Using Corpus Statistics and WordNet Relations for Sense Identification

Claudia Leacock[*]
Educational Testing Service

Martin Chodorow[†]
Hunter College of CUNY

George A. Miller[‡]
Princeton University

Corpus-based approaches to word sense identification have flexibility and generality but suffer from a knowledge acquisition bottleneck. We show how knowledge-based techniques can be used to open the bottleneck by automatically locating training corpora. We describe a statistical classifier that combines topical context with local cues to identify a word sense. The classifier is used to disambiguate a noun, a verb, and an adjective. A knowledge base in the form of WordNet's lexical relations is used to automatically locate training examples in a general text corpus. Test results are compared with those from manually tagged training examples.

1. Introduction

An impressive array of statistical methods have been developed for word sense identification. They range from dictionary-based approaches that rely on definitions (Véronis and Ide 1990; Wilks et al. 1993) to corpus-based approaches that use only word co-occurrence frequencies extracted from large textual corpora (Schütze 1995; Dagan and Itai 1994). We have drawn on these two traditions, using corpus-based co-occurrence and the lexical knowledge base that is embodied in the WordNet lexicon.

The two traditions complement each other. Corpus-based approaches have the advantage of being generally applicable to new texts, domains, and corpora without needing costly and perhaps error-prone parsing or semantic analysis. They require only training corpora in which the sense distinctions have been marked, but therein lies their weakness. Obtaining training materials for statistical methods is costly and time-consuming—it is a "knowledge acquisition bottleneck" (Gale, Church, and Yarowsky 1992a). To open this bottleneck, we use WordNet's lexical relations to locate unsupervised training examples.

Section 2 describes a statistical classifier, TLC (Topical/Local Classifier), that uses topical context (the open-class words that co-occur with a particular sense), local context (the open- and closed-class items that occur within a small window around a word), or a combination of the two. The results of combining the two types of context to disambiguate a noun (*line*), a verb (*serve*), and an adjective (*hard*) are presented. The following questions are discussed: When is topical context superior to local context (and vice versa)? Is their combination superior to either type alone? Do the answers to these questions depend on the size of the training? Do they depend on the syntactic category of the target?

[*] Division of Cognitive and Instructional Science, Princeton, NJ 08541; e-mail: cleacock@ets.org. The work reported here was done while the author was at Princeton University.

[†] Department of Psychology, 695 Park Avenue, New York, NY 10021; e-mail: mschc@cunyvm.cuny.edu

[‡] Cognitive Science Laboratory, 221 Nassau Street, Princeton, NJ 08542; e-mail: geo@clarity.princeton.edu

Manually tagged training materials were used in the development of TLC and the experiments in Section 2. The Cognitive Science Laboratory at Princeton University, with support from NSF-ARPA, is producing textual corpora that can be used in developing and evaluating automatic methods for disambiguation. Examples of the different meanings of one thousand common, polysemous, open-class English words are being manually tagged. The results of this effort will be a useful resource for training statistical classifiers, but what about the next thousand polysemous words, and the next? In order to identify senses of these words, it will be necessary to learn how to harvest training examples automatically.

Section 3 describes WordNet's lexical relations and the role that monosemous "relatives" of polysemous words can play in creating unsupervised training materials. TLC is trained with automatically extracted examples, its performance is compared with that obtained from manually tagged training materials.

2. Corpus-based Statistical Sense Identification

Work on automatic sense identification from the 1950s onward has been well summarized by Hirst (1987) and Dagan and Itai (1994). The discussion below is limited to work that is closely related to our research.

2.1 Some Recent Work

Hearst (1991) represents local context with a shallow syntactic parse in which the context is segmented into prepositional phrases, noun phrases, and verb groups. The target noun is coded for the word it modifies, the word that modifies it, and the prepositions that precede and follow it. Open-class items within ±3 phrase segments of the target are coded in terms of their relation to the target (modifier or head) or their role in a construct that is adjacent to the target. Evidence is combined in a manner similar to that used by the local classifier component of TLC. With supervised training of up to 70 sentences per sense, performance on three homographs was quite good (88–100% correct); with fewer training examples and semantically related senses, performance on two additional words was less satisfactory (73–77% correct).

Gale, Church, and Yarowsky (1992a) developed a topical classifier based on Bayesian decision theory. The only information the classifier uses is an unordered list of words that co-occur with the target in training examples. No other cues, such as part-of-speech tags or word order, are used. Leacock, Towell, and Voorhees (1993) compared this Bayesian classifier with a content vector classifier as used in information retrieval and a neural network with backpropagation. The classifiers were compared using different numbers of senses (two, three, or six manually tagged senses of *line*) and different amounts of training material (50, 100, and 200 examples). On the six-sense task, the classifiers averaged 74% correct answers. Leacock, Towell, and Voorhees (1993) found that the response patterns of the three classifiers converged, suggesting that each of the classifiers was extracting as much data as is available in purely topical approaches that look only at word counts from training examples. If this is the case, any technique that uses only topical information will not be significantly more accurate than the three classifiers tested.

Leacock, Towell, and Voorhees (1996) showed that performance of the content vector topical classifier could be improved with the addition of local templates—specific word patterns that were recognized as being indicative of a particular sense—in an extension of an idea initially suggested by Weiss (1973). Although the templates proved to be highly reliable when they occurred, all too often, none were found.

Yarowsky (1993) also found that template-like structures are very powerful indicators of sense. He located collocations by looking at adjacent words or at the first word to the left or right in a given part of speech and found that, with binary ambiguity, a word has only one sense in a given collocation with a probability of 90–99%.[1] However, he had an average of only 29% recall (i.e., the collocations were found in only 29% of the cases). When local information occurred it was highly reliable, but all too often, it did not occur.

Bruce and Wiebe (1994a, 1994b) have developed a classifier that represents local context by morphology (the inflection on the target word), the syntactic category of words within a window of ±2 words from the target, and collocation-specific items found in the sentence. The collocation-specific items are those determined to be the most informative, where an item is considered informative if the model for independence between it and a sense tag provided a poor fit to the training data. The relative probabilities of senses, available from the training corpus, are used in the decision process as prior probabilities. For each test example, the evidence in its local context is combined in a Bayesian-type model of the probability of each sense, and the most probable sense is selected. Performance ranges from 77–84% correct on the test words, where a lower bound for performance based on always selecting the most frequent sense for the same words (i.e., the sense with the greatest prior probability) would yield 53–80% correct.

