ONE SENSE PER COLLOCATION

*David Yarowsky**

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
yarowsky@unagi.cis.upenn.edu

ABSTRACT

Previous work [Gale, Church and Yarowsky, 1992] showed that with high probability a polysemous word has one sense per discourse. In this paper we show that for certain definitions of collocation, a polysemous word exhibits essentially only one sense per collocation. We test this empirical hypothesis for several definitions of sense and collocation, and discover that it holds with 90-99% accuracy for binary ambiguities. We utilize this property in a disambiguation algorithm that achieves precision of 92% using combined models of very local context.

1. INTRODUCTION

The use of collocations to resolve lexical ambiguities is certainly not a new idea. The first approaches to sense disambiguation, such as [Kelly and Stone 1975], were based on simple hand-built decision tables consisting almost exclusively of questions about observed word associations in specific positions. Later work from the AI community relied heavily upon selectional restrictions for verbs, although primarily in terms of features exhibited by their arguments (such as +DRINKABLE) rather than in terms of individual words or word classes. More recent work [Brown et al. 1991][Hearst 1991] has utilized a set of discrete local questions (such as *word-to-the-right*) in the development of statistical decision procedures. However, a strong trend in recent years is to treat a reasonably wide context window as an unordered bag of independent evidence points. This technique from information retrieval has been used in neural networks, Bayesian discriminators, and dictionary definition matching. In a comparative paper in this volume [Leacock et al. 1993], all three methods under investigation used words in wide context as a pool of evidence independent of relative position. It is perhaps not a coincidence that this work has focused almost exclusively on nouns, as will be shown in Section 6.2. In this study we will return again to extremely local sources of evidence, and show that models of discrete syntactic relationships have considerable advantages.

*This research was supported by an NDSEG Fellowship and by DARPA grant N00014-90-J-1863. The author is also affiliated with the Linguistics Research Department of AT&T Bell Laboratories, and greatly appreciates the use of its resources in support of this work. He would also like to thank Eric Brill, Bill Gale, Libby Levison, Mitch Marcus and Philip Resnik for their valuable feedback.

2. DEFINITIONS OF SENSE

The traditional definition of word sense is "One of several meanings assigned to the same orthographic string". As meanings can always be partitioned into multiple refinements, senses are typically organized in a tree such as one finds in a dictionary. In the extreme case, one could continue making refinements until a word has a slightly different sense every time it is used. If so, the title of this paper is a tautology. However, the studies in this paper are focused on the sense distinctions at the top of the tree. A good working definition of the distinctions considered are those meanings which are not typically translated to the same word in a foreign language.

Therefore, one natural type of sense distinction to consider are those words in English which indeed have multiple translations in a language such as French. As is now standard in the field, we use the Canadian Hansards, a parallel bilingual corpus, to provide sense tags in the form of French translations. Unfortunately, the Hansards are highly skewed in their sense distributions, and it is difficult to find words for which there are adequate numbers of a second sense. More diverse large bilingual corpora are not yet readily available.

We also use data sets which have been hand-tagged by native English speakers. To make the selection of sense distinctions more objective, we use words such as *bass* where the sense distinctions (*fish* and *musical instrument*) correspond to pronunciation differences ([bæs] and [beɪs]). Such data is often problematic, as the tagging is potentially subjective and error-filled, and sufficient quantities are difficult to obtain.

As a solution to the data shortages for the above methods, [Gale, Church and Yarowsky 1992b] proposed the use of "pseudo-words," artificial sense ambiguities created by taking two English words with the same part of speech (such as *guerilla* and *reptile*), and replacing each instance of both in a corpus with a new polysemous word *guerrilla/reptile*. As it is entirely possible that the concepts guerrilla and reptile are represented by the same orthographic string in some foreign language, choosing between these two meanings based on context is a problem a word sense disambiguation algorithm could easily face. "Pseudo-words" are very useful for developing and testing disambiguation methods because of their nearly unlimited availability and the known, fully reliable

ground truth they provide when grading performance.

