$$

where r is the relation that holds between c_1 and c_2 and r' is its inverse (*i.e.,* the relation that holds between c_2 and c_1). Finally, the semantic distance between two arbitrary nodes c_i and c_j is the sum of the distances between the pairs of adjacent nodes along the shortest path connecting them.

2.5.2 Wu and Palmer's Conceptual Similarity In a paper on translating English verbs into Mandarin Chinese, Wu and Palmer (1994) introduce a scaled metric for what they call *conceptual similarity* between a pair of concepts c_1 and c_2 in a hierarchy as

$$\text{sim}_{\text{WP}}(c_1, c_2) = \frac{2 \times \text{depth}(lso(c_1, c_2))}{\text{len}(c_1, lso(c_1, c_2)) + \text{len}(c_2, lso(c_1, c_2)) + 2 \times \text{depth}(lso(c_1, c_2))} \tag{5}$$

Note that $\text{depth}(lso(c_1, c_2))$ is the 'global' depth in the hierarchy; its role as a scaling factor can be seen more clearly, if we recast Equation 5 from *similarity* into *distance*:

$$\text{dist}_{\text{WP}}(c_1, c_2) = 1 - \text{sim}_{\text{WP}}(c_1, c_2) = \frac{\text{len}(c_1, lso(c_1, c_2)) + \text{len}(c_2, lso(c_1, c_2))}{\text{len}(c_1, lso(c_1, c_2)) + \text{len}(c_2, lso(c_1, c_2)) + 2 \times \text{depth}(lso(c_1, c_2))} \tag{6}$$

2.5.3 Leacock and Chodorow's Normalized Path Length Leacock and Chodorow (1998) proposed the following formula for computing the scaled semantic similarity between concepts c_1 and c_2 in WordNet:

$$\text{sim}_{\text{LC}}(c_1, c_2) = -\log \frac{\text{len}(c_1, c_2)}{2 \times \max\limits_{c \in WordNet} \text{depth}(c)} \; . \tag{7}$$

Here, the denominator includes the maximum depth of the hierarchy.

2.6 Information-based and integrated approaches

Like the methods in the preceding subsection, the final group of approaches that we present attempt to counter problems inherent in a general ontology by incorporating an additional, and qualitatively different, knowledge source, namely information from a corpus.

2.6.1 Resnik's information-based approach The key idea underlying Resnik's (1995) approach is the intuition that one criterion of similarity between two concepts is "the extent to which they share information in common", which in an IS-A taxonomy can be determined by inspecting the relative position of the most-specific concept that subsumes them both. This intuition seems to be indirectly captured by edge-counting methods (such as that of Rada and colleagues; Section 2.4 above), in that "if the minimal path of IS-A links between two nodes is long, that means it is necessary to go high in the taxonomy, to more abstract concepts, in order to find a least upper bound". An example given by Resnik is the difference in the relative positions of the most-specific subsumer of nickel and dime — coin — and that of nickel and credit card — medium of exchange, as seen in Figure 1.

In mathematical terms, for any concept c in the taxonomy, let $\mathrm{p}(c)$ be the probability of encountering an *instance* of concept c. Following the standard definition from information theory, the *information content* of c, $\mathrm{IC}(c)$, is then $-\log \mathrm{p}(c)$. Thus, we can define the semantic similarity of a pair of concepts c_1 and c_2, as

$$\mathrm{sim}_{\mathrm{R}}(c_1, c_2) = -\log \mathrm{p}(lso(c_1, c_2)) \; . \tag{8}$$

Notice that p is monotonic as one moves up the taxonomy: if c_1 IS-A c_2 then $\mathrm{p}(c_1) \leq \mathrm{p}(c_2)$. For example, whenever we encounter a nickel, we have encountered a coin (Figure 1), so $\mathrm{p}(\textsf{nickel}) \leq \mathrm{p}(\textsf{coin})$. As a consequence, the higher the position of the most specific subsumer for given two concepts in the taxonomy (*i.e.,* the more abstract it is), the lower

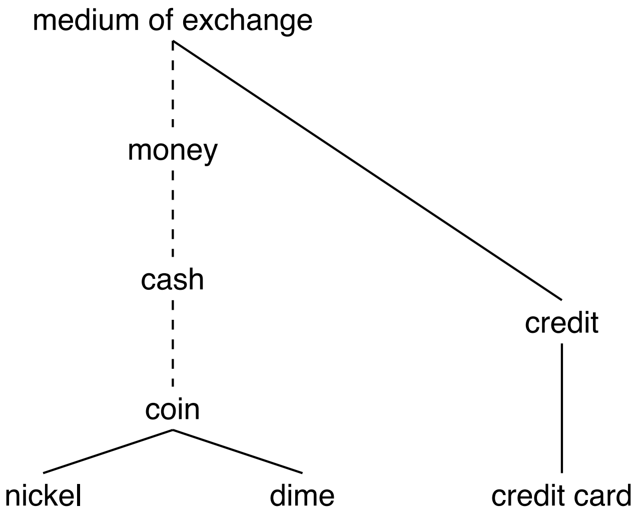

Figure 1
Fragment of the WordNet taxonomy, showing most-specific subsumers of nickel and dime and of nickel and credit card. Solid lines represent IS-A links; dashed lines indicate that some intervening nodes have been omitted. Adapted from Resnik (1995).

their similarity. In particular, if the taxonomy has a unique top node, its probability will be 1, so if the most specific subsumer of a pair of concepts is the top node, their similarity will be $-\log(1) = 0$, as desired.

In Resnik's experiments, the probabilities of concepts in the taxonomy were estimated from noun frequencies gathered from the one-million-word Brown Corpus of American English (Francis and Kučera, 1982). The key characteristic of his counting method is that an individual occurrence of any noun in the corpus "was counted as an occurrence of each taxonomic class containing it". For example, an occurrence of the noun *nickel* was, in accordance with Figure 1, counted towards the frequency of nickel, coin, and so forth. Notice that, as a consequence of using raw (non-disambiguated) data, encountering a polysemous word contributes to the counts of all its senses. So in case of *nickel*, the counts of both the coin and the metal senses will be increased. Formally,

$$\mathrm{p}(c) = \frac{\sum_{w \in W(c)} \mathrm{count}(w)}{N}\,, \tag{9}$$

where $W(c)$ is the set of words (nouns) in the corpus whose senses are subsumed by concept c, and N is the total number of word (noun) tokens in the corpus that are also present in WordNet.

Thus Resnik's approach attempts to deal with the problem of varying link distances (see Section 2.5) by generally downplaying the role of network edges in the determination of the degree of semantic proximity: edges are used solely for locating superordinates of a pair of concepts; in particular, the number of links does not figure in any of the formulas pertaining to the method; and numerical evidence comes from corpus statistics, which are associated with nodes. This rather selective use of the structure of the taxonomy has its drawbacks, one of which is the indistinguishability, in terms of semantic distance, of any two pairs of concepts having the same most-specific subsumer. For example, in Figure 1, we find that $\mathrm{sim}_{\mathrm{R}}(\textsf{money}, \textsf{credit}) = \mathrm{sim}_{\mathrm{R}}(\textsf{dime}, \textsf{credit card})$, because in each case the *lso* is medium of exchange, whereas, for an *edge-based* method such as Leacock and Chodorow's (Section 2.5.3), clearly this is not so, as the number of edges in each case is different.

2.6.2 Jiang and Conrath's combined approach Reacting to the disadvantages of Resnik's method, Jiang and Conrath's (1997) idea was to synthesize edge- and node-based techniques by restoring network edges to their dominant role in similarity computations, and using corpus statistics as a secondary, corrective factor. A complete exegesis of their work is presented by Budanitsky (1999); here we summarize only their conclusions.

In the framework of the IS-A hierarchy, Jiang and Conrath postulated that the semantic distance of the link connecting a child-concept c to its parent-concept $par(c)$ is proportional to the conditional probability $\mathrm{p}(c \mid par(c))$ of encountering an instance of c given an instance of $par(c)$. More specifically,

$$\mathrm{dist}_{\mathrm{JC}}(c, par(c)) = -\log \mathrm{p}(c \mid par(c)) \ . \tag{10}$$

By definition,

$$\mathrm{p}(c \,|\, par(c)) = \frac{\mathrm{p}(c \& par(c))}{\mathrm{p}(par(c))} \ . \tag{11}$$

If we adopt Resnik's scheme for assigning probabilities to concepts (Section 2.6.1), then $\mathrm{p}(c \& par(c)) = \mathrm{p}(c)$, since any instance of a child is automatically an instance of its parent. Then,

$$\mathrm{p}(c|par(c)) = \frac{\mathrm{p}(c)}{\mathrm{p}(par(c))} \ , \tag{12}$$

and, recalling the definition of information content,

$$\mathrm{dist}_{\mathrm{JC}}(c, par(c)) = \mathrm{IC}(c) - \mathrm{IC}(par(c)) \ . \tag{13}$$

