Journal of Artificial Intelligence Research x (1998) yy-zz Submitted x/98; published y/98

Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language

Philip Resnik RESNIK@UMIACS.UMD.EDU
Department of Linguistics and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742 USA

Abstract

This paper presents a measure of semantic similarity in an IS-A taxonomy based on the notion of shared information content. Experimental evaluation against a benchmark set of human similarity judgments demonstrates that the measure performs significantly better than the traditional edge counting approach. The paper presents algorithms that take advantage of taxonomic similarity in resolving syntactic and semantic ambiguity, along with experimental results demonstrating their effectiveness.

1. Introduction

Evaluating semantic relatedness using network representations is a problem with a long history in artificial intelligence and psychology, dating back to the spreading activation approach of Quillian (1968) and Collins and Loftus (1975). Semantic similarity represents a special case of semantic relatedness: for example, cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar. Rada et al. (Rada, Mili, Bicknell, & Blettner, 1989) suggest that the assessment of similarity in semantic networks can in fact be thought of as involving just taxonomic (IS-A) links, to the exclusion of other link types; that view will also be taken here, although admittedly it excludes some potentially useful information (Richardson, Smeaton, & Murphy, 1994).

A natural, time-honored way to evaluate semantic similarity in a taxonomy is to evaluate the distance between the nodes corresponding to the items being compared — the shorter the path from one node to another, the more similar they are. Given multiple paths, one takes the length of the shortest one (Lee, Kim, & Lee, 1993; Rada & Bicknell, 1989; Rada et al., 1989).

A widely acknowledged problem with this approach, however, is that it relies on the notion that links in the taxonomy represent uniform distances. Unfortunately, uniform link distance is difficult to define, much less to control. In real taxonomies, there is 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. For example, in WordNet (Miller, 1990), a widely used, broad-coverage semantic network for English, it is not at all difficult to find links that cover an intuitively narrow distance (RABBIT EARS IS-A TELEVISION ANTENNA) or an intuitively wide one (PHYTOPLANKTON IS-A LIVING THING). The same kinds of examples can be found in the Collins COBUILD Dictionary

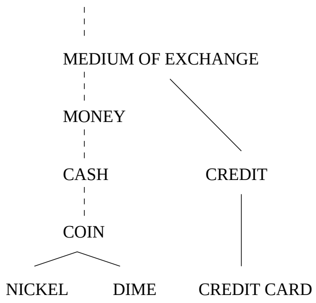

Figure 1: Fragment of the WordNet taxonomy. Solid lines represent IS-A links; dashed lines indicate that some intervening nodes were omitted to save space.

(Sinclair (ed.), 1987), which identifies superordinate terms for many words (e.g. SAFETY VALVE IS-A VALVE seems much narrower than KNITTING MACHINE IS-A MACHINE).

In the first part of this paper, I describe an alternative way to evaluate semantic similarity in a taxonomy, based on the notion of information content. Like the edge counting method, it is conceptually quite simple. However, it is not sensitive to the problem of varying link distances. In addition, by combining a taxonomic structure with empirical probability estimates, it provides a way of adapting a static knowledge structure to multiple contexts. Section 2 sets up the probabilistic framework and defines the measure of semantic similarity in information-theoretic terms, and Section 3 presents an evaluation of the similarity measure against human similarity judgments, using the simple edge-counting method as a baseline.

In the second part of the paper, Sections 4 and 5, I describe two applications of semantic similarity to problems of ambiguity in natural language. The first concerns a particular case of syntactic ambiguity that involves both coordination and nominal compounds, each of which is a pernicious source of structural ambiguity in English. Consider the phrase *food handling and storage procedures*: does it represent a conjunction of *food handling* and *storage procedures*, or does it refer to the *handling and storage* of food? The second application concerns the resolution of word sense ambiguity — not for words in running text, which is a large open problem (though cf. (Wilks & Stevenson, 1996)), but for groups of related words as are often discovered by distributional analysis of text corpora or found in dictionaries and thesauri. Finally, Section 6 discusses related work.

2. Similarity and Information Content

Let $\mathcal{C}$ be the set of concepts in an IS-A taxonomy, permitting multiple inheritance. Intuitively, one key to the similarity of two concepts is the extent to which they share information in common, indicated in an IS-A taxonomy by a highly specific concept that subsumes them both. The edge counting method captures this indirectly, since 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. For example, in WordNet, NICKEL and DIME are both subsumed by COIN, whereas the most specific superclass that NICKEL and CREDIT CARD share is MEDIUM OF EXCHANGE.[1] (See Figure 1.)

By associating probabilities with concepts in the taxonomy, it is possible to capture the same idea, but avoiding the unreliability of edge distances. Let the taxonomy be augmented with a function $\mathrm{p} : \mathcal{C} \rightarrow [0,1]$, such that for any $c \in \mathcal{C}$, $\mathrm{p}(c)$ is the probability of encountering an instance of concept c. This implies that p is monotonically nondecreasing as one moves up the taxonomy: if c_1 IS-A c_2, then $\mathrm{p}(c_1) \leq \mathrm{p}(c_2)$. Moreover, if the taxonomy has a unique top node then its probability is 1.

Following the standard argumentation of information theory (Ross, 1976), the *information content* of a concept c can be quantified as negative the log likelihood, $-\log \mathrm{p}(c)$. Notice that quantifying information content in this way makes intuitive sense in this setting: as probability increases, informativeness decreases, so the more abstract a concept, the lower its information content. Moreover, if there is a unique top concept, its information content is 0.

This quantitative characterization of information provides a new way to measure semantic similarity. The more information two concepts share in common, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them in the taxonomy. Formally, define

$$\mathrm{sim}(c_1, c_2) \quad = \quad \max_{c \,\in\, S(c_1,\, c_2)} \left[-\log \mathrm{p}(c)\right], \tag{1}$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2. A class that achieves the maximum value in Equation 1 will be termed a *most informative subsumer*; most often there is a unique most informative subsumer although this need not be true in the general case.

Notice that although similarity is computed by considering all upper bounds for the two concepts, the information measure has the effect of identifying minimal upper bounds, since no class is less informative than its superordinates. For example, in Figure 1, COIN, CASH, etc. are all members of S(NICKEL, DIME), but the concept that is structurally the minimal upper bound, COIN, will also be the most informative. This can make a difference in cases of multiple inheritance; for example, in Figure 2, METAL and CHEMICAL ELEMENT are not structurally distinguishable as upper bounds of NICKEL' and GOLD', but their information content may in fact be quite different.

In practice, one often needs to measure word similarity, rather than concept similarity. Using $s(w)$ to represent the set of concepts in the taxonomy that are senses of word w, define

$$\mathrm{sim}(w_1, w_2) \quad = \quad \max_{c1,\, c2} \left[\mathrm{sim}(c_1, c_2)\right], \tag{2}$$

where c_1 ranges over $s(w_1)$ and c_2 ranges over $s(w_2)$. This is consistent with Rada et al.'s (1989) treatment of "disjunctive concepts" using edge counting: they define the distance between two disjunctive sets of concepts as the minimum path length from any element of

1. In a feature-based setting (e.g. (Tversky, 1977)), this would be reflected by explicit shared features: nickels and dimes are both small, round, metallic, and so on. These features are captured implicitly by the taxonomy in categorizing NICKEL and DIME as subordinates of COIN.