Yarowsky (1994), building on his earlier work, designed a classifier that looks at words within $\pm k$ positions from the target; lemma forms are obtained through morphological analysis; and a coarse part-of-speech assignment is performed by dictionary lookup. Context is represented by collocations based on words or parts of speech at specific positions within the window or, less specifically, in any position. Also coded are some special classes of words, such as WEEKDAY, that might serve to distinguish among word senses. For each type of local-context evidence found in the corpus, a log-likelihood ratio is constructed, indicating the strength of the evidence for one form of the homograph versus the other. These ratios are then arranged in a sorted decision list with the largest values (strongest evidence) first. A decision is made for a test sentence by scanning down the decision list until a match is found. Thus, only the single best piece of evidence is used. The classifier was tested on disambiguating the homographs that result from accent removal in Spanish and French (e.g., *seria*, *sería*). In tests with the number of training examples ranging from a few hundred to several thousand, overall accuracy was high, above 90%.

Clearly, sense identification is an active area of research, and considerable ingenuity is apparent. But despite the promising results reported in this literature, the reality is that there still are no large-scale, operational systems for tagging the senses of words in text.

2.2 Topical/Local Classifier (TLC)

The statistical classifier, TLC, uses topical context, local context, or a combination of the two, for word sense identification. TLC's flexibility in using both forms is an important asset for our investigations.

A noun, a verb, and an adjective were tested in this study. Table 1 provides a synonym or brief gloss for each of the senses used. Training corpora and testing corpora were collected as follows:

1. Examples for *serve*, *hard*, and *line*, in base or inflected form, were located in on-line corpora. Examples containing *line* and *serve* were taken from the 1987–89 LDC

[1] Yarowsky does not use the idiomatic or noncompositional sense of collocation. Instead, he means co-occurrence of any words.

Table 1
Word senses used in the experiment and their relative frequencies.

serve (verb)		hard (adj)		line (noun)	
supply with food	.41	not easy (difficult)	.80	product	.54
hold an office	.29	not soft (metaphoric)	.12	phone	.10
function as something	.20	not soft (physical)	.08	text	.10
provide a service	.10			cord	.09
				division	.09
				formation	.08

Wall Street Journal corpus and from the American Printing House for the Blind corpus.[2] Examples for *hard* were taken from the LDC *San Jose Mercury News* (SJM) corpus. Each consisted of the sentence containing the target and one sentence preceding it. The resulting strings had an average length of 49 items.

2. Examples where the target was the head of an unambiguous collocation were removed from the files. Being unambiguous, they do not need to be disambiguated. These collocations, for example, *product line* and *hard candy* were found using WordNet. In Section 3, we consider how they can be used for unsupervised training. Examples where the target was part of a proper noun were also removed; for example, *Japan Air Lines* was not taken as an example of *line*.

3. Each occurrence of the target word was manually tagged with a WordNet sense until a large number of examples was obtained for six senses of *line*, four senses of *serve*, and three senses of *hard*. In the process of collecting and manually tagging examples, it was possible to determine the relative frequencies of the senses of each word. The less frequent senses, which do not appear in Table 1, occurred too rarely for us to collect the minimum number of examples needed to perform the experiment described in the next section.

4. Three sets of materials were prepared by partitioning the examples for each sense into training and test sets.[3] The size of the training set was varied by taking the first 25, 50, 100, and 200 examples of the least frequent sense, and examples from the other senses in numbers that reflected their relative frequencies in the corpus. As an illustration, in the smallest training set for *hard*, there were 25 examples of the least frequent sense, 37 examples of the second most frequent sense, and 256 examples of the most frequent sense. The test sets were of fixed size: each contained 150 of the least frequent sense and examples of the other senses in numbers that reflected their relative frequencies.

The operation of TLC consists of preprocessing, training, and testing. During preprocessing, examples are tagged with a part-of-speech tagger (Brill 1994); special tags are inserted at sentence breaks; and each open-class word found in WordNet is replaced with its base form. This step normalizes across morphological variants without

[2] This 25-million-word corpus is archived at IBM's T. J. Watson Research Center; it consists of stories and articles from books and general circulation magazines.

[3] When the examples were collected from the corpus, it was often the case that more than one was extracted from a given document. To prevent the classifier from being trained and then tested on sentences from the same text, care was taken to insure that the training and test materials were separate. Much of the corpus consisted of newspapers and periodicals, where it is common practice to repeat or paraphrase the same story on successive days. To minimize possible overlap due to repeated stories, the temporal order of the documents was preserved, and the test set was selected in such a way that it was not contiguous with the training materials.

resorting to the more drastic measure of stemming. Morphological information is not lost, since the part-of-speech tag remains unchanged.

Training consists of counting the frequencies of various contextual cues for each sense. Testing consists of taking a new example of the polysemous word and computing the most probable sense, based on the cues present in the context of the new item. A comparison is made to the sense assigned by a human judge, and the classifier's decision is scored as correct or incorrect.

TLC uses a Bayesian approach to find the sense s_i that is the most probable given the cues c_j contained in a context window of $\pm k$ positions around the polysemous target word. For each s_i, the probability is computed with Bayes' rule:

$$p(s_i \mid c_{-k}, \ldots, c_k) = \frac{p(c_{-k}, \ldots, c_k \mid s_i)p(s_i)}{p(c_{-k}, \ldots, c_k)}$$

As Golding (1995) points out, the term $p(c_{-k}, \ldots, c_k \mid s_i)$ is difficult to estimate because of the sparse data problem, but if we assume, as is often done, that the occurrence of each cue is independent of the others, then this term can be replaced with:

$$p(c_{-k}, \ldots, c_k \mid s_i) = \prod_{j=-k}^{k} p(c_j \mid s_i)$$

In TLC, we have made this assumption and have estimated $p(c_j \mid s_i)$ from the training. Of course, the sparse data problem affects these probabilities too, and so TLC uses the Good-Turing formula (Good 1953; Chiang, Lin, and Su 1995), to smooth the values of $p(c_j \mid s_i)$, including providing probabilities for cues that did not occur in the training.

TLC actually uses the mean of the Good-Turing value and the training-derived value for $p(c_j \mid s_i)$. When cues do not appear in training, it uses the mean of the Good-Turing value and the global probability of the cue $p(c_j)$, obtained from a large text corpus. This approach to smoothing has yielded consistently better performance than relying on the Good-Turing values alone.

TLC uses: (1) topical cues consisting of open-class words found in a wide window that includes the sentence in which the target is located plus the preceding sentence; (2) local open-class words found in a narrow window around the target; (3) local closed-class items; (4) local part-of-speech tags. The procedures for estimating $p(c_j \mid s_i)$ and $p(c_j)$ differ somewhat for the various types of cue.

1. The counts for open-class words (nouns, verbs, adjectives, and adverbs) from which the topical cue probabilities $p(c_j \mid s_i)$ and $p(c_j)$ are calculated are not sensitive to position within the wide window (the "bag-of-words" method). By contrast, the local cue probabilities do take into account position relative to the target.