Given this as the measure of semantic distance from a node to its immediate parent, the semantic distance between an arbitrary pair of nodes was taken, as per common practice, to be the sum of the distances along the shortest path that connects the nodes:

$$\mathrm{dist}_{\mathrm{JC}}(c_1, c_2) = \sum_{c \in Path(c_1,c_2) \setminus lso(c_1,c_2)} \mathrm{dist}_{\mathrm{JC}}(c, par(c)) \ , \tag{14}$$

where $Path(c_1, c_2)$ is the set of all the nodes in the shortest path from c_1 to c_2. The node $lso(c_1, c_2)$ is removed from $Path(c_1, c_2)$ in (14)), because it has no parent in the set. Expanding the sum in the right-hand side of Equation 14, plugging in the expression for parent–child distance from Equation 13, and performing necessary eliminations results in the following final formula for the semantic distance between concepts c_1 and c_2:

$$\begin{aligned} \mathrm{dist}_{\mathrm{JC}}(c_1, c_2) &= \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 \times \mathrm{IC}(lso(c_1, c_2)) && (15) \\ &= 2 \log \mathrm{p}(lso(c_1, c_2)) - (\log \mathrm{p}(c_1) + \log \mathrm{p}(c_2)) \ . && (16) \end{aligned}$$

2.6.3 Lin's universal similarity measure Noticing that all of the similarity measures known to him were tied to a particular application, domain, or resource, Lin (1998b) attempted to define a measure of similarity that would be both universal (applicable to arbitrary objects and "not presuming any form of knowledge representation") and theoretically justified ("derived from a set of assumptions", instead of "directly by a formula", so that "if the assumptions are deemed reasonable, the similarity measure necessarily follows"). He used the following three intuitions as a basis:

1. The similarity between arbitrary objects A and B is related to their commonality; the more commonality they share, the more similar they are.
2. The similarity between A and B is related to the differences between them; the more differences they have, the less similar they are.
3. The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

Lin defined the *commonality* between A and B as the information content of "the proposition that states the commonalities" between them, formally

$$\mathrm{IC}(\mathrm{comm}(A,B)) \ , \tag{17}$$

and the *difference* between A and B as

$$\mathrm{IC}(\mathrm{descr}(A,B)) - \mathrm{IC}(\mathrm{comm}(A,B)) \ , \tag{18}$$

where descr(A, B) is a proposition describing what A and B are.

Given these assumptions and definitions and the apparatus of information theory, Lin proved the following:

> **Similarity Theorem:** The similarity between A and B is measured by the ratio between the amount of information needed to state their commonality and the information needed to fully describe what they are:
>
> $$\mathrm{sim_L}(A,B) = \frac{\log \mathrm{p}(\mathrm{comm}(A,B))}{\log \mathrm{p}(\mathrm{descr}(A,B))} \ . \tag{19}$$

His measure of similarity between two concepts in a taxonomy is a corollary of this theorem:

$$\mathrm{sim_L}(c_1,c_2) = \frac{2 \times \log \mathrm{p}(\mathit{lso}(c_1,c_2))}{\log \mathrm{p}(c_1) + \log \mathrm{p}(c_2)} \ , \tag{20}$$

where the probabilities p(c) are determined in a manner analogous to Resnik's p(c) (Equation 9).

3 Evaluation methods

How can we reason about and evaluate computational measures of semantic relatedness? Three kinds of approaches are prevalent in the 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. 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. (In Section 4.1 below, we will employ the rather limited data that such experiments have obtained to date.)

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.

In the remainder of this paper, we will use the second and the third methods to compare several different measures (sections 4 and 5 respectively). We focus on measures that use WordNet (Fellbaum, 1998) as their knowledge source (to keep that as a constant) and that permit straightforward implementation as functions in a programming language. Therefore, we select the following five measures: Hirst and St-Onge's (Section 2.4), Jiang and Conrath's (Section 2.6.2), Leacock and Chodorow's (Section 2.5.3), Lin's (Section 2.6.3), Resnik's (Section 2.6.1).[5] The first is claimed as a measure of semantic relatedness because it uses all noun relations in WordNet; the others are claimed only as measures of similarity because they use only the hyponymy relation. We implemented each measure, and used the Brown Corpus as the basis for the frequency counts needed in the information-based approaches.[6]

4 Comparison with human ratings of semantic relatedness

In this section we compare the five chosen measures by how well they reflect human judgments of semantic relatedness. In addition, we will use the data that we obtain in this section to set closeness thresholds for the application-based evaluation of each measure in Section 5.

4.1 Data

As a part of an investigation into "the relationship between similarity of context and similarity of meaning (synonymy)", Rubenstein and Goodenough (1965) obtained "synonymy judgements" from 51 human subjects on 65 pairs of words. The pairs ranged from "highly synonymous" to "semantically unrelated", and the subjects were asked to rate them, on the scale of 0.0 to 4.0, according to their "similarity of meaning" (see

[5] We also attempted to implement Sussna's (1993; 1997) measure (Section 2.5.1), but ran into problems because a key element depended closely on the particulars of an earlier version of WordNet; see (Budanitsky, 1999) for details. We did not include Wu and Palmer's measure (Section 2.5.2) because Lin (1998b) has shown it to be a special case of his measure in which all child–parent probabilities are equal.

[6] In their original experiments, Lin and Jiang and Conrath used SemCor, a sense-tagged subset of the Brown Corpus, as their empirical data; but we decided to follow Resnik in using the full and untagged corpus. While this means trading accuracy for size, we believe that using a non-disambiguated corpus constitutes a more-general approach, as the availability and size of disambiguated texts such as SemCor is highly limited.

Table 1, columns 2 and 3). For a similar study, Miller and Charles (1991) chose 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 judgements from 38 subjects, given the same instructions as above, on those 30 pairs (see Table 2, columns 2 and 3).[7]

4.2 Method

For each of our five implemented measures, we obtained similarity or relatedness scores for the human-rated pairs. Where either or both of the words had more than one synset in WordNet, we took the most-related pair of synsets. For the measures of Resnik, Jiang and Conrath, and Lin, this replicates and extends a study by each of the original authors of their own measure.

4.3 Results

The mean ratings from Rubenstein and Goodenough's and Miller and Charles's original experiments (labeled 'Humans') and the ratings of the Rubenstein–Goodenough and Miller–Charles word pairs produced by (our implementations of) the Hirst–St-Onge, Jiang–Conrath, Leacock–Chodorow, Lin, and Resnik measures of relatedness are given in Tables 1 and 2, and in Figures 2 and 3.[8]

4.4 Discussion

When comparing two sets of ratings, we are interested in the strength of the linear association between two quantitative variables, so we follow Resnik (1995) in summarizing

[7] As a result of a typographical error that occurred in the course of either Miller and Charles's actual experiments or in its publication, the Rubenstein–Goodenough pair *cord–smile* became *chord–smile*. Probably because of the comparable degree of (dis)similarity, the error was not discovered and the latter pair has been used in all subsequent work.

[8] We have kept the original orderings of the pairs: from dissimilar to similar for the Rubenstein–Goodenough data and from similar to dissimilar for Miller–Charles. This explains why the two groups of graphs (Figures 2 and 3) as wholes have the opposite directions. Notice that because dist_{JC} measures *distance*, the Jiang–Conrath plot has a slope opposite to the rest of each group.

Table 1
Human and computer ratings of the Rubenstein–Goodenough set of word pairs (*part 1 of 2*).

#	Pair		Humans	rel$_{\mathrm{HS}}$	dist$_{\mathrm{JC}}$	sim$_{\mathrm{LC}}$	sim$_{\mathrm{L}}$	sim$_{\mathrm{R}}$
1	cord	smile	0.02	0	19.6	1.38	0.09	1.17
2	rooster	voyage	0.04	0	26.9	0.91	0.00	0.00
3	noon	string	0.04	0	22.6	1.50	0.00	0.00
4	fruit	furnace	0.05	0	18.5	2.28	0.14	1.85
5	autograph	shore	0.06	0	22.7	1.38	0.00	0.00
6	automobile	wizard	0.11	0	17.8	1.50	0.09	0.97
7	mound	stove	0.14	0	17.2	2.28	0.22	2.90
8	grin	implement	0.18	0	16.6	1.28	0.00	0.00
9	asylum	fruit	0.19	0	19.5	2.28	0.14	1.85
10	asylum	monk	0.39	0	25.6	1.62	0.07	0.97
11	graveyard	madhouse	0.42	0	29.7	1.18	0.00	0.00
12	glass	magician	0.44	0	22.8	1.91	0.07	0.97
13	boy	rooster	0.44	0	17.8	1.50	0.21	2.38
14	cushion	jewel	0.45	0	22.9	2.28	0.13	1.85
15	monk	slave	0.57	94	18.9	2.76	0.21	2.53
16	asylum	cemetery	0.79	0	28.1	1.50	0.00	0.00
17	coast	forest	0.85	0	20.2	2.28	0.12	1.50
18	grin	lad	0.88	0	20.8	1.28	0.00	0.00
19	shore	woodland	0.90	93	19.3	2.50	0.13	1.50
20	monk	oracle	0.91	0	22.7	2.08	0.18	2.53
21	boy	sage	0.96	93	19.9	2.50	0.20	2.53
22	automobile	cushion	0.97	98	15.0	2.08	0.27	2.90
23	mound	shore	0.97	91	12.4	2.76	0.49	6.19
24	lad	wizard	0.99	94	16.5	2.76	0.23	2.53
25	forest	graveyard	1.00	0	24.5	1.76	0.00	0.00
26	food	rooster	1.09	0	17.4	1.38	0.10	0.97
27	cemetery	woodland	1.18	0	25.0	1.76	0.00	0.00
28	shore	voyage	1.22	0	23.7	1.38	0.00	0.00
29	bird	woodland	1.24	0	18.1	2.08	0.13	1.50
30	coast	hill	1.26	94	10.8	2.76	0.53	6.19
31	furnace	implement	1.37	93	15.8	2.50	0.18	1.85
32	crane	rooster	1.41	0	12.8	2.08	0.58	8.88
33	hill	woodland	1.48	93	18.2	2.50	0.14	1.50
34	car	journey	1.55	0	16.3	1.28	0.00	0.00
35	cemetery	mound	1.69	0	23.8	1.91	0.00	0.00
36	glass	jewel	1.78	0	22.0	2.08	0.14	1.85
37	magician	oracle	1.82	98	1.0	3.50	0.96	13.58
38	crane	implement	2.37	94	15.6	2.76	0.27	2.90
39	brother	lad	2.41	94	16.3	2.76	0.23	2.53
40	sage	wizard	2.46	93	22.8	2.50	0.18	2.53
41	oracle	sage	2.61	0	26.2	2.08	0.16	2.53
42	bird	crane	2.63	97	7.4	3.08	0.70	8.88
43	bird	cock	2.63	150	5.4	4.08	0.76	8.88
44	food	fruit	2.69	0	10.2	2.28	0.22	1.50
45	brother	monk	2.74	93	19.2	2.50	0.20	2.53
46	asylum	madhouse	3.04	150	0.2	4.08	0.99	15.70
47	furnace	stove	3.11	0	20.5	2.08	0.13	1.85
48	magician	wizard	3.21	200	0.00	5.08	1.00	13.58
49	hill	mound	3.29	200	0.00	5.08	1.00	12.08
50	cord	string	3.41	150	2.2	4.08	0.89	9.25