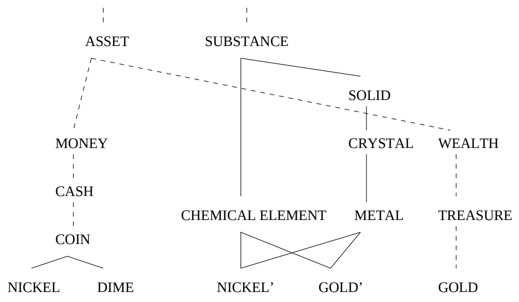

Figure 2: Another fragment of the WordNet taxonomy

the first set to any element of the second. Here, the word similarity is judged by taking the maximal information content over all concepts of which both words could be an instance. For example, Figure 2 illustrates how the similarity of words *nickel* and *gold* would be computed: the information content would be computed for all classes subsuming any pair in the cross product of {NICKEL,NICKEL'} and {GOLD,GOLD'}, and the information content of the most informative subsumer would be used to quantify the similarity of the two words.

3. Evaluation

3.1 Implementation

The work reported here used WordNet's taxonomy of concepts represented by nouns (and compound nominals) in English.[2] Frequencies of concepts in the taxonomy were estimated using noun frequencies from the Brown Corpus of American English (Francis & Kučera, 1982), a large (1,000,000 word) collection of text across genres ranging from news articles to science fiction. Each noun that occurred in the corpus was counted as an occurrence of each taxonomic class containing it.[3] For example, in Figure 1, an occurrence of the noun *dime* would be counted toward the frequency of DIME, COIN, and so forth. Formally,

$$\mathrm{freq}(c) \quad = \sum_{n \in \mathrm{words}(c)} \mathrm{count}(n), \tag{3}$$

where words(c) is the set of words subsumed by concept c. Concept probabilities were computed simply as relative frequency:

$$\hat{\mathrm{p}}(c) \quad = \quad \frac{\mathrm{freq}(c)}{N}, \tag{4}$$

where N was the total number of nouns observed (excluding those not subsumed by any WordNet class, of course). Naturally the frequency estimates in Equation 3 would be improved by taking into account the intended sense of each noun in the corpus — for example, an instance of *crane* can be a bird or a machine, but not both. Sense-tagged

2. *Concept* as used here refers to what Miller et al. (1990) call a *synset*, essentially a node in the taxonomy. The experiment reported in this section used the noun taxonomy from WordNet version 1.4, which has approximately 50,000 nodes.
3. Plural nouns counted as instances of their singular forms.

corpora are generally not available, however, and so the frequency estimates are done using this weaker but more generally applicable technique.

3.2 Task

Although there is no standard way to evaluate computational measures of semantic similarity, one reasonable way to judge would seem to be agreement with human similarity ratings. This can be assessed by using a computational similarity measure to rate the similarity of a set of word pairs, and looking at how well its ratings correlate with human ratings of the same pairs.

An experiment by Miller and Charles (1991) provided appropriate human subject data for the task. In their study, 38 undergraduate subjects were given 30 pairs of nouns that were chosen to cover high, intermediate, and low levels of similarity (as determined using a previous study (Rubenstein & Goodenough, 1965)), and those subjects were asked to rate "similarity of meaning" for each pair on a scale from 0 (no similarity) to 4 (perfect synonymy). The average rating for each pair thus represents a good estimate of how similar the two words are, according to human judgments.

In order to get a baseline for comparison, I replicated Miller and Charles's experiment, giving ten subjects the same 30 noun pairs. The subjects were all computer science graduate students or postdocs at the University of Pennsylvania, and the instructions were exactly the same as used by Miller and Charles, the main difference being that in this replication the subjects completed the questionnaire by electronic mail (though they were instructed to complete the whole thing in a single uninterrupted sitting). Five subjects received the list of word pairs in a random order, and the other five received the list in the reverse order. The correlation between the Miller and Charles mean ratings and the mean ratings in my replication was .96, quite close to the .97 correlation that Miller and Charles obtained between their results and the ratings determined by the earlier study.

For each subject in my replication, I computed how well his or her ratings correlated with the Miller and Charles ratings. The average correlation over the 10 subjects was $r = 0.88$, with a standard deviation of 0.08.[4] This value represents an upper bound on what one should expect from a computational attempt to perform the same task.

For purposes of evaluation, three computational similarity measures were used. The first is the similarity measurement using information content proposed in the previous section. The second is a variant on the edge counting method, converting it from distance to similarity by subtracting the path length from the maximum possible path length:

$$\mathrm{sim}_{\mathrm{edge}}(w_1, w_2) = (2 \times \textsc{max}) - \left[\min_{c_1, c_2} \mathrm{len}(c_1, c_2)\right] \tag{5}$$

where c_1 ranges over $s(w_1)$, c_2 ranges over $s(w_2)$, $\textsc{max}$ is the maximum depth of the taxonomy, and $\mathrm{len}(c_1, c_2)$ is the length of the shortest path from c_1 to c_2 . (Recall that $s(w)$ denotes the set of concepts in the taxonomy that represent senses of word w.) Note that the conversion from a distance to a similarity can be viewed as an expository convenience, and does not affect the evaluation: although the sign of the correlation coefficient changes

4. Inter-subject correlation in the replication, estimated using leaving-one-out resampling (Weiss & Kulikowski, 1991), was $r = .90, \mathrm{stdev} = 0.07$.

Similarity method	Correlation
Human judgments (replication)	$r = .9015$
Information content	$r = .7911$
Probability	$r = .6671$
Edge counting	$r = .6645$

Table 1: Summary of experimental results.

from positive to negative, its magnitude turns out to be just the same regardless of whether or not the minimum path length is subtracted from ($2 \times$ MAX).

The third point of comparison is a measure that simply uses the probability of a concept, rather than the information content:

$$\mathrm{sim}_{\mathrm{p}(c)}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [1 - \mathrm{p}(c)] \tag{6}$$

$$\mathrm{sim}_{\mathrm{p}(c)}(w_1, w_2) = \max_{c_1, c_2} \left[\mathrm{sim}_{\mathrm{p}(c)}(c_1, c_2)\right], \tag{7}$$

where c_1 ranges over $s(w_1)$ and c_2 ranges over $s(w_2)$ in (7). The probability-based similarity score is included in order to assess the extent to which similarity judgments might be sensitive to frequency *per se* rather than information content. Again, the difference between maximizing $1 - \mathrm{p}(c)$ and minimizing $\mathrm{p}(c)$ turns out not to affect the magnitude of the correlation. It simply ensures that the value can be interpreted as a similarity value, with high values indicating similar words.

3.3 Results

Table 1 summarizes the experimental results, giving the correlation between the similarity ratings and the mean ratings reported by Miller and Charles. Note that, owing to a noun missing from the WordNet 1.4 taxonomy, it was only possible to obtain computational similarity ratings for 28 of the 30 noun pairs; hence the proper point of comparison for human judgments is not the correlation over all 30 items ($r = .88$), but rather the correlation over the 28 included pairs ($r = .90$). The similarity ratings by item are given in Table 2.

3.4 Discussion

The experimental results in the previous section suggest that measuring semantic similarity using information content provides results that are significantly better than the traditional method of simply counting the number of intervening IS-A links.