2. For open-class words found in the three positions to the left of the target (i.e., $j = -3, -2, -1$), $p(c_j \mid s_i)$ is the probability that word c_j appears in any of these positions. This permits TLC to generalize over variations in the placement of premodifiers, for example. In a similar manner, there is generalization over the three positions to the right of the target. The local window does not extend beyond a sentence boundary. A window size of ± 3 was chosen on empirical grounds; a preliminary study using parts of the Brown corpus that had been manually tagged with senses in WordNet (Landes, Leacock, and Tengi 1998) and a version of TLC that looked only at local open-class words performed best with this width when tested on a large number of nouns, verbs, and adjectives.

3. Local closed-class items include those elements not assigned a noun, verb, adjective, or adverb tag. Among these are determiners, prepositions, pronouns, and punc-

tuation. For this cue type, $p(c_j \mid s_i)$ is the probability that item c_j appears precisely at location j for sense s_i. Positions $j = -2, -1, 1, 2$ are used. The global probabilities, for example $p(the_{-1})$, are based on counts of closed-class items found at these positions relative to the nouns in a large text corpus. The local window width of ±2 was selected after pilot testing on the semantically tagged Brown corpus. As in (2) above, the local window does not extend beyond a sentence boundary.

4. Part-of-speech tags in the positions $j = -2, -1, 0, 1, 2$ are also used as cues. The probabilities for these tags are computed for specific positions (e.g., $p(DT_{-1} \mid s_i)$, $p(DT_{-1})$) in a manner similar to that described in (3) above.

When TLC is configured to use only topical information, cue type (1) is employed. When it is configured for local information, cue types (2), (3), and (4) are used. Finally, in combined mode, the set of cues contains all four types.

2.3 Results

Figures 1 to 3 show the accuracy of the classifier as a function of the size of the training set when using local context, topical context, and a combination of the two, averaged across three runs for each training set. To the extent that the words used are representative, some clear differences appear as a function of syntactic category. With the verb *serve*, local context was more reliable than topical context at all levels of training (78% versus 68% with 200 training examples for the least frequent sense). The combination of local and topical context showed improvement (83%) over either form alone (see Figure 1). With the adjective *hard*, local context was much more reliable as an indicator of sense than topical context for all training sizes (83% versus 60% with 200 training examples) and the combined classifier's performance (at 83%) was the same as for local (see Figure 2). In the case of the noun *line*, topical was slightly better than local at all set sizes, but with 200 training examples, their combination yielded 84% accuracy, greater than either topical (78%) or local (67%) alone (see Figure 3).

To summarize, local context was more reliable than topical context as an indicator of sense for this verb and this adjective, but slightly less reliable for this noun. The combination of local and topical context showed improved or equal performance for all three words. Performance for all of the classifiers improved with increased training size. All classifiers performed best with at least 200 training examples per sense, but the learning curve tended to level off beyond a minimum 100 training examples.

These results are consistent with those of Yarowsky (1993), based on his experiments with pseudowords, homophones, and homonyms (discussed below). He observed that performance for verbs and adjectives dropped sharply as the window increased, while distant context remained useful for nouns. Thus one is tempted to conclude that nouns depend more on topic than do verbs and adjectives. But such a conclusion is probably an overgeneralization, inasmuch as some noun senses are clearly nontopical. Thus, Leacock, Towell, and Voorhees (1993) found that some senses of the noun *line* are not susceptible to disambiguation with topical context. For example, the 'textual' sense of *line* can appear with any topic, whereas the 'product' sense of *line* cannot. When it happens that a nontopical sense accounts for a large proportion of occurrences (in our study, all senses of *hard* are nontopical), then adding topical context to local will have little benefit and may even reduce accuracy.

One should not conclude from these results that the topical classifiers and TLC are inferior to the classifiers reviewed in Section 2. In our experiments, monosemous collocations in WordNet that contain the target word were systematically removed from the training and testing materials. This was done on the assumption that these words are not ambiguous. Removing them undoubtedly made the task more difficult than it would normally be. How much more difficult? An estimate is possible. We

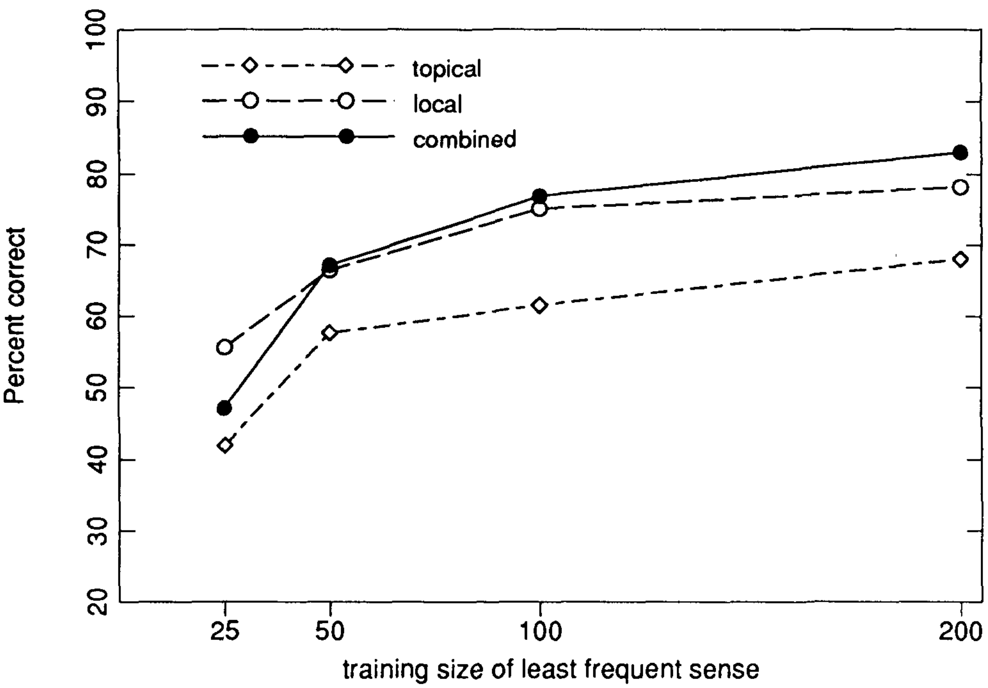

Figure 1
Classifier performance on four senses of the verb *serve*. Percentage accounted for by most frequent sense = 41%.

searched through 7,000 sentences containing *line* and found 1,470 sentences contained *line* as the head of a monosemous collocation in WordNet, i.e., *line* could be correctly disambiguated in some 21% of those 7,000 sentences simply on the basis of the WordNet entries in which it occurred. In other words, if these sentences had been included in the experiment—and had been identified by automatic lookup—overall accuracy would have increased from 83% to 87%.