Finally, we consider sense disambiguation for mediums other than clean English text. For example, we look at word pairs such as *terse/tense* and *cookie/rookie* which may be plausibly confused in optical character recognition (OCR). Homophones, such as *aid/aide*, and *censor/sensor*, are ideal candidates for such a study because large data sets with known ground truth are available in written text, yet they are true ambiguities which must be resolved routinely in oral communication.

We discover that the central claims of this paper hold for all of these potential definitions of sense. This corroborating evidence makes us much more confident in our results than if they were derived solely from a relatively small hand-tagged data set.

3. DEFINITIONS OF COLLOCATION

Collocation means the co-occurrence of two words in some defined relationship. We look at several such relationships, including direct adjacency and first word to the left or right having a certain part-of-speech. We also consider certain direct syntactic relationships, such as verb/object, subject/verb, and adjective/noun pairs. It appears that *content words* (nouns, verbs, adjectives, and adverbs) behave quite differently from *function words* (other parts of speech); we make use of this distinction in several definitions of collocation.

We will attempt to quantify the validity of the one-sense-per-collocation hypothesis for these different collocation types.

4. EXPERIMENTS

In the experiments, we ask two central, related questions:
For each definition of sense and collocation,

- What is the mean entropy of the distribution $Pr(Sense|Collocation)$?
- What is the performance of a disambiguation algorithm which uses only that collocation type as evidence?

We examine several permutations for each, and are interested in how the results of these questions differ when applied to polysemous nouns, verbs, and adjectives.

To limit the already very large number of parameters considered, we study only binary sense distinctions. In all cases the senses being compared have the same part of speech. The selection between different possible parts of speech has been heavily studied and is not replicated here.

4.1. Sample Collection

All samples were extracted from a 380 million word corpus collection consisting of newswire text (AP Newswire and Wall Street Journal), scientific abstracts (from NSF and the Department of Energy), the Canadian Hansards parliamentary debate records, Grolier's Encyclopedia, a medical encyclopedia, over 100 Harper & Row books, and several smaller corpora including the Brown Corpus, and ATIS and TIMIT sentences.[1]

- **Hand Tagged (homographs):** bass, axes, chi, bow, colon, lead, IV, sake, tear, ...
- **French Translation Distinctions:** sentence, duty, drug, language, position, paper, single, ...
- **Homophones:** aid/aide, cellar/seller, censor/sensor, cue/queue, pedal/petal, ...
- **OCR Ambiguities:** terse/tense, gum/gym, deaf/dear, cookie/rookie, beverage/leverage, ...
- **Pseudo-Words:** covered/waved, kissed/slapped, abused/escorted, cute/compatible, ...

Table 1: A sample of the words used in the experiments

The homophone pairs used were randomly selected from a list of words having the same pronunciation or which differed in only one phoneme. The OCR and pseudo-word pairs were randomly selected from corpus wordlists, with the former restricted to pairs which could plausibly be confused in a noisy FAX, typically words differing in only one character. Due to the difficulty of obtaining new data, the hand-tagged and French translation examples were borrowed from those used in our previous studies in sense disambiguation.

4.2. Measuring Entropies

When computing the entropy of $Pr(Sense|Collocation)$, we enumerate all collocations of a given type observed for the word or word pair being disambiguated. Table 2 shows the example of the homophone ambiguity *aid/aide* for the collocation type *content-word-to-the-left*. We list all words[2] appearing in such a collocation with either of these two "senses" of the homograph, and calculate the raw distributional count for each.

Note that the vast majority of the entries in Table 2 have zero as one of the frequency counts. It is not acceptable, however,

[1] Training and test samples were not only extracted from different articles or discourses but also from entirely different blocks of the corpus. This was done to minimize long range discourse effects such as one finds in the AP or Hansards.