The measure is not without its problems, however. One problem is that, like simple edge counting, the measure sometimes produces spuriously high similarity measures for words on the basis of inappropriate word senses. For example, Table 3 shows the word similarity for several words with *tobacco*. *Tobacco* and *alcohol* are similar, both being drugs, and *tobacco* and *sugar* are less similar, though not entirely dissimilar, since both can be classified as substances. The problem arises, however, in the similarity rating for *tobacco* with *horse*: the word *horse* can be used as a slang term for *heroin*, and as a result

Word Pair		Miller and Charles means	Replication means	sim	$\mathrm{sim}_{\mathrm{edge}}$	$\mathrm{sim}_{\mathrm{p}(c)}$
car	automobile	3.92	3.9	8.0411	30	0.9962
gem	jewel	3.84	3.5	14.9286	30	1.0000
journey	voyage	3.84	3.5	6.7537	29	0.9907
boy	lad	3.76	3.5	8.4240	29	0.9971
coast	shore	3.70	3.5	10.8076	29	0.9994
asylum	madhouse	3.61	3.6	15.6656	29	1.0000
magician	wizard	3.50	3.5	13.6656	30	0.9999
midday	noon	3.42	3.6	12.3925	30	0.9998
furnace	stove	3.11	2.6	1.7135	23	0.6951
food	fruit	3.08	2.1	5.0076	27	0.9689
bird	cock	3.05	2.2	9.3139	29	0.9984
bird	crane	2.97	2.1	9.3139	27	0.9984
tool	implement	2.95	3.4	6.0787	29	0.9852
brother	monk	2.82	2.4	2.9683	24	0.8722
crane	implement	1.68	0.3	2.9683	24	0.8722
lad	brother	1.66	1.2	2.9355	26	0.8693
journey	car	1.16	0.7	0.0000	0	0.0000
monk	oracle	1.10	0.8	2.9683	24	0.8722
food	rooster	0.89	1.1	1.0105	18	0.5036
coast	hill	0.87	0.7	6.2344	26	0.9867
forest	graveyard	0.84	0.6	0.0000	0	0.0000
monk	slave	0.55	0.7	2.9683	27	0.8722
coast	forest	0.42	0.6	0.0000	0	0.0000
lad	wizard	0.42	0.7	2.9683	26	0.8722
chord	smile	0.13	0.1	2.3544	20	0.8044
glass	magician	0.11	0.1	1.0105	22	0.5036
noon	string	0.08	0.0	0.0000	0	0.0000
rooster	voyage	0.08	0.0	0.0000	0	0.0000

Table 2: Semantic similarity by item.

n1	n2	sim(n1,n2)	class
tobacco	alcohol	7.63	DRUG
tobacco	sugar	3.56	SUBSTANCE
tobacco	horse	8.26	NARCOTIC

Table 3: Similarity with *tobacco* computed by maximizing information content

information-based similarity is maximized, and path length minimized, when the two words are both categorized as narcotics. This is contrary to intuition.

Cases like this are probably relatively rare. However, the example illustrates a more general concern: in measuring similarity between words, it is really the relationship among word *senses* that matters, and a similarity measure should be able to take this into account.

In the absence of a reliable algorithm for choosing the appropriate word senses, the most straightforward way to do so in the information-based setting is to consider *all* concepts to which both nouns belong rather than taking just the single maximally informative class. This suggests redefining similarity as follows:

$$\mathrm{sim}(c_1, c_2) \quad = \quad \sum_i \alpha(c_i)[-\log \mathrm{p}(c_i)], \tag{8}$$

where $\{c_i\}$ is the set of concepts dominating both c_1 and c_2, as before, and $\sum_i \alpha(c_i) = 1$. This measure of similarity takes more information into account than the previous one: rather than relying on the single concept with *maximum* information content, it allows *each* class to contribute information content according to the value of $\alpha(c_i)$. Intuitively, these α values measure relevance — for example, α(NARCOTIC) might be low in general usage but high in the context of a newspaper article about drug dealers. This idea is discussed further in Section 4.

4. Using Taxonomic Similarity in Resolving Syntactic Ambiguity

4.1 Clues for Resolving Coordination Ambiguity

Syntactic ambiguity is a pervasive problem in natural language. As Church and Patil (1982) point out, the class of "every way ambiguous" syntactic constructions — those for which the number of analyses is the number of binary trees over the terminal elements — includes such frequent constructions as prepositional phrases, coordination, and nominal compounds. In the last several years, researchers in natural language have made a great deal of progress in using quantitative information from text corpora to provide the needed constraints. Progress on broad-coverage prepositional phrase attachment ambiguity has been particularly notable, now that the dominant approach has shifted from structural strategies to quantitative analysis of lexical relationships (Whittemore, Ferrara, & Brunner, 1990; Hindle & Rooth, 1993; Brill & Resnik, 1994; Ratnaparkhi & Roukos, 1994; Li & Abe, 1995; Collins & Brooks, 1995; Merlo, Crocker, & Berthouzoz, 1997). Noun compounds have received comparatively less attention (Kobayasi, Takunaga, & Tanaka, 1994; Lauer, 1994, 1995), as has the problem of coordination ambiguity (Agarwal & Boggess, 1992; Kurohashi & Nagao, 1992).

In this section, I describe the application of semantic similarity to coordination ambiguities involving nominal compounds. In particular, I consider noun phrase coordinations of the form *noun1 and noun2 noun3*, as in (1), and noun phrase conjunctions of the form *noun1 noun2 and noun3 noun4*, as in (2).

(1) a. a (bank and warehouse) guard
    b. a (policeman) and (park guard)

(2) a. freshman ((business and marketing) major)

b. (food (handling and storage)) procedures
c. ((mail fraud) and bribery) charges
d. Clorets (gum and (breath mints))
e. (baby food) and (puppy chow)

Phrases like those in (1) admit two structural analyses, one in which *noun1* and *noun2* are the two noun phrase heads being conjoined (1a) and one in which the conjoined heads are *noun1* and *noun3* (1b). Adding an additional noun, there are five analyses, comprising two groups, those that conjoin heads *noun2* and *noun3* (a–c) and those that conjoin *noun2* and *noun4* (d–e). In this study I take the main problem to be identifying which two nouns are the conjoined heads.

Kurohashi and Nagao (Kurohashi & Nagao, 1992) point out that similarity of form and similarity of meaning are important cues to conjoinability. In English, similarity of form is to a great extent captured by agreement in number:

(3) a. several *business* and *university* groups
b. several *businesses* and university *groups*

Agreement of number between candidate conjoined heads can thus be thought of as a Boolean variable: either the candidate heads agree in number or they do not.

Similarity of meaning of the conjoined heads also appears to play an important role:

(4) a. a *television* and *radio* personality
b. a *psychologist* and sex *researcher*

Clearly *television* and *radio* are more similar than *television* and *personality*; correspondingly for *psychologist* and *researcher*. This similarity of meaning is captured well by semantic similarity in a taxonomy, and thus a second variable to consider when evaluating a coordination structure is the overlap in information content between the two head nouns.

In addition, for the constructions considered here, the appropriateness of noun-noun modification for *noun1* and *noun3* is relevant:

(5) a. *mail* and *securities* fraud
b. *corn* and peanut *butter*

One reason we prefer to conjoin *mail* and *securities* is that *mail fraud* is a salient compound nominal phrase. On the other hand, *corn butter* is not a familiar concept; compare to the change in perceived structure if the phrase were *corn and peanut crops*. In order to measure the appropriateness of noun-noun modification, I use a quantitative measure of selectional fit called *selectional association* (Resnik, 1996), which takes into account both lexical co-occurrence frequencies and semantic class membership in the WordNet taxonomy. Briefly, the selectional association of a word w with a WordNet class c is given by

$$A(w, c) \quad = \quad \frac{\mathrm{p}(c|w) \log \frac{\mathrm{p}(c|w)}{\mathrm{p}(c)}}{\mathrm{D}(\mathrm{p}(C|w) \parallel \mathrm{p}(C))} \tag{9}$$

where $\mathrm{D}(p_1 \parallel p_2)$ is the Kullback-Leibler distance (relative entropy) between probability distributions p_1 and p_2. Intuitively, A(w,c) is measuring the extent to which class

c is predicted by word w; for example, A(*wool*,CLOTHING) would have a higher value than, say, A(*wool*,PERSON). The selectional association $A(w_1, w_2)$ of two words is defined as the maximum of $A(w_1, c)$ taken over all classes c to which w_2 belongs. For example, A(*wool*,*glove*) would most likely be equal to A(*wool*,CLOTHING), as compared to, say, A(*wool*,SPORTS_EQUIPMENT) — the latter value corresponding to the sense of *glove* as something used in baseball, or in boxing.