Using topical context alone, TLC performs no worse than other topical classifiers. Leacock, Towell, and Voorhees (1993) report that the three topical classifiers tested averaged 74% accuracy on six senses of the noun *line*. With these same training and testing data, TLC performed at 73% accuracy. Similarly, when the content vector and neural network classifiers were run on manually tagged training and testing examples of the verb *serve*, they averaged 74% accuracy—as did TLC using only topical context. When local context is combined with topical, TLC is superior to the topical classifiers compared in the Leacock, Towell, and Voorhees (1993) study.

2.4 Improving the Precision of Sense Identification

Just how useful is a sense classifier whose accuracy is 85% or less? Probably not very useful if it is part of a fully automated NLP application, but its performance might be adequate in an interactive application (e.g., machine-assisted translation, on-line thesaurus functions in word processing, interactive information retrieval). In fact, when recall does not have to be 100% (as when a human is in the loop) the precision of the classifier can be improved considerably. The classifier described above always selects the sense that has the highest probability. We have observed that when

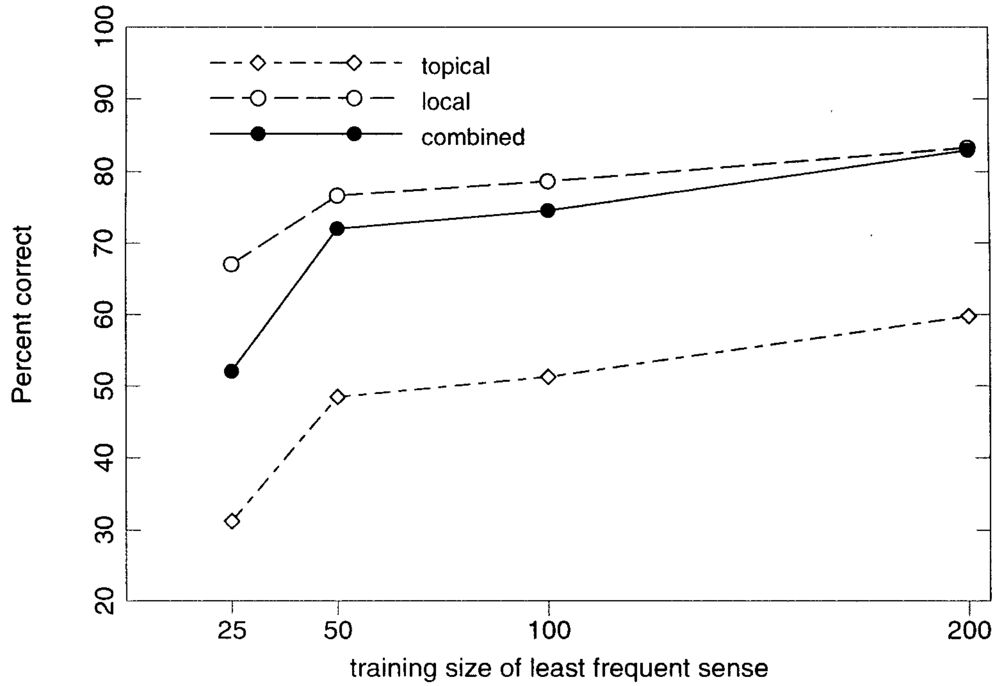

Figure 2
Classifier performance on three senses of the adjective *hard*. Percentage accounted for by most frequent sense = 80%.

the difference between the probability of this sense and that of the second highest is relatively small, the classifier's choice is often incorrect. One way to improve the precision of the classifier, though at the price of reduced recall, is to identify these situations and allow it to respond *do not know* rather than forcing a decision.

What is needed is a measure of the difference in the probabilities of the two senses. Following the approach of Dagan and Itai (1994), we use the log of the ratio of the probabilities $ln(p_1/p_2)$ for this purpose. Based on this value, a threshold Θ can be set to control when the classifier selects the most probable sense. For example, if $\Theta = 2$, then $ln(p_1/p_2)$ must be 2 or greater for a decision to be made. Dagan and Itai (1994) also describe a way to make the threshold dynamic so that it adjusts for the amount of evidence used to estimate p_1 and p_2. The basic idea is to create a one-tailed confidence interval so that we can state with probability $1 - \alpha$ that the true value of the difference measure is greater than Θ. When the amount of evidence is small, the value of the measure must be larger in order to insure that Θ is indeed exceeded.

Table 2 shows precision and recall values for *serve, hard,* and *line* at eight different settings of Θ using a 60% confidence interval. TLC was first trained on 100 examples of each sense, and it was then tested on separate 100-example sets. In all cases, precision was positively correlated with the square root of Θ (all r values $> .97$), and recall was negatively correlated with the square root of Θ (r values $< -.96$). As cross-validation, the equations of the lines that fit the precision and recall results on the test sample were used to predict the precision and recall at the various values of Θ on a second test sample. They provided a good fit to the new data, accounting for an average of 93% of the variance. The standard errors of estimate for *hard, serve,* and *line* were .028,

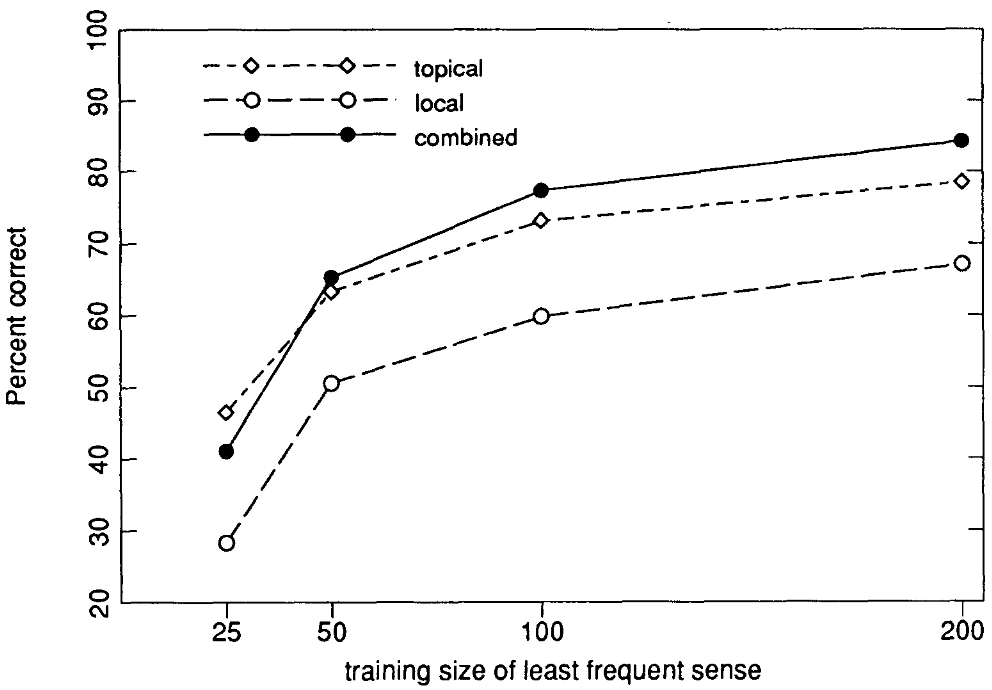

Figure 3
Classifier performance on six senses of the noun *line*. Percentage accounted for by most frequent sense = 54%.