Table 1
Human and computer ratings of the Rubenstein–Goodenough set of word pairs (*part 2 of 2*).

#	Pair		Humans	$\mathrm{rel}_{\mathrm{HS}}$	$\mathrm{dist}_{\mathrm{JC}}$	$\mathrm{sim}_{\mathrm{LC}}$	$\mathrm{sim}_{\mathrm{L}}$	$\mathrm{sim}_{\mathrm{R}}$
51	glass	tumbler	3.45	150	5.9	4.08	0.79	11.34
52	grin	smile	3.46	200	0.0	5.08	1.00	10.41
53	serf	slave	3.46	0	19.8	2.28	0.34	5.28
54	journey	voyage	3.58	150	5.2	4.08	0.74	7.71
55	autograph	signature	3.59	150	2.4	4.08	0.92	14.29
56	coast	shore	3.60	150	0.8	4.08	0.96	11.12
57	forest	woodland	3.65	200	0.0	5.08	1.00	11.23
58	implement	tool	3.66	150	1.1	4.08	0.91	6.20
59	cock	rooster	3.68	200	0.0	5.08	1.00	14.29
60	boy	lad	3.82	150	5.3	4.08	0.72	8.29
61	cushion	pillow	3.84	150	0.7	4.08	0.97	13.58
62	cemetery	graveyard	3.88	200	0.0	5.08	1.00	13.76
63	automobile	car	3.92	200	0.0	5.08	1.00	8.62
64	midday	noon	3.94	200	0.0	5.08	1.00	15.96
65	gem	jewel	3.94	200	0.0	5.08	1.00	14.38

Table 2
Human and computer ratings of the Miller–Charles set of word pairs.

#	Pair		Humans	$\mathrm{rel}_{\mathrm{HS}}$	$\mathrm{dist}_{\mathrm{JC}}$	$\mathrm{sim}_{\mathrm{LC}}$	$\mathrm{sim}_{\mathrm{L}}$	$\mathrm{sim}_{\mathrm{R}}$
1	car	automobile	3.92	200	0.0	5.08	1.00	8.62
2	gem	jewel	3.84	200	0.0	5.08	1.00	14.38
3	journey	voyage	3.84	150	5.2	4.08	0.74	7.71
4	boy	lad	3.76	150	5.3	4.08	0.72	8.29
5	coast	shore	3.70	150	0.9	4.08	0.96	11.12
6	asylum	madhouse	3.61	150	0.2	4.08	0.99	15.70
7	magician	wizard	3.50	200	0.0	5.08	1.00	13.58
8	midday	noon	3.42	200	0.0	5.08	1.00	15.96
9	furnace	stove	3.11	0	20.5	2.08	0.13	1.85
10	food	fruit	3.08	0	10.2	2.28	0.22	1.50
11	bird	cock	3.05	150	5.4	4.08	0.76	8.88
12	bird	crane	2.97	97	7.4	3.08	0.70	8.88
13	tool	implement	2.95	150	1.1	4.08	0.91	6.20
14	brother	monk	2.82	93	19.2	2.50	0.20	2.53
15	lad	brother	1.66	94	16.3	2.76	0.23	2.53
16	crane	implement	1.68	94	15.7	2.76	0.27	2.90
17	journey	car	1.16	0	16.3	1.28	0.00	0.00
18	monk	oracle	1.10	0	22.7	2.08	0.18	2.53
19	cemetery	woodland	0.95	0	25.0	1.76	0.00	0.00
20	food	rooster	0.89	0	17.4	1.38	0.10	0.97
21	coast	hill	0.87	94	10.9	2.76	0.53	6.19
22	forest	graveyard	0.84	0	24.6	1.76	0.00	0.00
23	shore	woodland	0.63	93	19.3	2.50	0.13	1.50
24	monk	slave	0.55	94	18.9	2.76	0.21	2.53
25	coast	forest	0.42	0	20.2	2.28	0.12	1.50
26	lad	wizard	0.42	94	16.5	2.76	0.23	2.53
27	chord	smile	0.13	0	20.2	1.62	0.18	2.23
28	glass	magician	0.11	0	22.8	1.91	0.07	0.97
29	rooster	voyage	0.08	0	26.9	0.91	0.00	0.00
30	noon	string	0.08	0	22.6	1.50	0.00	0.00

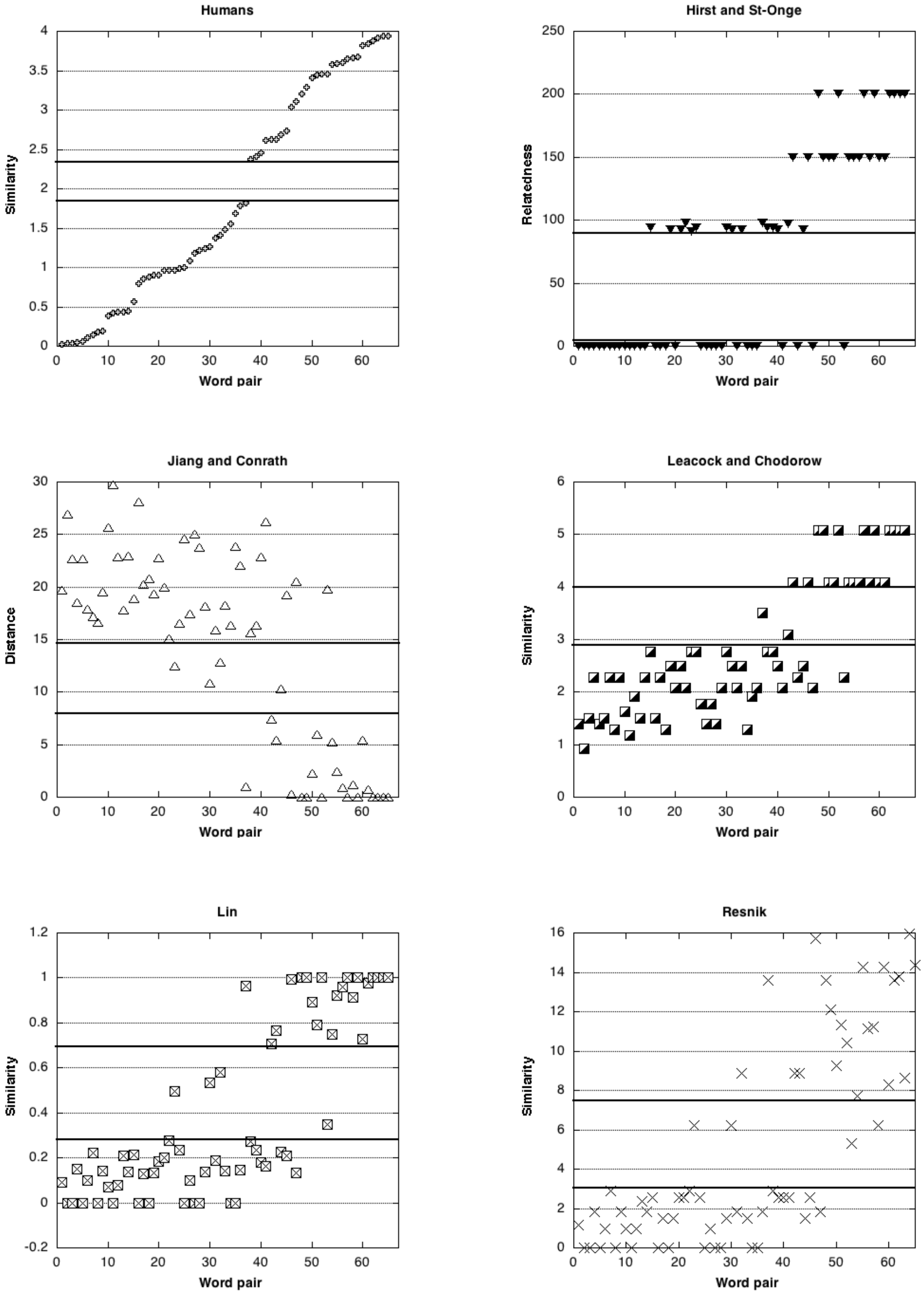

Figure 2
Human and computer ratings of the Rubenstein—Goodenough set of word pairs, with sparse bands marked (see text). *(a)* Rubenstein and Goodenough’s human ratings. *(b)* The word pairs rated by the Hirst—St-Onge similarity measure. *(c)* The word pairs rated by the Jiang—Conrath distance measure. *(d)* The word pairs rated by the Leacock–Chodorow similarity measure. *(e)* The word pairs rated by the Lin similarity measure. *(f)* The word pairs rated by the Resnik similarity measure.