4.2 Experiments in Resolving Coordination Ambiguity

I investigated the roles of these sources of information by conducting two straightforward disambiguation experiments using naturally occurring linguistic data. In the first experiment, two sets of 100 noun phrases of the form [NP *noun1 and noun2 noun3*] were extracted from the parsed *Wall Street Journal* (WSJ) corpus, as found in the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993). These were disambiguated by hand, with one set used for development and the other for testing.[5] A set of simple transformations were applied to all WSJ data, including the mapping of all proper names to the token *someone*, the expansion of month abbreviations, and the reduction of all nouns to their root forms.

Similarity of form, defined as agreement of number, was determined using a simple analysis of suffixes in combination with WordNet's database of nouns and noun exceptions.[6] Similarity of meaning was determined using the information-based measure of semantic similarity (Equation 2) — the noun class probabilities of Equation (1) were estimated using a sample of approximately 800,000 noun occurrences in Associated Press newswire stories.[7] For the purpose of determining semantic similarity, nouns not in WordNet were treated as instances of the class ⟨thing⟩. Co-occurrence frequencies used in measuring selectional association to assess appropriateness of noun-noun modification (Equation 9) were calculated using a sample of approximately 15,000 noun-noun compounds extracted from the WSJ corpus. (This sample did not include the test data.) Both selection of the modifier for the head and selection of the head for the modifier were considered.

Each of the three sources of information — form similarity, meaning similarity, and modification relationships — was used alone as a disambiguation strategy as described in Figure 3.[8]

In addition, I investigated several methods for combining the three sources of information. These included: (a) a simple form of "backing off" (i.e., given the form, modification, and meaning strategies in that order, use the first strategy that isn't undecided); (b) taking a "vote" among the three strategies and choosing the majority; (c) classifying using the results of a linear regression; and (d) constructing a decision tree classifier. The latter two methods are forms of supervised learning; in this experiment the development set was used as the training data.

5. Hand disambiguation was necessary because the Penn Treebank does not encode NP-internal structure. These phrases were disambiguated using the full sentence in which they occurred, plus the previous and following sentence, as context.
6. The experiments in this section used WordNet version 1.2.
7. I am grateful to Donald Hindle for making these data available.
8. Thresholds τ and σ were fixed before evaluating the test data.

- Form:
  - If noun1 and noun2 match in number
    and noun1 and noun3 do not
    then conjoin noun1 and noun2;
  - if noun1 and noun3 match in number
    and noun1 and noun2 do not
    then conjoin noun1 and noun3;
  - otherwise remain undecided.
- Meaning:
  - If sim(noun1,noun2) $>$ sim(noun1,noun3)
    then conjoin noun1 and noun2;
  - if sim(noun1,noun3) $>$ sim(noun1,noun2)
    then conjoin noun1 and noun3;
  - otherwise remain undecided.
- Modification:
  - If A(noun1,noun3) $> \tau$, a threshold, *or*
    if A(noun3,noun1) $> \tau$,
    then conjoin noun1 and noun3;
  - If A(noun1,noun3) $< \sigma$ *and* A(noun3,noun1) $< \sigma$
    then conjoin noun1 and noun2;
  - otherwise remain undecided.

Figure 3: Rules for using form, meaning, and modification in resolving syntactic ambiguity

Strategy	Coverage (%)	Accuracy (%)
Default	100.0	66.0
Form	53.0	90.6
Modification	75.0	69.3
Meaning	66.0	71.2
Backing off	95.0	81.1
Voting	89.0	78.7
Regression	100.0	79.0
ID3 Tree	100.0	80.0

Table 4: Syntactic disambiguation for items of the form *noun1 and noun2 noun3*

Strategy	Coverage (%)	Accuracy (%)
Default	100.0	44.9
Form	40.4	80.6
Meaning	69.7	77.4
Backing off	85.4	81.6

Table 5: Syntactic disambiguation for items of the form *noun1 noun2 and noun3 noun4*

The training set contained a bias in favor of conjoining *noun1* and *noun2*; therefore a "default" strategy, always choosing that bracketing, was used as a baseline for comparison. The results are shown in Table 4.

Not surprisingly, the individual strategies perform reasonably well on the instances they can classify, but coverage is poor; the strategy based on similarity of form is highly accurate, but arrives at an answer only half the time. Of the combined strategies, the "backing off" approach succeeds in answering 95% of the time and achieving 81.1% accuracy — a reduction of 44.4% in the baseline error rate.

In the second experiment, I used essentially the same approach in analyzing the more complex coordinations of the form [NP *noun1 noun2 and noun3 noun4*], in order to determine which nouns are the phrasal heads being conjoined. Three classification strategies were examined. The first was the form-based strategy described above. The second was a strategy based on semantic similarity; this time, however, selectional association was used to determine the α_i in Equation (8), incorporating modifier-head relationships into the semantic similarity strategy. The third strategy used "backing off" (from form similarity to semantic similarity) to combine the two individual strategies. As before, one set of items was used for development, and another set (89 items) was set aside for testing. As a baseline, results were evaluated against a simple default strategy of always choosing the group that was more common in the development set. The results are shown in Table 5.

In this case, the default strategy defined using the development set was misleading, leading to worse than chance accuracy. However, even if default choices were made using the bias found in the test set, accuracy would be only 55.1%. The results make it clear that the strategies using form and meaning are far more accurate, and that combining them leads to good coverage and accuracy.

The pattern of results in these two experiments demonstrates a significant reduction in syntactic misanalyses for this construction as compared to the simple baseline, and it confirms that form, meaning, and modification relationships all play a role in disambiguation. In addition, these results confirm the effectiveness of the proposed definitions of selectional association and semantic similarity.

5. Using Taxonomic Similarity in Word Sense Selection

5.1 Associating Word Senses with Noun Groupings

Knowledge about groups of related words plays a role in many natural language applications. As examples, query expansion using related words is a well studied technique in information retrieval (e.g., Harman, 1992; Grefenstette, 1992), clusters of similar words can play a role in smoothing stochastic language models for speech recognition (Brown, Della Pietra, deSouza, Lai, & Mercer, 1992), classes of verbs that share semantic structure form the basis for an approach to interlingual machine translation (Dorr, 1997), and clusterings of related words can be used in characterizing subgroupings of retrieved documents in large-scale Web searches (e.g. Digital Equipment Corporation, 1998). There is a wide body of research on the use of distributional methods for obtaining groups of related words (e.g., Bensch & Savitch, 1992; Brill, 1991; Brown et al., 1992; Grefenstette, 1992, 1994; McKeown & Hatzivassiloglou, 1993; Pereira, Tishby, & Lee, 1993; Schütze, 1993), and thesauri such as WordNet are another source of word relationships (e.g. Voorhees, 1994).