Table 2
Recall and precision at various levels of the threshold for one test sample, following training with 200 examples of each sense.

ThresholdValue	serve		hard		line	
for $ln(p_1/p_2)$	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)
0	100	78	100	77	100	76
.25	99	78	97	78	97	78
.5	97	79	95	79	92	80
1	94	80	90	80	88	82
2	88	82	79	82	77	86
4	72	89	62	88	61	91
8	45	94	39	98	38	97
16	14	98	18	100	14	100
Most Frequent Sense		41		80		54

.030, and .029 for precision, and .053, .068, and .041 for recall. This demonstrates that it is possible to produce accurate predictions of precision and recall as a function of Θ for new test sets.

When the threshold is set to a large value, precision approaches 100%. The criterion thus provides a way to locate those cases that can be identified automatically with very high accuracy. When TLC uses a high criterion for assigning senses, it can be used to augment the training examples by automatically collecting new examples from the test corpus.

In summary, the results obtained with TLC support the following preliminary conclusions: (a) improvement with training levels off after about 100 training examples for the least frequent sense; (b) the high predictive power of local context for the verb and adjective indicate that the local parameters effectively capture syntactically mediated relations, e.g., the subject and object or complement of verbs, or the noun that an adjective modifies; (c) nouns may be more "topical" than verbs and adjectives, and therefore benefit more from the combination of topical and local context; (d) the precision of TLC can be considerably improved at the price of recall, a trade-off that may be desirable in some interactive NLP applications.

A final observation we can make is that when topical and local information is combined, what we have called "nontopical senses" can reduce overall accuracy. For example, the 'textual' sense of *line* is relatively topic-independent. The results of the *line* experiment were not affected too adversely because the nontopical sense of *line* accounted for only 10% of the training examples. The effects of nontopical senses will be more serious when most senses are nontopical, as in the case of many adjectives and verbs.

The generality of these conclusions must, of course, be tested with additional words, which brings us to the problem of obtaining training and testing corpora. On one hand, it is surprising that a purely statistical classifier can "learn" how to identify a sense of a polysemous word with as few as 100 example contexts. On the other hand, anyone who has manually built such sets knows that even collecting 100 examples of each sense is a long and tedious process. The next section presents one way in which the lexical knowledge in WordNet can be used to extract training examples automatically.

3. Unsupervised Training

Corpus-based word sense identifiers are data hungry—it takes them mere seconds to digest all of the information contained in training materials that take months to prepare manually. So, although statistical classifiers are undeniably effective, they are not feasible until we can obtain reliable unsupervised training data. In the Gale, Church, and Yarowsky (1992a) study, training and testing materials were automatically acquired using an aligned French-English bilingual corpus by searching for English words that have two different French translations. For example, English tokens of *sentence* were translated as either *peine* or *phrase*. They collected contexts of *sentence* translated as *peine* to build a corpus for the judicial sense, and collected contexts of *sentence* translated as *phrase* to build a corpus for the grammatical sense. One problem with relying on bilingual corpora for data collection is that bilingual corpora are rare, and aligned bilingual corpora are even rarer. Another is that since French and English are so closely related, different senses of polysemous English words often translate to the same French word. For example, *line* is equally polysemous in French and English—and most senses of *line* translate into French as *ligne*.

Several artificial techniques have been used so that classifiers can be developed and tested without having to invest in manually tagging the data: Yarowsky (1993) and Schütze (1995) have acquired training and testing materials by creating pseudowords from existing nonhomographic forms. For example, a pseudoword was created by combining *abused/escorted*. Examples containing the string *escorted* were collected to train on one sense of the pseudoword and examples containing the string *abused* were collected to train on the other sense. In addition, Yarowsky (1993) used homophones (e.g., *cellar/seller*) and Yarowsky (1994) created homographs by stripping accents from French and Spanish words. Although these latter techniques are useful in their own right (e.g., spoken language systems or corrupted transmissions), the resulting materials do not generalize to the acquisition of tagged training for real polysemous or even homographic words. The results of disambiguation strategies reported for pseudowords and the like are consistently above 95% overall accuracy, far higher than those reported for disambiguating three or more senses of polysemous words (Wilks et al. 1993; Leacock, Towell, and Voorhees 1993).

Yarowsky (1992) used a thesaurus to collect training materials. He tested the unsupervised training materials on 12 nouns with almost perfect results on homonyms (95-99%), 72% accuracy for four senses of *interest,* and 77% on three senses of *cone.* The training was collected in the following manner. Take a *Roget's* category—his examples were TOOL and ANIMAL—and collect sentences from a corpus (in this case, *Grolier's Encyclopedia*) using the words in each category. Consider the noun *crane,* which appears in both the *Roget's* categories TOOL and ANIMAL. To represent the TOOL category, Yarowsky extracted contexts from *Grolier's Encyclopedia.* For example, contexts with the words *adz, shovel, crane, sickle,* and so on. Similarly he collected sentences with names of animals from the ANIMAL category. In these samples, *crane* and *drill* appeared under both categories. Yarowsky points out that the resulting noise will be a problem only when one of the spurious senses is salient, dominating the training set, and he uses frequency-based weights to minimize these effects. We propose to minimize spurious training by using monosemous words and collocations—on the assumption that, if a word has only one sense in WordNet, it is monosemous.

Schütze (1995) developed a statistical topical approach to word sense identification that provides its own automatically extracted training examples. For each occurrence *t* of a polysemous word in a corpus, a context vector is constructed by summing all the vectors that represent the co-occurrence patterns of the open-class words in *t*'s context (i.e., topical information is expressed as a kind of second-order co-occurrence). These context vectors are clustered, and the centroid of each cluster is used to represent a "sense." When given a new occurrence of the word, a vector of the words in its context is constructed, and this vector is compared to the sense representations to find the closest match. Schütze has used the method to disambiguate pseudowords, homographs, and polysemous words. Performance varies depending, in part, on the number of clusters that are created to represent senses, and on the degree to which the distinctions correspond to different topics. This approach performs very well, especially with pseudowords and homographs. However, there is no automatic means to map the sense representations derived from the system onto the more conventional word senses found in dictionaries. Consequently, it does not provide disambiguated examples that can be used by other systems.