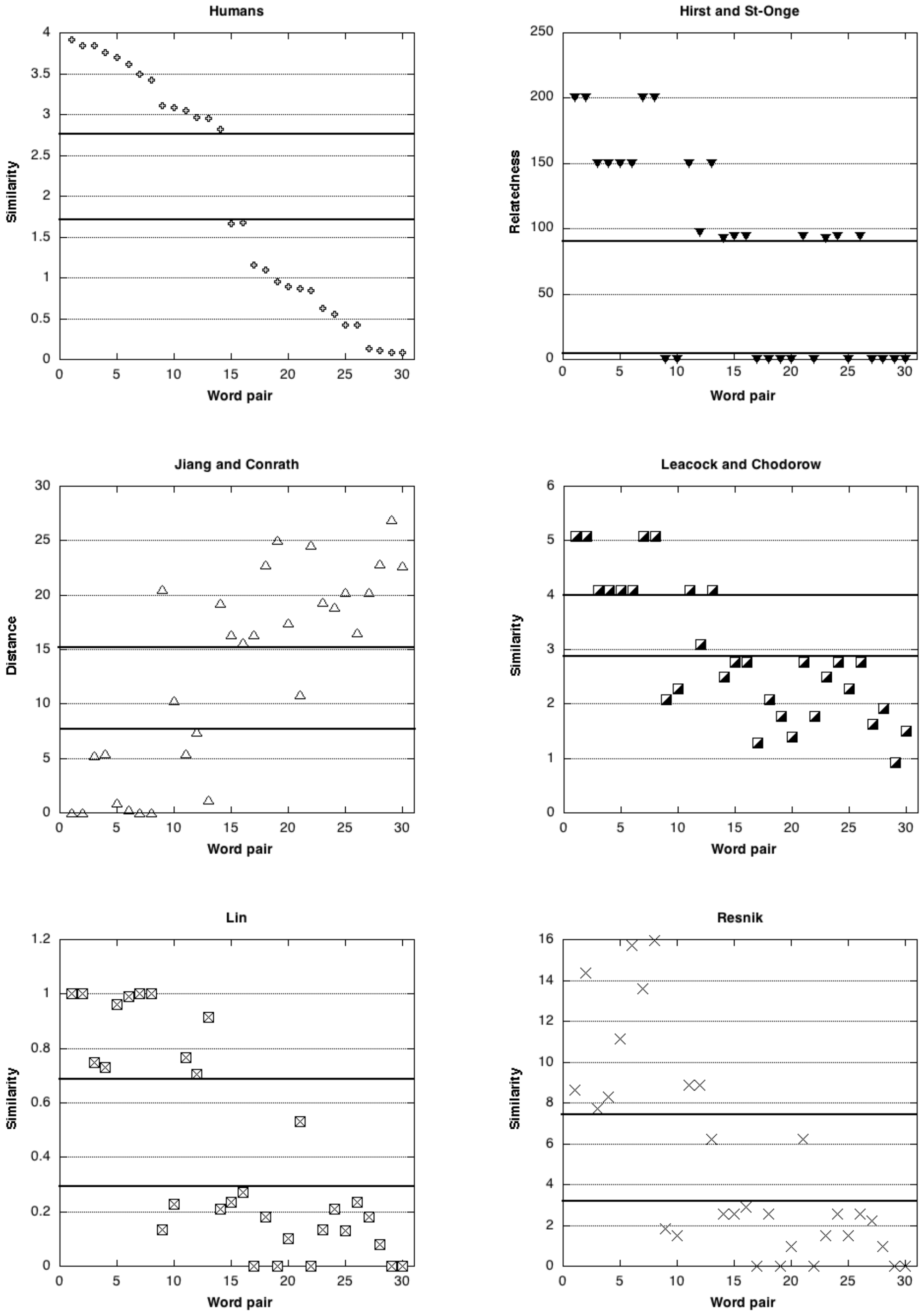

Figure 3
Human and computer ratings of the Miller–Charles set of word pairs, with sparse bands marked (see text). *(a)* Miller and Charles's human ratings. *(b)* The word pairs rated by the Hirst—St-Onge similarity measure. *(c)* The word pairs rated by the Jiang—Conrath distance measure. *(d)* The word pairs rated by the Leacock–Chodorow similarity measure. *(e)* The word pairs rated by the Lin similarity measure. *(f)* The word pairs rated by the Resnik similarity measure.

Table 3
The absolute values of the coefficients of correlation between human ratings of similarity (by Miller and Charles and by Rubenstein and Goodenough) and the five computational measures.

Measure	M&C	R&G
Hirst and St-Onge, rel_{HS}	.744	.786
Jiang and Conrath, dist_{JC}	.850	.781
Leacock and Chodorow, sim_{LC}	.816	.838
Lin, sim_{L}	.829	.819
Resnik, sim_{R}	.774	.779

the comparison results by means of the coefficient of correlation of each computational measure with the human ratings; see Table 3. (For Jiang and Conrath's measure, the coefficients are negative because their measure returns distance rather than similarity; so for convenience, we show absolute values in the table.) [9]

4.4.1 Comparison to upper bound To get an idea of the upper bound on performance of a computational measure, we can again refer to human performance. We have such an upper bound for the Miller and Charles word pairs (but not for the complete set of Rubenstein and Goodenough pairs): Resnik (1995) replicated Miller and Charles's experiment with 10 subjects and found that the average correlation with the Miller–Charles mean ratings over his subjects was 0.8848. While the difference between the (absolute) values of the highest and lowest correlation coefficients in the "M&C" column of Table 3 is of the order of 0.1, all of the coefficients compare quite favorably with this estimate of the upper bound; furthermore, the difference diminishes almost twofold as we consider the larger Rubenstein–Goodenough dataset (column "R&G" of Table 3).[10]

[9] Resnik (1995), Jiang and Conrath (1997), and Lin (1998b) report the coefficients of correlation between their measures and the Miller–Charles ratings to be 0.7911, 0.8282, and 0.8339, respectively, which differ slightly from the corresponding figures in Table 3. These discrepancies can be explained by possible minor differences in implementation (*e.g.*, the compound-word recognition mechanism used in collecting the frequency data), differences between the versions of WordNet used in the experiments (Resnik), and differences in the corpora used to obtain the frequency data (Jiang and Conrath, Lin). Also, the coefficients reported by Resnik and Lin are actually based on only 28 out of the 30 Miller–Charles pairs because of a noun missing from an earlier version of WordNet. Jarmasz and Szpakowicz (2003) repeated the experiment, obtaining similar results to ours in some cases and markedly different results in others; in their experiment, the correlations obtained with their measure that uses the hierarchy of *Roget's Thesaurus* exceeded those of all the WordNet measures.

[10] None of the differences in either column are statistically significant at the .05 level.

In fact, the measures are divided in their reaction to increasing the size of the dataset: the correlations of $\mathrm{rel}_{\mathrm{HS}}$, $\mathrm{sim}_{\mathrm{LC}}$, and $\mathrm{sim}_{\mathrm{R}}$ improve but those of $\mathrm{dist}_{\mathrm{JC}}$ and $\mathrm{sim}_{\mathrm{L}}$ deteriorate. This division might not be arbitrary: the last two depend on the same three quantities, $\log \mathrm{p}(c_1)$, $\log \mathrm{p}(c_2)$, and $\log \mathrm{p}(lso(c_1, c_2))$ (see Equations 16 and 20). (In fact, the coefficient for $\mathrm{sim}_{\mathrm{R}}$, which depends on only one of the three quantities, $\log \mathrm{p}(lso(c_1, c_2))$, improves only in the third digit.) However, with the present paucity of evidence, this connection remains hypothetical.

4.4.2 Differences in the performance and behavior of the measures We now examine the results of each of the measures and the differences between them. To do this, we will sometimes look at differences in their behavior on individual word pairs.

Looking at the graphs in Figures 2 and 3, we see that the discrete nature of the Hirst–St-Onge and Leacock–Chodorow measures is much more apparent than that of the others: *i.e.,* the values that they can take on are just a fixed number of *levels*. This is, of course, a result of their being based on the same highly discrete factor: the path length. As a matter of fact, a more substantial correspondence between the two measures can be recognized from the graphs and explained in the same way. In each dataset, the upper portions of the two graphs are identical: namely, the sets of pairs affording the highest and second-highest values of the two measures ($\mathrm{rel}_{\mathrm{HS}} \geq 150$, $\mathrm{sim}_{\mathrm{LC}} > 4$). This happens because these sets are composed of WordNet synonym and parent-child pairs, respectively.[11]

Further down the Y-axis, we find that for the Miller–Charles data, the two graphs still follow each other quite closely in the middle region (2.4–3.2 for $\mathrm{sim}_{\mathrm{LC}}$ and 90–100 for $\mathrm{rel}_{\mathrm{HS}}$). For the larger set of Rubenstein and Goodenough's, however, differences appear.