Distributional techniques can sometimes do a good job of identifying groups of related words (see Resnik, 1998b, for an overview and critical discussion), but for some tasks the relevant relationships are not among words, but among word *senses*. For example, Brown et al. (1992) illustrate the notion of a distributionally derived, "semantically sticky" cluster using an automatically derived word group containing *attorney*, *counsel*, *trial*, *court*, and *judge*. Although the semantic coherence of this cluster "pops out" for a human reader, a naive computational system has no defense against word sense ambiguity: using this cluster for query expansion could result in retrieving documents involving advice (one sense of *counsel*) and royalty (as one sense of *court*).[9]

Resnik (1998a) introduces an algorithm that uses taxonomically-defined semantic similarity in order to derive grouping relationships among word senses from grouping relationships among words. Formally, the problem can be stated as follows. Consider a set of words $W = \{w_1, \ldots, w_n\}$, with each word w_i having an associated set $S_i = \{s_{i,1}, \ldots, s_{i,m}\}$ of possible senses. Assume that there exists some set $W' \subseteq \bigcup S_i$, representing the set of word *senses* that an ideal human judge would conclude belong to the group of senses corresponding to the word grouping W. The goal is then to define a membership function φ that takes $s_{i,j}$, w_i, and W as its arguments and computes a value in $[0, 1]$, representing the confidence with which one can state that sense $s_{i,j}$ belongs in sense grouping W'. Note that,

9. See (Krovetz & Croft, 1992) and (Voorhees, 1993) for experimentation and discussion of the effects of word sense ambiguity in information retrieval.

in principle, nothing precludes the possibility that multiple senses of a word are included in W'.

For example, consider again the group

attorney, counsel, trial, court, judge.

Restricting attention to noun senses in WordNet, every word but *attorney* is polysemous. Treating this word group as W, a good algorithm for computing φ should assign a value of 1 to the unique sense of *attorney*, and it should assign a high value to the sense of *counsel* as

a lawyer who pleads cases in court.

Similarly, it should assign high values to the senses of *trial* as

legal proceedings consisting of the judicial examination of issues by a competent tribunal

the determination of a person's innocence or guilt by due process of law.

It should also assign high values to the senses of *court* as

an assembly to conduct judicial business

a room in which a law court sits.

And it should assign a high value to the sense of *judge* as

a public official authorized to decide questions brought before a court of justice.

It should assign low values of φ to the various word senses of words in this cluster that are associated with the group to a lesser extent or not at all. These would include the sense of *counsel* as

direction or advice as to a decision or course of action;

similarly, a low value of φ should be assigned to the sense of *court* as

a yard wholly or partly surrounded by walls or buildings.

The disambiguation algorithm for noun groups is given in Figure 4. Intuitively, when two polysemous words are similar, their most informative subsumer provides information about which sense of each word is the relevant one. This observation is similar in spirit to other approaches to word sense disambiguation based on maximizing relatedness of meaning (e.g., Lesk, 1986; Sussna, 1993). As an example, although WordNet lists *doctor* as meaning either a medical doctor or someone holding a Ph.D., and lists *nurse* as meaning either a health professional and a nanny, when the two words are considered together, the shared element of meaning for the two relevant senses emerges in the form of the most informative subsumer: HEALTH PROFESSIONAL.[10] The key idea behind the algorithm is to consider pairwise all combinations of words for the nouns in a word group, and all combinations

Algorithm. Given $W = \{w[1], \ldots, w[n]\}$, a set of nouns:

```
for i and j = 1 to n, with i ¡ j
{
   v[i, j] = sim(w[i], w[j])
   c[i, j] = the most informative subsumer for w[i] and w[j]

   for k = 1 to num_senses(w[i])
       if c[i, j] is an ancestor of sense[i, k]
           increment support[i, k] by v[i, j]

   for k' = 1 to num_senses(w[j])
       if c[i, j] is an ancestor of sense[j, k']
           increment support[j, k'] by v[i, j]

   increment normalization[i] by v[i, j]
   increment normalization[j] by v[i, j]
}

for i = 1 to n
   for k = 1 to num_senses(w[i])
   {
       if (normalization[i] > 0.0)
           phi[i, k] = support[i, k] / normalization[i]
       else
           phi[i, k] = 1 / num_senses(w[i])
   }
```

Figure 4: Disambiguation algorithm for noun groupings

of their word senses, and to assign "credit" to senses on the basis of shared information content, as measured using the information content of the most informative subsumer.

Resnik (1998a) illustrates the algorithm of Figure 4 using word groupings from a variety of sources, including several of the sources on distributional clustering cited above, and evaluates the algorithm more rigorously on the task of associating WordNet senses with nouns in Roget's thesaurus, based on their thesaurus category membership. On average, the algorithm achieved approximately 89% of the performance of human annotators performing the same task.[11] In the remainder of this section I describe a new application of the algorithm, and evaluate its performance.

5.2 Linking to WordNet using a Bilingual Dictionary

Multilingual resources for natural language processing can be difficult to obtain, although some promising efforts are underway in projects like EuroWordNet (Vossen, Diez-Orzas, & Peters, 1997). For many languages, however, such large-scale resources are unlikely to be available in the near future, and individual research efforts will have to continue to build from scratch or to adapt existing resources such as bilingual dictionaries (e.g., Klavans & Tzoukermann, 1995). In this section I describe an application of the algorithm of Figure 4 to the English definitions in a Chinese-English bilingual dictionary (MRM Incorporated, 199x). The ultimate task, being undertaken in the context of a Chinese-English machine translation project, will be to associate Chinese vocabulary items with nodes in WordNet, much in the same way that vocabulary in Spanish, Dutch, and Italian are associated with interlingual taxonomy nodes derived from the American WordNet, in the EuroWordNet project. The present study investigates the extent to which semantic similarity might be useful in partially automating the process.

For example, consider the following dictionary entries:

> Word 阿伯: 1. ⟨lit⟩ brother-in-law (husband's elder brother) 2. ⟨reg⟩ father 3. ⟨reg⟩ uncle (father's elder brother) 4. uncle (form of address to an older man)
>
> Word 唱旦的: actress, player of female roles

In order to associate Chinese terms such as these with the WordNet noun taxonomy, it is important to avoid associations with inappropriate senses — for example, the first word 阿伯 should clearly not be associated with *father* in its WordNet senses as Church Father, priest, God-the-Father, or founding father.

Although one traditional approach to using dictionary entries has been to compute word overlap with respect to dictionary definitions (e.g. Lesk, 1986), the English glosses in the Optilex dictionary are generally too short to take advantage of word overlap in this fashion.

10. The sense of nurse as a nanny is more common in British than American usage.
11. The task was performed independently by two human judges. Treating judge 1 as the benchmark the accuracies achieved by judge 2, the algorithm, and random selection were respectively 65.7%, 58.6%, and 34.8%; treating judge 2 as the benchmark the accuracies achieved by judge 1, the algorithm, and random selection were respectively 68.6%, 60.5%, and 33.3%. As the relatively low accuracies for human judges demonstrate, disambiguation using WordNet's fine-grained senses is quite a bit more difficult than disambiguation to the level of homographs (Hearst, 1991; Cowie, Guthrie, & Guthrie, 1992). Resnik and Yarowsky (1997) discuss the implications of WordNet's fine-grainedness for evaluation of word sense disambiguation, and consider alternative evaluation methods.

However, many of the definitions do have a useful property: they possess multiple subdefinitions that are similar in meaning, as in the cases illustrated above. Although one cannot always assume that this is so, e.g.,

Word 字盘: 1. case (i.e., upper case or lower case) 2. dial (of a watch, etc.),

inspection of the dictionary confirms that when multiple definitions are present they tend more toward polysemy than homonymy.