Yarowsky (1995) has proposed automatically augmenting a small set of experimenter-supplied **seed** collocations (e.g., *manufacturing plant* and *plant life* for two different senses of the noun *plant*) into a much larger set of training materials. He resolved the problem of the sparseness of his collocations by iteratively bootstrapping acquisition of training materials from a few seed collocations for each sense of a homograph. He locates examples containing the seeds in the corpus and analyzes these to find new predictive patterns in these sentences and retrieves examples containing these patterns. He repeats this step iteratively. Results for the 12 pairs of homographs reported are almost perfect. In his paper, Yarowsky suggests WordNet as a source for the seed collocations—a suggestion that we pursue in the next section.

WordNet is particularly well suited to the task of locating sense-relevant context because each word sense is represented as a node in a rich semantic lexical network with synonymy, hyponymy, and meronymy links to other words, some of them polysemous and others monosemous. These lexical "relatives" provide a key to finding relevant training sentences in a corpus. For example, the noun *suit* is polysemous, but one sense of it has *business suit* as a monosemous daughter and another has *legal proceeding* as a hypernym. By collecting sentences containing the unambiguous nouns *business suit* and *legal proceeding* we can build two corpora of contexts for the respective senses of the polysemous word. All the systems described in Section 2.1 could benefit from the additional training materials that monosemous relatives can provide.

3.1 WordNet: A Lexical Database for English

The WordNet on-line lexical database (Miller 1990, 1995) has been developed at Princeton University over the past 10 years.[4] Like a standard dictionary, WordNet contains the definitions of words. It differs from a standard dictionary in that, instead of being organized alphabetically, WordNet is organized conceptually. The basic unit in WordNet is a synonym set, or **synset**, which represents a lexicalized concept. For example, WordNet Version 1.5 distinguishes between two senses of the noun *shot* with the synsets {shot, snapshot} and {shot, injection}. In the context, "The photographer took a *shot* of Mary," the word *snapshot* can be substituted for one sense of *shot.* In the context, "The nurse gave Mary a flu *shot,"* the word *injection* can be substituted for another sense of *shot.*

Nouns, verbs, adjectives, and adverbs are each organized differently in WordNet. All are organized in synsets, but the semantic relations among the synsets differ depending on the grammatical category, as can be seen in Table 3.

Nouns are organized in a hierarchical tree structure based on hypernymy/hyponymy. The hyponym of a noun is its subordinate, and the relation between a hyponym and its hypernym is an *is a kind of* relation. For example, *maple* is a hyponym of *tree,* which is to say that a *maple* is a kind of *tree.* Hypernymy (supername) and its inverse, hyponymy (subname), are transitive semantic relations between synsets. Meronymy (part-name), and its inverse holonymy (whole-name), are complex semantic relations that distinguish component parts, substantive parts, and member parts.

The verbal hierarchy is based on troponymy, the *is a manner of* relation. For example, *stroll* is a troponym of *walk,* which is to say that strolling is a manner of walking. Entailment relations between verbs are also coded in WordNet.

The organization of attributive adjectives is based on the antonymy relation. Where direct antonyms exist, adjective synsets point to antonym synsets. A head adjective is one that has a direct antonym (e.g., *hot* versus *cold* or *long* versus *short*). Many adjectives, like *sultry,* have no direct antonyms. When an adjective has no direct antonym, its synset points to a head that is semantically similar to it. Thus *sultry* and *torrid* are similar in meaning to *hot,* which has the direct antonym of *cold.* So, although *sultry* has no direct antonym, it has *cold* as its indirect antonym.

Relational adjectives do not have antonyms; instead they point to nouns. Consider the difference between a *nervous disorder* and a *nervous student.* In the former, *nervous* pertains to a noun, as in *nervous system,* whereas the latter is defined by its relation to other adjectives—its synonyms (e.g., *edgy*) and antonyms (e.g., *relaxed*).

Adverbs have synonymy and antonymy relations. When the adverb is morphologically related to an adjective (when an *-ly* suffix is added to an adjective) and semantically related to the adjective as well, the adverb points to the adjective.

We have had some success in exploiting WordNet's semantic relations for word sense identification. Since the main problem with classifiers that use local context is

[4] Available by anonymous ftp from clarity.princeton.edu cd pub/wordnet or http://www.cogsci.princeton.edu/~wn/

Table 3
Semantic relations in WordNet.

Semantic Relation	Syntactic Category	Examples
Synonymy	Noun	pipe, tube
(similar)	Verb	rise, ascend
	Adj	sad, unhappy
	Adv	rapidly, speedily
Antonymy	Adj	wet, dry
(opposite)	Adv	rapidly, slowly
	Noun	top, bottom
	Verb	rise, fall
Hyponymy	Noun	sugar maple, maple
(subordinate)		maple, tree
		tree, plant
Meronymy	Noun	brim, hat
(part)		gin, martini
		ship, fleet
Troponymy	Verb	march, walk
(manner)		whisper, speak
Entailment	Verb	drive, ride
		divorce, marry
Derivation	Adj	magnetic, magnetism
	Adv	simply, simple

the sparseness of the training data, Leacock and Chodorow (1998) used a proximity measure on the hypernym relation to replace the subject and complement of the verb *serve* in the testing examples with the subject and complement from training examples that were "closest" to them in the noun hierarchy. For example, one of the test sentences was "Sauerbraten is usually served with dumplings," where neither *sauerbraten* nor *dumpling* appeared in any training sentence. The similarity measures on WordNet found that *sauerbraten* was most similar to *dinner* in the training, and *dumpling* to *bacon*. These nouns were substituted for the novel ones in the test sets. Thus the sentence "*Dinner* is usually served with *bacon*" was substituted for the original sentence. Augmentation of the local context classifier with WordNet similarity measures showed a small but consistent improvement in the classifier's performance. The improvement was greater with the smaller training sets.

Resnik (1992) uses an information-based measure, the most informative class, on the WordNet taxonomy. A class consists of the synonyms found at a node and the synonyms at all the nodes that it dominates (all of its hyponyms). Based on verb/object pairs collected from a corpus, Resnik found, for example, that the objects for the verb *open* fall into two classes: receptacle and oral communication. Conversely, the class of a verb's object could be used to determine the appropriate sense of that verb.

The experiments in the next section depend on a subset of the WordNet lexical relations, those involving monosemous relatives, so we were interested in determining just what proportion of word senses have such relatives. We examined 8,500 polysemous nouns that appeared in a moderate-size, 25-million-word corpus. In all, these 8,500 nouns have more than 24,000 WordNet senses. Restricting the relations to syn-

Table 4
Training materials and their frequencies for five senses of *line*.

product		formation		text		cord		phone	
product line	95	picket line	46	headline	52	rope	28	hot line	48
business line	5	line of		punch line	23	ropes	27	phone line	48
		succession	10	opening line	10	clothesline	12	private line	2
		bread line	7	tag line	7	fishing line	11	toll line	2
		single file	7	line of poetry	4	shoelaces	4		
		conga line	7	newspaper		twine	3		
		reception line	3	headline	2	dental floss	2		
		ticket line	3	gag line	2	high wire	2		
		chow line	2			jump rope	2		
		rivet line	1			lasso	2		
		single files	1			lead line	1		
		trap line	1			mooring line	1		
						...			

onyms, immediate hyponyms (i.e., daughters), and immediate hypernyms (parents), we found that about 64% (15,400) have monosemous relatives attested in the corpus. With larger corpora (e.g., with text obtained by Web crawling) and more lexical relations (e.g., meronymy), this percentage can be expected to increase.