[11] More generally, the inverse image of the second highest value for $\mathrm{sim}_{\mathrm{LC}}$ is a proper subset of that for $\mathrm{rel}_{\mathrm{HS}}$, for the latter would also include all the antonym and meronym–holonym pairs. The two datasets at hand, however, do not contain any instances from these categories.

The pair *automobile–cushion* (22), for instance, is ranked even with *magician–oracle* (37) by the Hirst–St-Onge measure but far below both *magician–oracle* (37) and *bird–crane* (42) by Leacock–Chodorow (and, in fact, by all the other measures). The cause of such a high ranking in the former case is the following meronymic connection in WordNet:

> automobile/. . . /car HAS-A suspension/suspension system ('a system of springs or shock absorbers connecting the wheels and axles to the chassis of a wheeled vehicle') HAS-A cushion/shock absorber/shock ('a mechanical damper; absorbs energy of sudden impulses').

Since rel_{HS} is the only measure that takes meronymy (and other WordNet relations beyond IS-A) into account, no other measure detected this connection—nor did the human judges, whose task was to assess *similarity*, not generic *relatedness*; see Section 4.1).

Finally, at the bottom portion of these two graphs, the picture becomes very different, because rel_{HS} assigns all weakly-related pairs the value of zero. (In fact, it is this cut-off that we believe to be largely responsible for the relatively low ranking of the correlation coefficient of the Hirst–St-Onge measure.) In contrast, two other measures, Resnik's and Lin's, behave quite similarly to each other in the low-similarity region. In particular, their sets of zero-similarity pairs are identical, because the definitions of both measures include the term $\log p(lso(c_1, c_2))$, which is zero for the pairs in question.[12] For instance, for the pair *rooster–voyage* (M&C #29, R&G #2), the synsets rooster and voyage have different 'unique beginners', and hence their *lso*—in fact their sole common subsumer—is the (fake) global root (see Section 2.5.3), which is the only concept whose probability is 1:

> cock/rooster ('adult male chicken') IS-A . . . IS-A domestic

[12] Again (*cf* footnote 11), the former set actually constitutes a proper subset of the latter, as $\text{sim}_{\text{L}}(c_1, c_2)$ will also be zero if either concept does not occur in the frequency-corpus (see Equation 20). However, no such instances appear in the data.

> fowl/…/poultry IS-A … IS-A bird IS-A … IS-A animal/animate being/…/fauna IS-A life form/…/living thing ('any living entity') IS-A entity ('something having concrete existence; living or nonliving') IS-A global root,
>
> voyage IS-A journey/journeying IS-A travel/…/traveling IS-A change of location/…/motion IS-A change ('the act of changing something') IS-A action ('something done (usually as opposed to something said)') IS-A act/human action/human activity ('something that people do or cause to happen') IS-A global root.

Analogously, although perhaps somewhat more surprisingly for a human reader, the same is true of the pair *asylum–cemetery* (R&G #16):

> asylum/insane asylum/…/mental hospital IS-A hospital/infirmary IS-A medical building ('a building where medicine is practiced') IS-A building/edifice IS-A … IS-A artifact/artefact ('a man-made object') IS-A object/inanimate object/physical object ('a nonliving entity') IS-A entity IS-A global root,
>
> cemetery/graveyard/…/necropolis ('a tract of land used for burials') IS-A site ('the piece of land on which something is located (or is to be located)') IS-A position/place ('the particular portion of space occupied by a physical object') IS-A … IS-A location ('a point or extent in space') IS-A global root.

Looking back at the high-similarity portion of the graphs, but now taking into consideration the other three measures, we can make a couple more observations. First, the graphs of all of the measures except Resnik's exhibit a 'line' of synonyms (comprising four points for the Miller–Charles dataset and nine points for Rubenstein–Goodenough) at the top (bottom for Jiang and Conrath's measure). In the case of Resnik's measure, $\mathrm{sim}_{\mathrm{R}}(c,c) = -\log \mathrm{p}(lso(c,c)) = -\log \mathrm{p}(c)$ (see Equation 8), and hence the similarity of a concept to itself varies from one concept to another. Second, these 'lines' are not continuous, as one might expect from the graphs of the human judgments: for the Miller–Charles set, for instance, the line includes pairs 1, 2, 7, and 8, but omits pairs 3–6. This peculiarity is due entirely to WordNet, according to which *gem* and *jewel* (# 2) and *magician* and *wizard* (# 7) are synonyms, whereas *journey* and *voyage* (# 3), *boy* and *lad* (pair 4), and even *asylum* and *madhouse* (# 6) are not, but rather are related by IS-A :

> voyage ('a journey to some distant place') IS-A journey/journeying ('the act of traveling from one place to another'),
>
> lad/laddie/cub/sonny/sonny boy ('a male child (a familiar term of address to a boy)') IS-A boy/male child/child ('a young male person'),
>
> madhouse/nuthouse/. . . /sanatorium ('pejorative terms for an insane asylum') IS-A asylum/insane asylum/. . . /mental hospital ('a hospital for mentally incompetent or unbalanced persons').

Although, as we saw above, already for two measures the details of their medium-similarity regions differ, there appears to be an interesting commonality at the level of general structure: in the vicinity of sim = 2, the plots of human similarity ratings for both the Miller–Charles and the Rubenstein–Goodenough word pairs display a very clear horizontal band that contains no points. For the Miller–Charles data (Figure 3), the band separates the pair *crane–implement* (16) from *brother–monk* (14),[13] and for the Rubenstein-Goodenough set (Figure 2), it separates *magician–oracle* (37) from *crane–implement* (38).

[13] For some reason, Miller and Charles, while generally ordering their pairs from least to most similar, put *crane–implement* (16) after *lad–brother*(15), even though the former was rated more similar.

On the graphs of the computed ratings, these empty bands correspond to regions with at most a few points—no more than two points for the Miller–Charles set and no more than four for the Rubenstein–Goodenough set. These regions are shown in Figures 3(b)–(f) and 2(b)–(f). This commonality among the measures suggests that if we were to partition the set of all word pairs into those that are deemed to be related and those that are deemed unrelated, the boundary between the two subsets for each measure (and for the human judgments, for that matter) would lie somewhere within these regions.

4.4.3 The limitations of this analysis 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.[14]

Even more importantly, there are serious methodological problems with this approach. It was implicit in the Rubenstein–Goodenough and Miller–Charles experiments that subjects were to use the dominant sense of the target words or mutually triggering related senses. But often 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.[15]

[14] Evgeniy Gabrilovich has recently made available a dataset of similarity judgments of 353 English word pairs that were used by Finkelstein et al (2002). Unfortunately, this set is still very small, and, as Jarmasz and Szpakowicz (2003) point out, is culturally and politically biased. And the scarcely larger set of synonymy norms for nouns created by Whitten, Suter, and Frank (1979) covers only words with quite closely related senses, and hence is not useful here either.

[15] In their creation of a set of synonymy norms for nouns, Whitten, Suter, and Frank (1979) observed frequent artifacts stemming from the order of presentation of the stimuli that seem to be due to the

5 An application-based evaluation of measures of relatedness

We now turn to 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*.

While malapropism correction is also a useful application in its own right, it is particularly appropriate for evaluating measures of semantic relatedness. 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; Morris and Hirst, 2004). 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 — that is, introduding semantic anomalies such as malapropisms — and looking at the ability of the different relatedness measures to detect and correct the perturbations.

5.1 Malapropism detection and correction as a testbed

Our malapropism corrector (Hirst and Budanitsky, 2005) is based on the idea behind that of Hirst and St-Onge (1998): look for semantic anomalies that can be removed by small changes to spelling.[16] 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

practical impossibility of forcing a context of interpretation in the experimental setting.

[16] 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.

found to be semantically unrelated to accepted senses of all other nearby words, but some sense of a *spelling variation* 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; a user would be warned of this possibility. 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, as is the precise definition of *spelling variation*; see Hirst and Budanitsky (2005).

This method makes the following assumptions:

- A real-word spelling error is unlikely to be semantically related to the text.[17]
- 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. Regardless of the degree of adequacy of its performance, it is a "level playing field" for comparison of the measures. Hirst and Budanitsky (2005) discuss the practical aspects of the method and compare it with other approaches to the same problem.

5.2 Method

To test the measures in this application, we need a sufficiently large corpus of malapropisms in their context, each identified and annotated with its correction. Since

[17] In fact, there is a semantic bias in human typing errors (Fromkin, 1980), but not in the malapropism generator to be described below.

no such corpus of naturally occurring malapropisms exists, we created one artificially. 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.[18] For example, in a sentence beginning *To win the case, which was filed shortly after the indictment and is pending in Manhattan federal court . . .* , the word *case* was replaced by *cage*. This gave us a corpus with 1408 malapropisms among 107,233 candidates.[19] We then tried to detect and correct the malapropisms by the algorithm outlined above, using in turn each of the five measures of semantic relatedness. For each, we used four different *search scopes*, *i.e.,* 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 complete article (scope = MAX).

We also needed to set a *threshold* of "relatedness" for each of the measures. This is because the malapropism-detection algorithm requires a boolean *related–unrelated* judgment, but each of the measures that we tested instead returns a numerical value of relatedness or similarity, and nothing in the measure (except for the Hirst–St-Onge measure) indicates which values count as "close". Moreover, 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 the Rubenstein–Goodenough pairs (the near-synonyms) from the lower level, as we described in Section 4.4.2.

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

[18] Articles too small to warrant such a replacement (19 in total) were excluded from further consideration.

[19] We assume that the original *WSJ*, being carefully edited text, contains essentially no malapropisms of its own.