Based on this observation, I conducted an experiment to assess the extent to which the word sense disambiguation algorithm of Figure 4 can be used to identify relevant noun senses in WordNet for Chinese words in the Optilex dictionary, using the English definitions as the source of similar nouns to disambiguate. Nouns heading definitional noun phrases were extracted automatically via simple (but effective) heuristic methods, for a randomly-selected sample of 100 dictionary entries containing multiple definitions. For example, the noun groups associated with the definitions above would be

(阿伯) uncle brother-in-law father

(唱旦的) actress player

WordNet's noun database was used to automatically identify compound nominals where possible. So, for example, a word defined as "record player" would have the compound *record_player* rather than *player* as its head noun because *record_player* is a compound noun known to WordNet.[12]

Two independent judges were recruited for assistance in evaluating the algorithm, one a native Chinese speaker, and the second a Chinese language expert for the United States government. These judges independently evaluated the 100 test items. For each item, the judge was given the Chinese word, its full Optilex definition, and a list of all the WordNet sense descriptions associated with any head noun in its associated noun group. For example:

Word 急讯: urgent message, urgent dispatch

- message, content, subject_matter, substance: what a communication that is about something is about
- dispatch, expedition, expeditiousness, fastness: subconcept of celerity , quickness , rapidity
- dispatch, despatch, communique: an official report (usually sent in haste)
- message: a communication (usually brief) that is written or spoken or signaled; "he sent a three-word message"
- dispatch, despatch, shipment: the act of sending off something
- dispatch, despatch: the murder or execution of someone

For each item, the judge was first asked whether he knew that Chinese word in that meaning; if the response was negative, he was instructed to proceed to the next item. For items with known words, the judges were instructed as in Figure 5.

12. WordNet version 1.5 was used for this experiment.

For each WordNet definition, you will see 6 boxes: 1, 2, 3, 4, 5, and IS-A. For each definition:

- if you think the Chinese word *can have that meaning*, select the number corresponding to your confidence in that choice, where 1 is **lowest** confidence and 5 is **highest** confidence.
- If the Chinese word cannot have that meaning, but can have a *more specific* meaning, select IS-A. For example, if the Chinese word means "truck" and the WordNet definition is "automotive vehicle: self-propelled wheeled vehicle", you would select this option. (That is, it makes sense to say that this Chinese word describes a concept that IS A KIND OF AUTOMOTIVE VEHICLE.) **Then** pick 1, 2, 3, 4, or 5 as your confidence in this decision, again with 1 as **lowest** confidence and 5 as **highest** confidence.
- If neither of the above cases apply for this WordNet definition, don't check off anything for this definition.

Figure 5: Instructions for human judges selecting WordNet senses associated with Chinese words

Although the use of the IS-A selection was not used in the analysis of the results, it was important to include it because it provided the judges with a way to indicate where a Chinese word could best be classified in the WordNet noun taxonomy, without having to assert translational equivalence between the Chinese concept and a close WordNet (English) concept. So, for example, a judge could classify the word 春节 (the spring festival, lunar new year, Chinese new year) as belonging under the WordNet sense glossed as

festival: a day or period of time set aside for feasting and celebration,

the most sensible choice given that "Chinese New Year" does not appear as a WordNet concept. It is also important because the algorithm was working on groups of head nouns, thereby potentially losing information pointing to a more specific concept reading. For example, the definition

Word 钢管: steel tube, steel pipe

would be given to the algorithm as a group containing head nouns *tube* and *pipe*.

Evaluation was done according to two paradigms. In both we assume that we have a reference standard for each Chinese test item that correctly specifies, for that item, which WordNet senses are to be considered *correct*, and which are *incorrect*. An algorithm or judge being tested identifies, for each sense of each item, whether that sense should be *included* for that item or whether it should be *excluded*.

For the *selection paradigm*, the goal is to identify WordNet senses to *include*. We can therefore define precision in that paradigm as

$$P_{\text{selection}} = \frac{\text{number of correctly included senses}}{\text{number of included senses}}.$$

and recall as

$$R_{\text{selection}} = \frac{\text{number of correctly included senses}}{\text{number of correct senses}}.$$

These correspond directly to the use of precision and recall in information retrieval. Precision begins with the set of senses included by some method, and computes the proportion of these that are correct. Recall begins with the set of senses that *should* have been included, and computes the proportion of these that the method actually managed to choose.

Since the number of potential WordNet senses for an item can be quite large, an equally valid alternative to the selection paradigm is what I will call the *filtering* paradigm, according to which the goal is to identify WordNet senses to *exclude*. One can easily imagine this being the more relevant paradigm — for example, in a semi-automated setting where one wishes to reduce the burden of a user selecting among alternatives. In the filtering paradigm one can define filtering precision as

$$P_{\text{filtering}} = \frac{\text{number of correctly excluded senses}}{\text{number of excluded senses}}.$$

and filtering recall as

$$R_{\text{filtering}} = \frac{\text{number of correctly excluded senses}}{\text{number of senses labeled incorrect}}.$$

In the filtering paradigm, precision begins with the set of senses that the method filtered out, and computes the proportion that were *correctly* filtered out. And recall in filtering begins with the set of senses that *should* have been excluded (i.e. the incorrect ones), and computes the proportion of these that the method actually managed to exclude.[13]

Table 6 shows the precision/recall figures using the judgments of Judge 1, the native Chinese speaker, as a reference standard, considering only known items selected with confidence 3 and above.[14] (The algorithm recorded all 100 items as known, and its confidence values were scaled linearly from continuous values in range [0,1] to discrete values from 1 to 5.) As a lower bound for comparison, an algorithm was implemented that considered each word sense for each item, selecting that sense probabilistically in such a way as to make the average number of senses per item as close as possible to the average number of senses per item in the reference standard (1.3 senses). Figures for the random baseline are the average over 10 runs.

The fact that Judge 2 has such low precision and recall for selection indicates that matching the choices of an independent judge is indeed a difficult task. This is unsurprising, given previous experience with the problem of selecting among WordNet's fine-grained senses (Resnik, 1998a; Resnik & Yarowsky, 1997). Inspection of the patterns of disagreement suggests that in some cases, if one judge excluded a sense and the other included it, the excluded sense is close enough in meaning that it could reasonably be included. For example, for the item

13. Filtering recall corresponds to *fallout* as that term is used in the IR community.

14. Judge 1, the native speaker of Chinese, identified 65 of the words as known to him; Judge 2 identified 69. This on-line dictionary was constructed from a large variety of other lexical resources, and includes a great many uncommon words, archaic usages, regionalisms, and the like.

	Sense Selection		Sense Filtering	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Random	29.5	31.2	88.0	87.1
Algorithm	36.9	69.9	93.8	79.3
Judge 2	54.8	55.6	91.9	91.7

Table 6: Evaluation using Judge 1 as the reference standard, considering items selected with confidence 3 and above.

> Word 本文: 1. this text, this article, etc. 2. main text (of a book as contrasted to notes, index, etc.) 3. original text (as contrasted to translation)

Judge 1 selected the WordNet sense glossed as

> text, textual_matter: the words of something written

while Judge 2 selected the WordNet sense glossed as

> text, textbook, text_edition, schoolbook, school_text: a book prepared for use in schools or colleges.

Judge 2 comments, however, that the first sense also seems perfectly valid.

The results in Table 6 clearly show that the algorithm is better than the baseline, but also indicate that it is overgenerating senses, which hurts selection precision. In terms of filtering, when the algorithm chooses to filter out a sense it tends to do so reliably (filtering precision). However, its propensity toward overgeneration is reflected in its below-baseline performance on filtering recall; that is, the algorithm is choosing to allow in senses that it should be filtering out.