[2] Note: the entries in this table are *lemmas* (uninflected root forms), rather than raw words. By treating the verbal inflections *squander, squanders, squandering, and squandered* as the same word, one can improve statistics and coverage at a slight cost of lost subtlety. Although we will refer to "words in collocation" throughout this paper for simplicity, this should always be interpreted as "lemmas in collocation."

Collocation	Frequency as **Aid**	Frequency as **Aide**
foreign	718	1
federal	297	0
western	146	0
provide	88	0
covert	26	0
oppose	13	0
future	9	0
similar	6	0
presidential	0	63
chief	0	40
longtime	0	26
aids-infected	0	2
sleepy	0	1
disaffected	0	1
indispensable	2	1
practical	2	0
squander	1	0

Table 2: A typical collocational distribution for the homophone ambiguity *aid/aide*.

to treat these as having zero probability and hence a zero entropy for the distribution. It is quite possible, especially for the lower frequency distributions, that we would see a contrary example in a larger sample. By cross-validation, we discover for the *aid/aide* example that for collocations with an observed 1/0 distribution, we would actually expect the minor sense to occur 6% of the time in an independent sample, on average. Thus a fairer distribution would be .94/.06, giving a cross-validated entropy of .33 bits rather than 0 bits. For a more unbalanced observed distribution, such as 10/0, the probability of seeing the minor sense decreases to 2%, giving a cross-validated entropy of H(.98,.02) = .14 bits. Repeating this process and taking the weighted mean yields the entropy of the full distribution, in this case .09 bits for the *aid/aide* ambiguity.

For each type of collocation, we also compute how well an observed probability distribution predicts the correct classification for novel examples. In general, this is a more useful measure for most of the comparison purposes we will address. Not only does it reflect the underlying entropy of the distribution, but it also has the practical advantage of showing how a working system would perform given this data.

5. ALGORITHM

The sense disambiguation algorithm used is quite straightforward. When based on a single collocation type, such as the object of the verb or word immediately to the left, the procedure is very simple. One identifies if this collocation type exists for the novel context and if the specific words found are listed in the table of probability distributions (as computed above). If so, we return the sense which was most frequent for that collocation in the training data. If not, we return the sense which is most frequent overall.

When we consider more than one collocation type and combine evidence, the process is more complicated. The algorithm used is based on decision lists [Rivest, 1987], and was discussed in [Sproat, Hirschberg, and Yarowsky 1992]. The goal is to base the decision on the single best piece of evidence available. Cross-validated probabilities are computed as in Section 4.2, and the different types of evidence are sorted by the absolute value of the log of these probability ratios: $Abs(Log(\frac{Pr(Sense_1|Collocation_i)}{Pr(Sense_2|Collocation_i)}))$. When a novel context is encountered, one steps through the decision list until the evidence at that point in the list (such as *word-to-left*="presidential") matches the current context under consideration. The sense with the greatest listed probability is returned, and this cross-validated probability represents the confidence in the answer.

This approach is well-suited for the combination of multiple evidence types which are clearly not independent (such as those found in this study) as probabilities are never combined. Therefore this method offers advantages over Bayesian classifier techniques which assume independence of the features used. It also offers advantages over decision tree based techniques because the training pools are not split at each question. The interesting problems are how one should re-estimate probabilities conditional on questions asked earlier in the list, or how one should prune lower evidence which is categorically subsumed by higher evidence or is entirely conditional on higher evidence. [Bahl et al. 1989] have discussed some of these issues at length, and there is not space to consider them here. For simplicity, in this experiment no secondary smoothing or pruning is done. This does not appear to be problematic when small numbers of independent evidence types are used, but performance should increase if this extra step is taken.

6. RESULTS AND DISCUSSION

6.1. One Sense Per Collocation

For the collocations studied, it appears that the hypothesis of one sense per collocation holds with high probability for binary ambiguities. The experimental results in the *precision* column of Table 3 quantify the validity of this claim. Accuracy varies from 90% to 99% for different types of collocation and part of speech, with a mean of 95%. The significance of these differences will be discussed in Section 6.2.