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 *correct suspect* if it is indeed a malapropism and a *false suspect* if it isn't. At the second stage, we say that, given a suspect, an *alarm* is raised when a spelling variation of the suspect is judged to be related to a nearby word or words; and if an alarm word is a malapropism, we say that the alarm is a *true alarm* and that the malapropism has been *detected*; otherwise, it is a *false alarm*. Then we can define precision (P), recall (R), and F-measure (F) for suspicion ($_S$), involving only the first stage, and detection ($_D$), involving both stages, as follows:

Suspicion:

$$P_S = \frac{\text{number of correct suspects}}{\text{number of suspects}}, \tag{21}$$

$$R_S = \frac{\text{number of correct suspects}}{\text{number of malapropisms in text}}, \tag{22}$$

$$F_S = \frac{2 \times P_S \times R_S}{P_S + R_S}. \tag{23}$$

Detection:

$$P_D = \frac{\text{number of true alarms}}{\text{number of alarms}}, \tag{24}$$

$$R_D = \frac{\text{number of true alarms}}{\text{number of malapropisms in text}}, \tag{25}$$

$$F_D = \frac{2 \times P_D \times R_D}{P_D + R_D}. \tag{26}$$

The precision, recall, and F values are computed as the mean values of these statistics across our collection of 481 articles, which constitute a random sample from the population of all *WSJ* articles. All the comparisons that we make below, except for comparisons to baseline, are performed with the Bonferroni multiple-comparison technique

(Agresti and Finlay, 1997), with an *overall* significance level of .05.

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.

Table 4 and Figure 4 show suspicion precision, recall, and F for each of the 5 × 4 combinations of measure and scope. The values of precision range from 3.3% (Resnik, scope = 1) to 11% (Jiang–Conrath, scope = MAX), with a mean of 6.2%, increasing with scope, as expected, for all measures except Hirst–St-Onge. More specifically, differences in precision are statistically significant for the difference between scope = 5 and scope = MAX for Leacock–Chodorow and between 1 and larger scopes for Lin, Resnik, and Jiang–Conrath; there are no significant differences for Hirst–St-Onge, which hence appears flat overall. The values of recall range from just under 6% (Hirst–St-Onge, scope = MAX) to more than 72% (Resnik, scope = 1), with a mean of 39.7%, decreasing with scope, as expected. All differences in recall are statistically significant, except between scope = 3 and scope = 5 for all measures other than Resnik's. F ranges from 5% (Hirst–St-Onge, scope = MAX) to 14% (Jiang–Conrath, scope = 5), with a mean of just under 10%. Even though values at the lower ends of these ranges appear small, they are still significantly ($p < .001$) better than chance, for which precision, recall, and F are all 1.29%.

Table 4
Precision (P_S), recall (R_S), and F-measure (F_S) 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

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 focused on the results for F_S by measure and scope to determine whether the performance of the five measures was significantly different and whether scope of search for relatedness made a significant difference.

Scope differences: For Jiang–Conrath and Resnik, the analysis confirms only that these 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 significantly

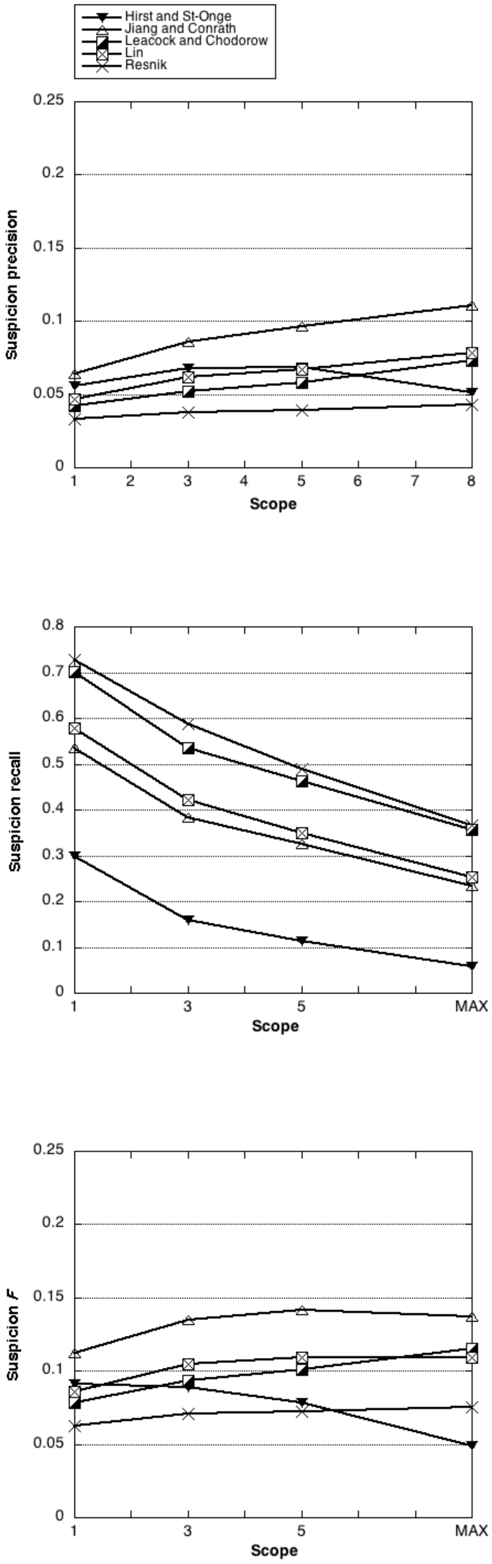

Figure 4
Precision (P_S), recall (R_S), and F-measure (F_S) for malapropism suspicion by measure and scope.

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.)

Differences between measures: Jiang–Conrath significantly outperforms the others in all scopes (except for Leacock–Chodorow and Lin at scope MAX, where it does better but not significantly so), followed by Lin and Leacock–Chodorow (whose performances are not significantly different from each other), 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 does best for the suspicion phase (and is optimal with scope = 5).

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 can only decrease (or, more precisely, not increase) from that for suspicion (*cf* equations 22 and 25). However, if a given measure of semantic relatedness is good, we expect the proportion of false alarms to reduce more considerably — far fewer false suspects will become alarms than correct suspects — thus resulting in higher precision for detection than for suspicion (*cf* equations 21 and 24).

Table 5 and Figure 5, show precision, recall, and F for each of the 5 × 4 measure–scope combinations, determined by the same method as those for suspicion. The values of recall range from 5.9% (Hirst–St-Onge, scope = MAX) to over 60% (Leacock–Chodorow, scope = 1). While these values are, as expected, lower than those for suspicion recall—R_D for each measure–scope combination is from 1 to 16 percentage points lower than the corresponding R_S—the decline is statistically significant for only 3 out of the 20 combinations.

The values of precision range from 6.7% (Hirst–St-Onge, scope = MAX) to just under 25% (Jiang–Conrath, scope = MAX), increasing, as expected, from suspicion precision; each combination increases from 1 to 14 percentage points; the increase is statistically significant for 18 out of the 20 combinations. Moreover, the increase in precision outweighs the decline in recall, and thus 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 precision, recall, and F ranges are significantly ($p < .001$) better than chance (which again is 1.29% for each), and the highest results are quite good (*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 fact that the method is inherently limited in the ways described earlier (Section 5.1). (See Hirst and Budanitsky (2005) for discussion of the practical usefulness of the method.)

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 — that is, scope doesn't affect the results. (Hence we can choose 1 as the optimal scope, since it involves the least amount of work, and Jiang and Conrath's method with scope = 1 as the optimal parameter combination for the malapropism detector.)

Differences between measures: The relative position of each measure's precision, recall, and F graphs for detection appears identical to that for suspicion, except for the precision and F graphs for Hirst–St-Onge, which slide further down. Statistical testing for F confirms this, with Jiang–Conrath leading, followed by Lin and Leacock–Chodorow together, Resnik, and then Hirst–St-Onge.

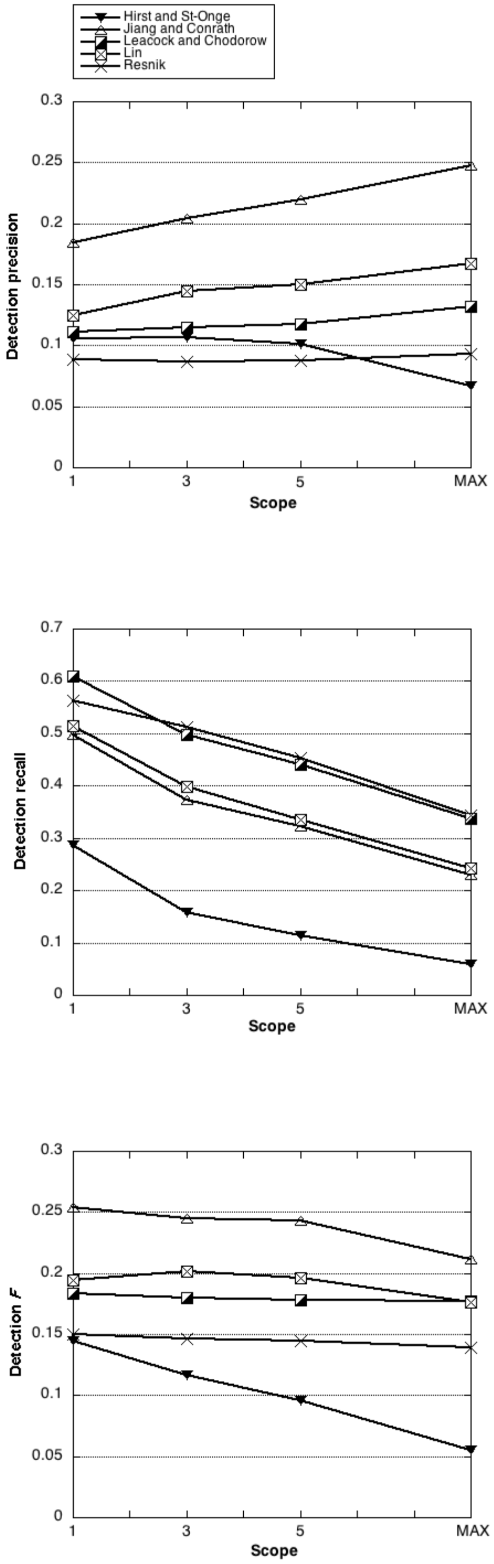

Figure 5
Precision (P_D), recall (R_D), and F-measure (F_D) for malapropism detection by measure and scope.