This pattern of results suggests that the best use of this algorithm at its present level of performance would be as a filter for a lexical acquisition process with a human in the loop, dividing candidate WordNet senses for dictionary entries according to higher and lower priority. For Chinese-English dictionary entries that serve as appropriate input to the algorithm (of which there are approximately 37000 in Optilex), if a WordNet sense is not selected by the algorithm with a confidence at least equal to 3 it should be demoted to the lower priority group in the presentation of alternatives, since the algorithm's choice to exclude a sense is correct approximately 93% of the time. Those senses that are selected by the algorithm are not *necessarily* to be included — the human judge is still needed to make the selection, since selection precision is low — but the algorithm tends to err on the side of caution and so correct senses will be found in the higher priority group some 70% of the time.

6. Related Work

Although the counting of edges in IS-A taxonomies seems to be something many people have tried, there seem to be few published descriptions of attempts to directly evaluate the effectiveness of this method. A number of researchers have attempted to make use of conceptual distance in information retrieval. For example, Rada et al. (1989, 1989) and Lee et al. (1993) report experiments using conceptual distance, implemented using the edge counting metric, as the basis for ranking documents by their similarity to a query. Sussna (1993) uses semantic relatedness measured with WordNet in word sense disambiguation, defining a measure of distance that weights different types of links and also explicitly takes depth in the taxonomy into account.

Following the original proposal to measure semantic similarity using information content (Resnik, 1993b, 1993a), a number of related proposals have been explored. Leacock and Chodorow (Leacock & Chodorow, 1994) define a measure resembling information content, but using the normalized path length between the two concepts being compared rather than the probability of a subsuming concept. Specifically, they define

$$\text{sim}_{\text{ndist}}(w_1, w_2) \quad = \quad -\log \left[\frac{\min\limits_{c_1, c_2} \text{len}(c_1, c_2)}{(2 \times \text{MAX})} \right]. \tag{10}$$

(The notation above is the same as for Equation (5).) In addition to this definition, they also include several special cases, most notably to avoid infinite similarity when c_1 and c_2 are exact synonyms and thus have a path length of 0. Leacock and Chodorow have experimented with this measure and the information content measure described here in the context of word sense disambiguation, and found that they yield roughly similar results. Implementing their method and testing it on the task reported in Section 3, I found that it actually outperformed the information-based measure slightly on that data set; however, in a followup experiment using a different and larger set of noun pairs (100 items), the information-based measure performed significantly better (Table 7).

Analyzing the differences between the two studies is illuminating. In the followup experiment, I used netnews archives to gather highly frequent nouns within related topic areas (to ensure that similar noun pairs occurred) and then selected noun pairings at random (in order to avoid biasing the followup study in favor of either algorithm). There is, therefore, a predominance of low-similarity noun pairs in the test data. Looking at the distribution of ratings for the noun pairs, as given by the two measures, it is evident that the Leacock and Chodorow measure is overestimating semantic similarity for many of the predominantly non-similar pairs. This stands to reason since the measure is identical whenever the edge distance is identical, regardless of whether the pair is high or low in the taxonomy (e.g. the distance between *plant* and *animal* is the same as the distance between *white oak* and *red oak*). In contrast, the information-based measure is sensitive to the difference, and better at avoiding spuriously high similarity values for non-similar pairs. This comparison between the followup study and the original Miller and Charles data illustrates quite clearly how the utility of a similarity measure can depend upon the distribution of items given by the task.

Lin (1997a, 1997b) has recently proposed an alternative information-theoretic similarity measure, derived from a set of basic assumptions about similarity in a style reminiscent of the way in which entropy/information itself has a formal definition derivable from a set of

Similarity method	Correlation
Information content	$r = .6894$
Leacock and Chodorow	$r = .4320$
Edge counting	$r = .4101$

Table 7: Summary of experimental results in followup study.

basic properties (Khinchin, 1957). Formally, Lin defines similarity in a taxonomy as:

$$\mathrm{sim}_{\mathtt{Lin}}(c_1, c_2) \quad = \quad \frac{2 \times \log \mathrm{p}(\bigcap_i C_i)}{\log \mathrm{p}(c_1) + \log \mathrm{p}(c_2)} \tag{11}$$

where the C_i are the "maximally specific superclasses" of both c_1 and c_2. Although the possibility of multiple inheritance makes the intersection $\bigcap_i C_i$ necessary in principle, multiple inheritance is in fact so rare in WordNet that in practice one computes Equation 11 separately for each common ancestor C_i, using $\mathrm{p}(C_i)$ in the numerator, and then takes the maximum (Dekang Lin, p.c.). Other than the multiplicative constant of 2, therefore, Lin's method for determining similarity in a taxonomy is essentially the information-based similarity measure of Equation 1, but *normalized* by the combined information content of the two concepts assuming their independence. Put another way, Lin's measure is taking into account not only commonalities but differences between the items being compared, expressing both in information-theoretic terms.

Lin's measure is theoretically well motivated and elegantly derived. Moreover, Lin points out that his measure will by definition yield the same value for $\mathrm{sim}_{\mathtt{Lin}}(x, x)$ regardless of the identity of x — unlike information content, which has been criticized on the grounds that the value of self-similarity depends on how specific a concept x is, and that two non-identical items x and y can be rated more similar to each other than a third item z is to itself (Richardson et al., 1994). Interestingly, however, similarity comparisons involving self-similarity ("Robins are similar to robins"), as well as subclass relationships ("Robins are similar to birds"), have themselves been criticized by psychologists as anomalous (Medin, Goldstone, & Gentner, 1993), and Tversky (1977) observes that minimality, i.e. the assumption that $\mathrm{sim}(x, y) \geq sim(x, x) = 0$, fails to hold when the behavioral measure is the probability of judging two identical stimuli as "same" rather than "different." "Moreover," Tversky points out, "in recognition experiments the off-diagonal entries often exceed the diagonal entries; that is, an object is identified as another object more frequently than it is identified as itself. If identification probability is interpreted as a measure of similarity, then these observations violate minimality..." (p. 328). It would appear that minimality, although intuitively appealing, may not be a necessary property of a similarity measure faithful to human judgments; since Lin's measure ensures minimality and the information content measure does not, further experimentation is clearly needed to evaluate their competing predictions.

Wu and Palmer (1994) propose a similarity measure that is based on edge distances, but related to Lin's measure in the way it takes into account the most specific node dominating c_1 and c_2, characterizing their commonalities, while normalizing in a way that accounts for

Similarity method	Correlation
Information content	$r = .7947$
$\text{sim}_{\texttt{Wu\&Palmer}}$	$r = .8027$
$\text{sim}_{\texttt{Lin}}$	$r = .8339$

Table 8: Summary of Lin's results comparing alternative similarity measures

their differences. Revising Wu and Palmer's notation slightly, their measure is:

$$\text{sim}_{\texttt{Wu\&Palmer}}(c_1, c_2) = \frac{2 \times \text{d}(c_3)}{\text{d}(c_1) + \text{d}(c_2)} \tag{12}$$

where c_3 is the maximally specific superclass of c_1 and c_2, $\text{d}(c_3)$ is its depth, i.e. distance from the root of the taxonomy, and $\text{d}(c_1)$ and $\text{d}(c_2)$ are the depths of c_1 and c_2 on the path through c_3.

Lin (1997a) repeats the experiment of Section 3 for the information content measure, $\text{sim}_{\texttt{Lin}}$, and $\text{sim}_{\texttt{Wu\&Palmer}}$, reporting the results that appear in Table 8. Lin uses a sense-tagged corpus to estimate frequencies, and smoothed probabilities rather than simple relative frequency. His results show a somewhat higher correlation for $\text{sim}_{\texttt{Lin}}$ than the other measures. Further experimentation is needed in order to assess the alternative measures, particularly with respect to their competing predictions. What seems clear, however, is that all these measures perform far better than the edge-counting measure, when evaluated using a taxonomy consisting only of IS-A links and a dataset that maintains a balance among high-, medium-, and low-similarity test items.