These precision values have several interpretations. First, they reflect the underlying probability distributions of sense

Collocation Type	Part of Sp.	Ent	Prec	Rec	No Coll	No Data
Content word to immediate right [A]	ALL	.18	.97	.29	.57	.14
	Noun		.98	.25	.66	.09
	Verb		.95	.14	.71	.15
	Adj		.97	.51	.27	.22
Content word to immediate left [B]	ALL	.24	.96	.26	.58	.16
	Noun		.99	.33	.56	.11
	Verb		.91	.23	.47	.30
	Adj		.96	.15	.75	.10
First Content Word to Right	ALL	.33	.94	.51	.09	.40
	Noun		.94	.49	.13	.38
	Verb		.91	.44	.05	.51
	Adj		.96	.58	.04	.38
First Content Word to Left	ALL	.40	.92	.50	.06	.44
	Noun		.96	.58	.06	.36
	Verb		.87	.37	.05	.58
	Adj		.90	.45	.06	.49
Subject ↔ Verb Pairs	Noun	.33	.94	.13	.87	.06
	Verb	.43	.91	.28	.33	.38
Verb ↔ Object Pairs	Noun	.46	.90	.07	.81	.07
	Verb	.29	.95	.36	.32	.32
Adj ↔ Noun	Adj	.14	.98	.54	.20	.26
A&B Above	ALL	–	.97	.47	.31	.21
All Above	ALL	–	.92	.98	.00	.02

Table 3: Includes the entropy of the $Pr(Sense|Collocation)$ distribution for several types of collocation, and the performance achieved when basing sense disambiguation solely on that evidence. Results are itemized by the part of speech of the ambiguous word (*not* of the collocate). Precision (Prec.) indicates percent correct and Recall (Rec.) refers to the percentage of samples for which an answer is returned. Precision is measured on this subset. No collocation (No Coll) indicates the failure to provide an answer because no collocation of that type was present in the test context, and "No Data" indicates the failure to return an answer because no data for the observed collocation was present in the model. See Section 7.3 for a discussion of the "All Above" result. The results stated above are based on the average of the different types of sense considered, and have a mean prior probability of .69 and a mean sample size of 3944.

conditional on collocation. For example, for the collocation type *content-word-to-the-right*, the value of .97 indicates that on average, given a specific collocation we will expect to see the same sense 97% of the time. This mean distribution is also reflected in the *entropy* column.

However, these numbers have much more practical interpretations. If we actually build a disambiguation procedure using exclusively the content word to the right as information, such a system performs with 97% precision on new data where a content word appears to the right and for which there is information in the model.[3] This is considerably higher than the

[3]The correlation between these numbers is not a coincidence. Because the probability distributions are based on cross-validated tests on independent data and weighted by collocation frequency, if on average we find that 97% of samples of a given collocation exhibit the same sense, this is the expected precision of a disambiguation algorithm which assumes one sense per collocation, when applied to new samples of these collocations.

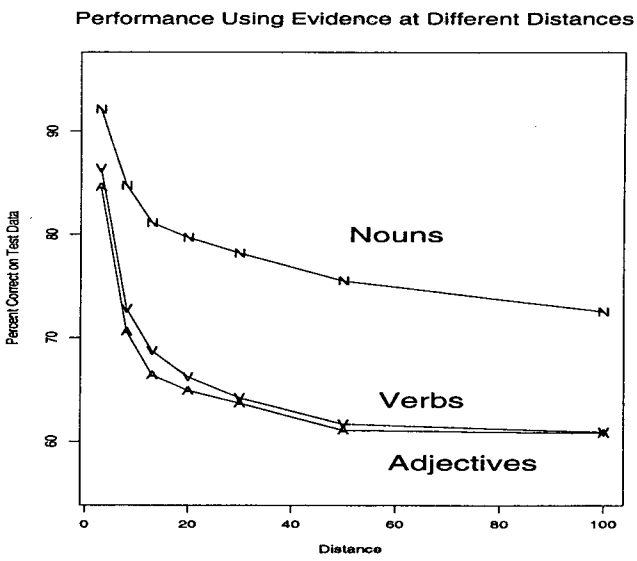

Figure 1: Comparison of the performance of nouns, verbs and adjectives based strictly on a 5 word window centered at the distance shown on the horizontal axis.