Table 5
Precision (P_D), recall (R_D), and F-measure (F_D) 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

5.4 Interpretation of the results

In our interpretation, we focus largely on the results for suspicion; those for detection 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 'under-relates', 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.[20] Indeed, for every scope, Resnik's measure generates

[20] It is debatable whether this metonymic sense of *house* should appear in WordNet at all, though given that

more suspects than any other measure—*e.g.,* an average of 62.5 per article for scope = 1, compared to a range of 15 to 47, with an 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 ability at discerning relatedness.

The same comparison can be made between the Jiang–Conrath and Lin measures. Even though both use the same information-content–based components, albeit in different arithmetic combinations, and show similar recall, the Jiang–Conrath measure shows superior precision and is 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 4 evidently reflects the corresponding difference in precision behavior: the Hirst–St-Onge suspicion precision graph is statistically flat, unlike the others. Ironically, given that this measure is the only one of the five that promises the semantic relatedness that we want rather than mere similarity, this poor performance appears to be 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. By comparison, for the other measures, the

it does, its relationship to *crowd* follows, and, as it happens, this sense was the correct one in the context for this particular case.

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.

6 Related work

6.1 Other applications of WordNet-based measures

Since the first publication of the initial results of this work (Budanitsky and Hirst, 2001), Pedersen and his colleagues (Pedersen et al., 2004) have made available a Perl implementation of the five WordNet-based measures (plus Wu and Palmer's and their own; see below) that has been used by a number of researchers in published work on other NLP applications. Generally, these results are consistent with our own. For example, Stevenson and Greenwood (2005) found Jiang–Conrath to be the best measure (out of "several", which they do not list) for their task of pattern induction for information extraction. Similarly, Kohomban and Lee (2005) found Jiang–Conrath the best (out of "various schemes", which they do not list) for their task of learning coarse-grained semantic classes. In word-sense disambiguation, Patwardhan, Banerjee, and Pedersen (2003) found Jiang–Conrath to be clearly the best of the five measures evaluated here, albeit edged out by their own new "Lesk" measure based on gloss overlaps;[21] and McCarthy et al (2004) found that the Jiang–Conrath and Lesk measures gave the best accuracy in their task of finding predominant word senses, with the results of the two being "comparable" but Jiang–Conrath being far more efficient. On the other hand, Corley and Mihalcea (2005) found little difference between the measures when using them in an algorithm for computing text similarity.

[21] Patwardhan et al's measure is based on the idea, originally due to Lesk (1986), of measuring the degree of relatedness of two words by the number of string overlaps in their dictionary definitions or glosses. Patwardhan et al extend this idea by also including overlaps with definitions of words that are one WordNet edge away from the comparison words. It is thus a hybrid method, with characteristics of both dictionary-based and network-based methods (see sections 2.1 and 2.3 above).

6.2 Measures of distributional similarity as proxies for measures of semantic relatedness

In Section 1.1, we mentioned that the lexical semantic relatedness or similarity that we have dealt with in this paper is a notion distinct from that of lexical distributional or co-occurrence similarity. However, a number of researchers, such as Dagan (2000), have promoted the hypothesis that distributional similarity can act as a useful proxy for semantic relatedness in many applications because it is based on corpus-derived data rather than manually created lexical resources; indeed, it could perhaps be used to automatically *create* (first-draft) lexical resources (Grefenstette, 1994). It is therefore natural to ask how distributional-similarity measures compare with the WordNet-based measures that we have looked at above.

Formally, by *distributional similarity* (or *co-occurrence similarity*) of two words w_1 and w_2, we mean that they tend to occur in similar contexts, for some definition of *context*; or that the set of words that w_1 tends to co-occur with is similar to the set that w_2 tends to co-occur with; or that if w_1 is substituted for w_2 in a context, its “plausibility” (Weeds, 2003; Weeds and Weir, 2005) is unchanged. The context considered may be a small or large window around the word, or an entire document; or it may be a syntactic relationship. For example, Weeds (2003; Weeds and Weir, 2005) (see below) took verbs as contexts for nouns in object position: so they regarded two nouns to be similar to the extent that they occur as direct objects of the same set of verbs. Lin (1998b; 1998a) considered other syntactic relationships as well, such as subject–verb and modifier–noun, and looked at both roles in the relationship.

Given this framework, many different methods of measuring distributional similarity have been proposed; see Dagan (2000), Weeds (2003), or Mohammad and Hirst (2005) for a review. For example, the set of words that co-occur with w_1 and those that co-occur with w_2 may be regarded as a feature vector of each and their similarity measured as the cosine between the vectors; or a measure may be based on the Kullback–Leibler divergence between the probability distributions $P(w \mid w_1)$ and $P(w \mid w_2)$, as, for example, Lee's (1999) α-skew divergence. Lin (1998b) uses his similarity theorem (equation 19 above) to derive a measure based on the degree of overlap of the sets of words with which w_1 and w_2, respectively, have positive mutual information.[22]

Words that are distributionally similar do indeed often represent semantically related concepts, and vice versa, as the following examples demonstrate. Weeds (2003), in her study of 15 distributional-similarity measures, found that words distributionally similar to *hope* (noun) included *confidence, dream, feeling,* and *desire*; Lin (1998b) found pairs such as *earnings–profit, biggest–largest, nylon–silk,* and *pill–tablet*. It is intuitively clear why these results occur: if two concepts are similar or related, it is likely that their role in the world will be similar, so similar things will be said about them, and so the contexts of occurrence of the corresponding words will be similar. And conversely (albeit with less certainty), if the contexts of occurrence of two words are similar, then similar things are being said about each, so they are playing similar roles in the world and hence are semantically similar — at least to the extent of these roles. Nonetheless, the limitations of this observation will become clear in our discussion below.

Three differences between semantic relatedness and distributional similarity are immediately apparent. First, while semantic relatedness is inherently a relation on concepts, as we emphasized in Section 1.1, distributional similarity is a (corpus-dependent) relation on words. In theory, of course, if one had a large-enough sense-tagged corpus, one could derive distributional similarities of word-senses. But in practice, apart from the lack of such corpora, distributional similarities are promoted exactly for applications such as various kinds of ambiguity resolution in which it is words rather than senses

[22] Do not confound Lin's distributional similarity measure with his semantic relatedness measure, sim_{L}, which has been discussed in earlier sections of this paper; but observe that both are derived from the same theorem.

that are available (see Weeds (2003) for an extensive list).

Second, whereas semantic relatedness is symmetric, distributional similarity is a potentially asymmetrical relationship. If distributional similarity is conceived of as substitutability, as Weeds and Weir(2005) and Lee (1999) emphasize, then asymmetries arise when one word appears in a subset of the contexts in which the other appears; for example, the adjectives that typically modify *apple* are a subset of those that modify *fruit*, so *fruit* substitutes for *apple* better than *apple* substitutes for *fruit*. While some distributional similarity measures, such as cosine, are symmetric, many, such as α-skew divergence and the co-occurrence retrieval models developed by Weeds and Weir, are not. But this is simply not an adequate model of semantic relatedness, for which substitutability is far too strict a requirement: *window* and *house* are semantically related, but they are not plausibly substitutable in most contexts.

Third, lexical semantic relatedness depends on a pre-defined lexicographic or other knowledge resource, whereas distributional similarity is relative to a corpus. In each case, matching the measures to the resource is a research problem in itself, as this paper and Weeds (2003) show, and anomalies can arise.[23] But the knowledge source for semantic relatedness is created by humans and may be presumed to be (in a weak sense) “true, unbiased, and complete”. A corpus, on the other hand, is not. Imbalance in the corpus and data sparseness is an additional source of anomalous results even for “good” measures. For example, Lin (1998b) found “peculiar” similarities that were “reasonable” for his corpus of news articles, such as *captive–westerner* (because in the news articles, more than half of the “westerners” mentioned were being held captive) and *audition–rite* (because both were infrequent and were modified by *uninhibited*).

[23] We have already remarked in Section 5.4 above on the promiscuity of the Hirst–St-Onge measure and its tendency to find connections such as cation–group. Similarly, one of the poorer measures that Weeds experimented with returned this list as the ten words most distributionally similar to *hope*: *hem, dissatisfaction, dismay, scepticism, concern, outrage, break, warrior, optimism, readiness*.