3.2 Training on WordNet's Monosemous Relatives

The approach we have used is related to that of Yarowsky (1992) in that training materials are collected using a knowledge base, but it differs in other respects, notably in the selection of training and testing materials, the choice of a knowledge base, and use of both topical and local classifiers. Yarowsky collects his training and testing materials from a specialized corpus, *Grolier's Encyclopedia*. It remains to be seen whether a statistical classifier trained on a topically organized corpus such as an encyclopedia will perform in the same way when tested on general unrestricted text, such as newspapers, periodicals, and books. One of our goals is to determine whether automatic extraction of training examples is feasible using general corpora. In his experiment, Yarowsky uses an updated on-line version of *Roget's Thesaurus* that is not generally available to the research community. The only generally available version of *Roget's* is the 1912 edition, which contains many lexical gaps. We are using WordNet, which can be obtained via anonymous ftp. Yarowsky's classifier is purely topical, but we also examine local context. Finally, we hope to avoid inclusion of spurious senses by using monosemous relatives.

In this experiment we collected monosemous relatives of senses of 14 nouns. Training sets are created in the following manner. A program called AutoTrain retrieves from WordNet all of the monosemous relatives of a polysemous word sense, samples and retrieves example sentences containing these monosemous relatives from a 30-million-word corpus of the *San Jose Mercury News*, and formats them for TLC. The sampling process retrieves the "closest" relatives first. For example, suppose that the system is asked to retrieve 100 examples for each sense of the noun *court*. The system first looks for the strongest or top-level relatives: for monosemous synonyms of the sense (e.g., *tribunal*) and for daughter collocations that contain the target word as the head (e.g., *superior court*) and tallies the number of examples in the corpus for each. If the corpus has 100 or more examples for these top-level relatives, it retrieves a sampling of them and formats them for TLC. If there are not enough top-level examples,

Table 5
TLC's performance when training on (1) manually tagged data and (2) monosemous relatives of polysemous words.

Target Word	Sense and Priors (%)		Manually Tagged Training	Monosemous Relative Training
bill			89.1%	88.5%
	legal	85	97	98
	invoice	15	56	35
duty			94.3%	93.5%
	tariff	56	95	97
	obligation	44	99	91
line			82.6%	74.7%
	product	67	97	86
	phone	10	67	51
	cord	9	79	74
	formation	8	49	26
	text	6	67	52
rate			81%	79%
	monetary	65	90	80
	frequency	35	66	77
shot			89.8%	89.1%
	sports	74	99	95
	gunshot	17	77	87
	opportunity	8	36	47
work			75.3%	65.2%
	activity	55	81	81
	product	45	68	46

the remainder of the target's monosemous relatives are inspected in the order: all other daughters; hyponym collocations that contain the target; all other hyponyms; hypernyms; and, finally, sisters. AutoTrain takes as broad a sampling as possible across the corpus and never takes more than one example from an article. The number of examples for each relative is based on the relative proportion of its occurrences in the corpus. Table 4 shows the monosemous relatives that were used to train five senses of the noun *line*—the monosemous relatives of the sixth sense in the original study, *line* as an abstract division, are not attested in the SJM corpus.

The purpose of the experiment was to see how well TLC performed using unsupervised training and, when possible, to compare this with its performance when training on the manually tagged materials being produced at Princeton's Cognitive Science Laboratory.[5] When a sufficient number of examples for two or more senses were available, 100 examples of each sense were set aside to use in training. The remainder were used for testing. Only the topical and local open-class cues were used, since preliminary tests showed that performance declined when using local closed-class and part-of-speech cues obtained from the monosemous relatives. This is not surprising, as many of the relatives are collocations whose local syntax is quite different from that

[5] These materials are being produced under the direction of Shari Landes. Monosemous collocations that contain the target as the head have not been removed from the materials reported here—except for the noun *line*.

Table 6
TLC's performance when training on monosemous relatives of polysemous words. (Manually tagged data were used for scoring but there were not enough examples to train on.)

Target Word	Sense and Priors (%)		Monosemous Relative Training	Target Word	Sense and Priors (%)		Monosemous Relative Training
bank			92.8%	security			67.4%
	institution	92	98		certificate	67	85
	land form	8	35		precaution	37	63
company			86.9%	stock			99.9%
	business	87	90		capital	95	100
	troupe	7	67		broth	5	98
	guests	6	70				
court			97.5%	strike			80.8%
	tribunal	96	99		work stoppage	78	98
	sports	4	71		attack	23	21
party			87.8%	trade			80.2%
	political	77	96		commerce	81	92
	social	23	59		swap	19	31

of the polysemous word in its typical usage. For example, the 'formation' sense of *line* is often followed by an *of*-phrase as in *a line of children*, but its relative, *picket line*, is not. Prior probabilities for the sense were taken from the manually tagged materials.

Table 5 shows the results when TLC was trained on monosemous relatives and on manually tagged training materials. Baseline performance is when the classifier always chooses the most frequent sense. Eight additional words had a sufficient number of manually tagged examples for testing but not for training TLC. These are shown in Table 6.

For four of the examples in Table 5, training with relatives produced results within 1% or 2% of manually tagged training. *Line* and *work*, however, showed a substantial decrease in performance. In the case of *line*, this might be due to overly specific training contexts. Almost half of the training examples for the 'formation' sense of *line* come from one relative, *picket line*. In fact, all of the monosemous relatives, except for *rivet line* and *trap line*, are human formations. This may have skewed training so that the classifier performs poorly on other uses of *line* as formation.

In order to compare our results with those reported in Yarowsky (1992), we trained and tested on the same two senses of the noun *duty* that Yarowsky had tested ('obligation' and 'tax'). He reported that his thesaurus-based approach yielded 96% precision with 100% recall. TLC used training examples based on monosemous WordNet relatives and correctly identified the senses with 93.5% precision at 100% recall.

Table 6 shows TLC's performance on the other eight words after training with monosemous relatives and testing on manually tagged examples. Performance is about the same as, or only slightly better than, the highest prior probability. In part, this is due to the rather high probability of the most frequent sense for this set.