performance of 69% one would expect simply by chance due to the unbalanced prior probability of the two senses.

It should be noted that such precision is achieved at only partial recall. The three rightmost columns of Table 3 give the breakdown of the recall. On average, the model *content-word-to-right* could only be applied in 29% of the test samples. In 57% of the cases, no content word appeared to the right, so this collocational model did not hold. In 14% of the cases, a content word did appear to the right, but no instances of that word appeared in the training data, so the model had no information on which to base a decision. There are several solutions to both these deficiencies, and they are discussed in Section 7.

6.2. Part of Speech Differences

It is interesting to note the difference in behavior between different parts of speech. Verbs, for example, derive more disambiguating information from their objects (.95) than from their subjects (.90). Adjectives derive almost all of their disambiguating information from the nouns they modify (.98). Nouns are best disambiguated by directly adjacent adjectives or nouns, with the content word to the left indicating a single sense with 99% precision. Verbs appear to be less useful for noun sense disambiguation, although they are relatively better indicators when the noun is their object rather than their subject.

Figure 1 shows that nouns, verbs and adjectives also differ in their ability to be disambiguated by wider context. [Gale et al. 1993] previously showed that nouns can be disambiguated based strictly on distant context, and that useful information was present up to 10,000 words away. We replicated an experiment in which performance was calculated for disambiguations based strictly on 5 word windows centered at various distances (shown on the horizontal axis). Gale's observation was tested only on nouns; our experiment also shows that reasonably accurate decisions may be made for nouns using exclusively remote context. Our results in this case are based on test sets with equal numbers of the two senses. Hence chance performance is at 50%. However, when tested on verbs and adjectives, precision drops off with a much steeper slope as the distance from the ambiguous word increases. This would indicate that approaches giving equal weight to all positions in a broad window of context may be less well-suited for handling verbs and adjectives. Models which give greater weight to immediate context would seem more appropriate in these circumstances.

A similar experiment was applied to function words, and the dropoff beyond strictly immediate context was precipitous, converging at near chance performance for distances greater than 5. However, function words did appear to have predictive power of roughly 5% greater than chance in directly adjacent positions. The effect was greatest for verbs, where the function word to the right (typically a preposition or particle) served to disambiguate at a precision of 13% above chance. This would indicate that methods which exclude function words from models to minimize noise should consider their inclusion, but only for restricted local positions.

6.3. Comparison of Sense Definitions

Results for the 5 different definitions of sense ambiguity studied here are similar. However they tend to fluctuate relative to each other across experiments, and there appears to be no consistent ordering of the mean entropy of the different types of sense distributions. Because of the very large number of permutations considered, it is not possible to give a full breakdown of the differences, and such a breakdown does not appear to be terribly informative. The important observation, however, is that the basic conclusions drawn from this paper hold for *each* of the sense definitions considered, and hence corroborate and strengthen the conclusions which can be drawn from any one.