We now turn to the hypothesis that distributional similarity can usefully stand in for semantic relatedness in NLP applications such as malapropism detection. Weeds (2003) considered the hypothesis in detail. She carried out a number of experiments using data gathered from the British National Corpus on the distribution of a set of 2000 nouns with respect to the verbs of which they were direct objects, comparing a large number of proposed measures of distributional similarity. She applied ten of these measures to the Miller and Charles word-pairs (see Section 4.1 above); the absolute values of the correlations with the Miller and Charles human judgments was at best .62 (and at worst .26), compared with .74 to .85 for the semantic measures (table 3 above). Weeds also compared these measures on their ability to predict the *k* words that are semantically closest to a target word in WordNet, as measured by Lin's semantic similarity measure, $\mathrm{sim}_{\mathrm{L}}$. She found performance to be "generally fairly poor" (p. 162), and undermined by the effects of varying word frequencies.

Last, Weeds experimented with distributional measures in real-word spelling correction, much as we have defined it in Hirst and Budanitsky (2005) and in Section 5.1 above, but replacing the semantic relatedness measures with distributional similarity measures. However, she varied the experimental procedure in a number of ways, with the consequence that her results are not directly comparable to ours: her test data was the British National Corpus; scope was measured in words, not paragraphs; and relatedness thresholds were replaced by considering the *k* words most similar to the target word (and *k* was a parameter). The most significant difference, however, arose from the limitations due to data sparseness that are inherent in methods based on distributional similarity: the very small size of the set of words that could be corrected. Specifically, only malapropisms for which both the error and the correction occurred in the set of 2000 words for which Weeds had distributional data could be considered; and the ability to detect and correct the malapropism depended on other members of that set also being within the scope of the target word. It is therefore not surprising that the results were generally poor (and so were results for sim_{L} run under the same conditions). This severe limitation on the data means that this was not really a fair test of the principles underlying the hypothesis; a fair test would require data allowing the comparison of any two nouns (or better still, any two words) in WordNet, but obtaining such data for less-frequent words (possibly using the Web as the corpus) would be a massive task.

7 Conclusion

Our goal in this paper has been to evaluate resource-based measures of lexical semantic distance, or, equivalently, semantic relatedness, for use in natural language processing applications. Most of the work, however, was limited to the narrower notion of measures of similarity and how well they fill the broader role, because those measures are what current resources support best and hence what most current research has focused on. But ultimately it is the more-general idea of relatedness, not just similarity, that we need for most NLP methods and applications, because the goal, in one form or another, is to determine whether two smaller or larger pieces of text share a topic or some kind of closeness in meaning, and this need not depend on the presence of words that denote similar concepts. In word sense disambiguation, such an association with the context is frequently a sufficient basis for selecting or rejecting candidate senses (Banerjee and Pedersen, 2003); in our malapropism corrector, a word should be considered non-anomalous in the context of another if there is any kind of semantic relationship at all apparent between them. These relationships include not just hyponymy and the non-hyponymy relationships in WordNet such as meronymy but also *associative* and *ad hoc* relationships. As mentioned in the introduction, these can include just about any kind

Table 6
From Spellman, Holyoak, and Morrison's (2001) list of associative semantic relations.

Name	Example
IS-USED-TO	*bed–sleep*
WORKS-IN	*judge–court*
LIVES-IN	*camel–desert*
IS-THE-OUTSIDE-OF	*husk–corn*

of functional relation or frequent association in the world.[24]

For the last century, many researchers have attempted to enumerate these kinds of relationships. Some elements from a typical list (that of Spellman, Holyoak, and Morrison (2001)) are shown in Table 6. Morris and Hirst (2004; 2005) have termed these *non-classical* lexical semantic relationships (following Lakoff's (1987) non-classical categories), and Morris has shown in experiments with human subjects that around 60% of the lexical relationships that readers perceive in a text are of this nature (Morris, 2005). There is presently no catalogue of instances of these kinds of relationships let alone any incorporation of such relationships into a quantification of semantic distance. Nonetheless, there are clear intuitions to be captured here, and this should be a focus for future research.

But lists of such relationships can never be exhaustive, as lexical relationships can also arise ad hoc in context (Barsalou, 1983; Barsalou, 1989) — in particular, as comembership of an *ad hoc category*. For example, Morris's subjects reported that the words *sex, drinking,* and *drag racing* were semantically related, by all being "dangerous behaviors", in the context of an article about teenagers emulating what they see in movies. Thus lexical semantic relatedness is sometimes *constructed* in context and cannot always be determined purely from an a priori lexical resource such as WordNet.[25] It's very unclear how ad hoc semantic relationships could be quantified in any meaningful way,

[24] Don't confound "frequent association in the world" with the lexical co-occurrences that underlie the distributional similarity of Section 6.2.

[25] Indeed Murphy (2003) has suggested that semantic relations (of all types) are best conceived of as *metalexical*: derived from a (pre-existing) lexicon, but not part of it.

let alone compared with prior quantifications of the classical and non-classical relationships. However, ad hoc relationships accounted for only a small fraction of those reported by Morris's subjects (Morris, 2005). Their fact of their existence does not undermine the usefulness of computational methods for quantifying semantic distances for non–ad hoc relationships, and the continued development of such methods is an important direction for research.

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, Saif Mohammad, Jane Morris, Radford Neal, Manabu Okumura, Gerald Penn, Philip Resnik, David St-Onge, Suzanne Stevenson, Michael Sussna, and some of the anonymous reviewers of *Computational Linguistics*.

References

Alan Agresti and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences*. Prentice Hall, Upper Saddle River, NJ, 3rd edition.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, August.

Lawrence W. Barsalou. 1983. Ad hoc categories. *Memory and Cognition*, 11:211–227.

Lawrence W. Barsalou. 1989. Intra-concept similarity and its implications for inter-concept similarity. In Stella Vosniadou and Andrew Ortony, editors, *Similarity and analogical reasoning*, pages 76–121. Cambridge University Press.

J.R.L. Bernard, editor. 1986. *The Macquarie thesaurus*. Macquarie Library, Sydney, Australia.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 29–34.

Alexander Budanitsky. 1999. Lexical semantic relatedness and its application in natural language processing. Technical Report CSRG-390, Computer Systems Research Group, University of Toronto, August.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, MI, June.

Ido Dagan, Lillian Lee, and Fernando C.N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69, February.

Ido Dagan. 2000. Contextual word similarity. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 459–475. Marcel Dekker Inc.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20:116–131.

Winthrop Nelson Francis and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.

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

Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. English Language Series. Longman, New York.

Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11:87–111.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 13, pages 305–332. The MIT Press, Cambridge, MA.

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

Mario Jarmasz and Stan Szpakowicz. 2003. *Roget's Thesaurus* and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 212–219, September.

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 (ROCLING X), Taiwan.

Upali Sathyajith Kohomban and Wee Sun Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 34–41, Ann Arbor, MI, June.

Hideki Kozima and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL-93)*, pages 232–239, Utrecht.

Hideki Kozima and Akira Ito. 1997. Context-sensitive word distance by adaptive scaling of a semantic space. In Ruslan Mitkov and Nicolas Nicolov, editors, *Recent Advances in Natural Language Processing: Selected Papers from RANLP'95*, volume 136 of *Amsterdam Studies in the Theory and History of Linguistic Science: Current Issues in Linguistic Theory*, chapter 2, pages 111–124. John Benjamins Publishing Company, Amsterdam / Philadelphia.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What categories reveal about the mind*. University of Chicago Press, Chicago.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 11, pages 265–283. The MIT Press, Cambridge, MA.

Joon Ho Lee, Myong Ho Kim, and Yoon Joon Lee. 1993. Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2):188–207, June.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL-99)*, pages 25–31.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings, Fifth International Conference on Systems Documentation (SIGDOC '86)*, pages 24–26.

Dekang Lin. 1998a. Automatic retreival 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 '98)*, August.

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

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, pages 280–287.

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

Saif Mohammad and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. *Submitted for publication*.

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, March.

Jane Morris and Graeme Hirst. 2004. 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*, pages 46–51, May.

Jane Morris and Graeme Hirst. 2005. The subjectivity of lexical cohesion in text. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing attitude and affect in text*, pages ???–??? Springer, New York.

Jane Morris. 2005. *Readers' Perceptions of Lexical Cohesion and Lexical Semantic Relations in Text*. Ph.D. thesis, University of Toronto.

M. Lynne Murphy. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, February.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity — measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pages 1024–1025.

Paul Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman.

Roy Rada and Ellen Bicknell. 1989. Ranking documents with a thesaurus. *JASIS*, 40(5):304–310, September.

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, February.

Philip Resnik and Mona Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 399–404.

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, Canada, August.

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

Barbara A. Spellman, Keith J. Holyoak, and Robert G. Morrison. 2001. Analogical priming via semantic relations. *Memory and Cognition*, 29(3):383–393.

David St-Onge. 1995. Detecting and correcting malapropisms with lexical chains. Master's thesis, University of Toronto, March. Published as Technical Report CSRI-319.

Mark Stevenson and Mark Greenwood. 2005. A semantic approach to IE pattern induction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 379–386, Ann Arbor, MI, June.

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, Virginia.

Michael John Sussna. 1997. *Text Retrieval Using Inference in Semantic Metanetworks*. Ph.D. thesis, University of California, San Diego.

Julie Weeds and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(?):???–???

Julie E. Weeds. 2003. *Measures and applications of lexical distributional similarity*. Ph.D. thesis, University of Sussex, September.

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

William B. Whitten, W. Newton Suter, and Michael L. Frank. 1979. Bidirectional synonym ratings of 464 noun pairs. *Journal of Learning and Verbal Behavior*, 18:109–127.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, June.