The values in the table are based on decisions made on all test examples. If a threshold is set for TLC (see Section 2.4), precision of the classifier can be increased substantially, at the expense of recall. Table 7 shows recall levels when TLC is trained on monosemous relatives and the value of Θ is set for 95% precision. Operating in this mode, the classifier can gather new training materials, automatically, and with

Table 7
Percentage of recall when the precision is 95%.

Word	Recall at 95% Precision	Word	Recall at 95% Precision
bank	95%	rate	36%
bill	73%	security	35%
company	72%	shot	77%
duty	93%	strike	37%
line	45%	trade	42%
party	78%	work	2%

high precision. This is a particularly good way to find clear cases of the most frequent sense.

The results also show that not all words are well suited to this kind of operation. Little can be gained for a word like *work*, where the two senses, 'activity' and 'product,' are closely related and therefore difficult for the classifier to distinguish, due to a high degree of overlap in the training contexts. Problems of this sort can be detected even before testing, by computing correlations between the vectors of open-class words for the different senses. The cosine correlation between the 'activity' and 'product' senses of *work* is $r = .49$, indicating a high degree of overlap. The mean correlation between pairs of senses for the other words in Table 7 is $r = .31$.

4. Conclusion

Our evidence indicates that local context is superior to topical context as an indicator of word sense when using a statistical classifier. The benefits of adding topical to local context alone depend on syntactic category as well as on the characteristics of the individual word. The three words studied yielded three different patterns; a substantial benefit for the noun *line*, slightly less for the verb *serve*, and none for the adjective *hard*. Some word senses are simply not limited to specific topics, and appear freely in many different domains of discourse. The existence of nontopical senses also limits the applicability of the "one sense per discourse" generalization of Gale, Church, and Yarowsky (1992b), who observed that, within a document, a repeated word is almost always used in the same sense. Future work should be directed toward developing methods for determining when a word has a nontopical sense. One approach to this problem is to look for a word that appears in many more topical domains than its total number of senses.

Because the supply of manually tagged training data will always be limited, we propose a method to obtain training data automatically using commonly available materials: exploiting WordNet's lexical relations to harvest training examples from LDC corpora or even the World Wide Web. We found this method to be effective, although not as effective as using manually tagged training. We have presented the components of a system for acquiring unsupervised training materials that can be used with any statistical classifier.

The components can be fit together in the following manner. For a polysemous word, locate the monosemous relatives for each of its senses in WordNet and extract examples containing these relatives from a large corpus. Senses whose contexts greatly overlap can be identified with a simple cosine correlation. Often, correlations are high between senses of a word that are systematically related, as we saw for the 'activity' and 'product' senses of *work*. In some cases, the contexts for the two closely related senses may be combined.

Since the frequencies of the monosemous relatives do not correlate with the frequencies of the senses, prior probabilities must be estimated for classifiers that use them. In the experiments of Section 3.2, these were estimated from the testing materials. They can also be estimated from a small manually tagged sample, such as the parts of the Brown corpus that have been tagged with senses in WordNet.

When the threshold is set to maximize precision, the results are highly reliable and can be used to support an interactive application, such as machine-assisted translation, with the goal of reducing the amount of interaction.

Although we have looked at only a few examples, it is clear that, given WordNet and a large enough corpus, the methods outlined for training on monosemous relatives can be generalized to build training materials for thousands of polysemous words.

Acknowledgments

We are indebted to the other members of the WordNet group who have provided advice and technical support: Christiane Fellbaum, Shari Landes, and Randee Tengi. We are also grateful to Paul Bagyenda, Ben Johnson-Laird and Joshua Schecter. We thank Scott Wayland, Tim Allison and Jill Hollifield for tagging the *serve* and *hard* corpora. Finally we are grateful to the three anonymous CL reviewers for their comments and advice.

This material is based upon work supported in part by the National Science Foundation under NSF Award No. IRI95-28983 and by the Defense Advanced Research Projects Agency, Grant No. N00014-91-1634.

References

Brill, Eric. 1994. Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle. AAAI.

Bruce, Rebecca and Janyce Wiebe. 1994a. A new approach to word sense disambiguation. In *Proceedings of the ARPA Workshop on Human Language Technology*, San Francisco, CA, Morgan Kaufman.

Bruce, Rebecca and Janyce Wiebe. 1994b. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting*, Las Cruces, NM. Association for Computational Linguistics.

Chiang, T-H., Y-C. Lin, and K-Y Su. 1995. Robust learning, smoothing, and parameter tying on syntactic ambiguity resolution. *Computational Linguistics*, 21(3):321–349.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4).

Gale, William, Kenneth W. Church, and David Yarowsky. 1992a. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26.

Gale, William, Kenneth W. Church, and David Yarowsky. 1992b. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*, San Francisco, CA, Morgan Kaufmann.

Golding, Andrew. 1995. A Bayesian hybrid method for context-sensitive spelling correction. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA. ACL.

Good, I. F. 1953. The population frequencies of species and the estimation of population parameters. *Biometrica*, 40:237–264.

Hearst, Marti A. 1991. Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, pages 1–22, Oxford.

Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press, Cambridge, MA.

Landes, Shari, Claudia Leacock, and Randee Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: A Lexical Reference System and its Application*. MIT Press, Cambridge, MA.

Leacock, Claudia and Martin Chodorow. 1998. Combining local context with WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: A Lexical Reference System and its Application*. MIT Press, Cambridge, MA.

Leacock, Claudia, Geoffrey Towell, and Ellen M. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings* *of the ARPA Workshop on Human Language Technology*, San Francisco, CA, Morgan Kaufman.

Leacock, Claudia, Geoffrey Towell, and Ellen M. Voorhees. 1996. Towards building contextual representations of word senses using statistical models. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA.

Miller, George A., editor, 1990. *WordNet: An On-Line Lexical Database*. Volume 3(4) of the *International Journal of Lexicography*. Oxford University Press.

Miller, George A. 1995. WordNet: An on-line lexical database. *Communications of the ACM*, 38(11).

Resnik, Philip. 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In *Workshop on Statistically-Based Natural-Language-Processing Techniques*, San Jose, July.

Schütze, Hinrich. 1995. *Ambiguity and Language Learning: Computational and Cognitive Models*. Ph. D. thesis, Stanford University.

Vẽronis, Jean and Nancy Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of COLING-90*.

Weiss, Stephen. 1973. Learning to disambiguate. *Information Storage and Retrieval*, 9.

Wilks, Yorick, Dan Fass, Cheng Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. 1993. Machine tractable dictionary tools. In James Pustejovsky, editor, *Semantics and the Lexicon*. Kluwer, Dordrecht.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, Nantes, France.

Yarowsky, David. 1993. One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Technology*, San Francisco, CA, Morgan Kaufman.

Yarowsky, David. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting*. Las Cruces, NM. Association for Computational Linguistics.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting*, Cambridge, MA. Association for Computational Linguistics.