7. Conclusions

This paper has presented a measure of semantic similarity in an IS-A taxonomy, based on the notion of information content. Experimental evaluation was performed using a large, independently constructed corpus, an independently constructed taxonomy, and previously existing human subject data, and the results suggest that the measure performs encouragingly well, and significantly better than the traditional edge-counting approach. Semantic similarity, as measured using information content, was shown to be useful in resolving two forms of linguistic ambiguity, one syntactic and one semantic.

Later work on semantic similarity in taxonomies has improved on the information content measure when evaluated against the same human subject data, but also raises questions about the underlying assumptions one should make when attempting to faithfully model human judgments, as well as the kinds of data sets for which one measure or another is better adapted. Moreover, although the IS-A relationship underlies most knowledge bases for which semantic similarity is of interest, many also include other kinds of relational links not taken into account by the similarity measures discussed here, which might be of use (e.g., Sussna, 1993; Richardson et al., 1994).

Acknowledgements

This paper is a revised and significantly extended version of (Resnik, 1995). Section 4 includes material from (Resnik, 1993a), which also was presented in (Resnik, 1993b). Parts of this research were done at the University of Pennsylvania with the partial support of an IBM Graduate Fellowship and grants ARO DAAL 03-89-C-0031, DARPA N00014-90-J-1863, NSF IRI 90-16592, and Ben Franklin 91S.3078C-1; parts of this research were also done at Sun Microsystems Laboratories in Chelmsford, Massachusetts; and parts of this work were supported at the University of Maryland by Department of Defense contract MDA90496C1250, DARPA/ITO Contract N66001-97-C-8540, Army Research Laboratory contract DAAL03-91-C-0034 through Battelle, and a research grant from Sun Microsystems Laboratories. The author gratefully acknowledges helpful discussions with Claudia Leacock and Dekang Lin, as well as comments and criticism received during various presentations of this work.

References

Agarwal, R., & Boggess, L. (1992). A simple but useful approach to conjunct identification. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 15–21. Association for Computational Linguistics.

Bensch, P. A., & Savitch, W. J. (1992). An occurrence-based model of word categorization. Presented at 3rd Meeting on Mathematics of Language (MOL3).

Brill, E. (1991). Discovering the lexical features of a language. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA*.

Brill, E., & Resnik, P. (1994). A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*.

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–480.

Church, K. W., & Patil, R. (1982). Coping with syntactic ambiguity or how to put the block in the box on the table. *American Journal of Computational Linguistics*, *8*(3-4), 139–149.

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Collins, M., & Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. In *Third Workshop on Very Large Corpora*. Association for Computational Linguistics. cmp-lg/9506021.

Cowie, J., Guthrie, J., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pp. 359–365 Nantes, France.

Digital Equipment Corporation (1998). AltaVista web page: Refine, or cow9?.. `http://altavista.digital.com/av/content/about_our_technology_cow9.htm`.

Dorr, B. J. (1997). Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, *12*(4), 271–322.

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 89–97.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer.

Harman, D. (1992). Relevance feedback revisited. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1–10.

Hearst, M. (1991). Noun homograph disambiguation using local context in large corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research* Oxford.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, *19*(1), 103–120.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. New York: Dover Publications. Translated by R. A. Silverman and M. D. Friedman.

Klavans, J. L., & Tzoukermann, E. (1995). Dictionaries and Corpora: Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, *10*, 185–218.

Kobayasi, Y., Takunaga, T., & Tanaka, H. (1994). Analysis of Japanese compound nouns using collocational information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*.

Krovetz, R., & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, *10*(2), 115–141.

Kurohashi, S., & Nagao, M. (1992). Dynamic programming method for analyzing conjunctive structures in Japanese. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)* Nantes, France.

Lauer, M. (1994). Conceptual association for compound noun analysis. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* Las Cruces, New Mexico. Student Session.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University, Sydney, Australia.

Leacock, C., & Chodorow, M. (1994). Filling in a sparse training space for word sense identification. ms.

Lee, J. H., Kim, M. H., & Lee, Y. J. (1993). Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, *49*(2), 188–207.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pp. 24–26.

Li, H., & Abe, N. (1995). Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of the International Conference on Recent Advances in NLP* Velingrad, Bulgaria.

Lin, D. (1997a). An information-theoretic definition of similarity.. submitted.

Lin, D. (1997b). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, *19*, 313–330.

McKeown, K., & Hatzivassiloglou, V. (1993). Augmenting lexicons automatically: Clustering semantically related adjectives. In Bates, M. (Ed.), *ARPA Workshop on Human Language Technology*. Morgan Kaufmann.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278.

Merlo, P., Crocker, M., & Berthouzoz, C. (1997). Attaching multiple prepositional phrases: Generalized backed-off estimation. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*. cmp-lg/9710005.

Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4). (Special Issue).

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

MRM Incorporated (Ed.). (199x). *Optilex*.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)* Morristown, New Jersey. Association for Computational Linguistics.

Quillian, M. R. (1968). Semantic memory. In Minsky, M. (Ed.), *Semantic Information Processing*. MIT Press, Cambridge, MA.

Rada, R., & Bicknell, E. (1989). Ranking documents with a thesaurus. *JASIS*, *40*(5), 304–310.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, *19*(1), 17–30.

Ratnaparkhi, A., & Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceddings of the ARPA Workshop on Human Language Technology* Plainsboro, NJ.

Resnik, P. (1993a). *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania. (`ftp://ftp.cis.upenn.edu/pub/ircs/tr/93-42.ps.Z`).

Resnik, P. (1993b). Semantic classes and syntactic ambiguity. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*. Morgan Kaufmann.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*. (cmp-lg/9511007).

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*, 127–159.

Resnik, P. (1998a). Disambiguating noun groupings with respect to Wordnet senses. In Armstrong, S., Church, K., Isabelle, P., Tzoukermann, E., & Yarowsky, D. (Eds.), *Natural Language Processing Using Very Large Corpora*. Kluwer.

Resnik, P. (1998b). WordNet and class-based probabilities. In Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database*. MIT Press.

Resnik, P., & Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *ANLP Workshop on Tagging Text with Lexical Semantics* Washington, D.C.

Richardson, R., Smeaton, A. F., & Murphy, J. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. Working paper CA-1294, Dublin City University, School of Computer Applications, Dublin, Ireland. `ftp://ftp.compapp.dcu.ie/pub/w-papers/1994/CA1294.ps.Z`.

Ross, S. (1976). *A First Course in Probability*. Macmillan.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *CACM*, *8*(10), 627–633.

Schütze, H. (1993). Word space. In Hanson, S. J., Cowan, J. D., & Giles, C. L. (Eds.), *Advances in Neural Information Processing Systems 5*, pp. 895–902. Morgan Kaufmann Publishers, San Mateo CA.

Sinclair (ed.), J. (1987). *Collins COBUILD English Language Dictionary.* Collins: London.

Sussna, M. (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)* Arlington, Virginia.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In Korfhage, R., Rasmussen, E., & Willett, P. (Eds.), *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171–180.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *17th International Conference on Research and Development in Information Retrieval (SIGIR '94)* Dublin, Ireland.

Vossen, P., Diez-Orzas, P., & Peters, W. (1997). The multilingual design of EuroWordNet. In *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* Madrid.

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* Morgan Kaufmann, San Mateo, CA.

Whittemore, G., Ferrara, K., & Brunner, H. (1990). Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 23–30. Pittsburgh, Pennsylvania.

Wilks, Y., & Stevenson, M. (1996). The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?.. Technical Report CS-96-05, cmp-lg/9607028.

Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* Las Cruces, New Mexico.