6.4. Performance Given Little Evidence

One of the most striking conclusions to emerge from this study is that for the local collocations considered, decisions based on a single data point are highly reliable. Normally one would consider a 1/0 sense distribution in a 3944 sample training set to be noise, with performance based on this information not

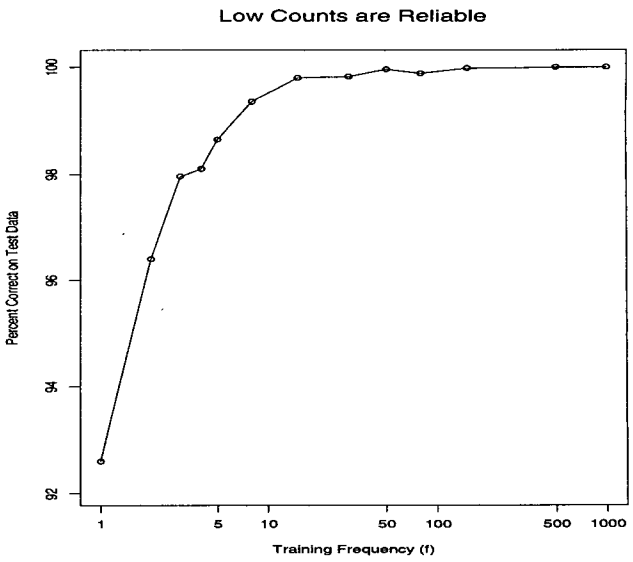

Figure 2: Percentage correct for disambiguations based solely on a single *content-word-to-the-right* collocation seen f times in the training data without counter-examples.

likely to much exceed the 69% prior probability expected by chance. But this is not what we observe. For example, when tested on the *word-to-the-right* collocation, disambiguations based solely on a single data point exceed 92% accuracy, and performance on 2/0 and 3/0 distributions climb rapidly from there, and reach nearly perfect accuracy for training samples as small as 15/0, as shown in Figure 2. In contrast, a collocation 30 words away which also exhibits a 1/0 sense distribution has a predictive value of only 3% greater than chance. This difference in the reliability of low frequency data from local and wide context will have implications for algorithm design.

7. APPLICATIONS

7.1. Training Set Creation and Verification

This last observation has relevance for new data set creation and correction. Collocations with an ambiguous content word which have frequency greater than 10-15 and which do not belong exclusively to one sense should be flagged for human reinspection, as they are most likely in error. One can speed the sense tagging process by computing the most frequent collocates, and for each one assigning all examples to the same sense. For the data in Table 2, this will apparently fail for the *foreign Aid/Aide* example in 1 out of 719 instances (still 99.9% correct). However, in this example the model's classification was actually correct; the given usage was a misspelling in the 1992 AP Newswire: "Bush accelerated foreign *aide* and weapons sales to Iraq.". It is quite likely that if were indeed a foreign assistant being discussed, this example would also have another collocation (with the verb, for example),

which would indicate the correct sense. Such inconsistencies should also be flagged for human supervision. Working from the most to least frequent collocates in this manner, one can use previously tagged collocates to automatically suggest the classification of other words appearing in different collocation types for those tagged examples. The one sense per discourse constraint can be used to refine this process further. We are working on a similar use of these two constraints for unsupervised sense clustering.

7.2. Algorithm Design

Our results also have implications for algorithm design. For the large number of current approaches which treat wide context as an unordered bag of words, it may be beneficial to model certain local collocations separately. We have shown that reliability of collocational evidence differs considerably between local and distant context, especially for verbs and adjectives. If one one is interested in providing a probability with an answer, modeling local collocations separately will improve the probability estimates and reduce cross entropy.

Another reason for modeling local collocations separately is that this will allow the reliable inclusion of evidence with very low frequency counts. Evidence with observed frequency distributions of 1/0 typically constitute on the order of 50% of all available evidence types, yet in a wide context window this low frequency evidence is effectively noise, with predictive power little better than chance. However, in very local collocations, single data points carry considerable information, and when used alone can achieve precision in excess of 92%. Their inclusion should improve system recall, with a much-reduced danger of overmodeling the data.

7.3. Building a Full Disambiguation System

Finally, one may ask to what extent can local collocational evidence alone support a practical sense disambiguation algorithm. As shown in Table 3, our models of single collocation types achieve high precision, but individually their applicability is limited. However, if we combine these models as described in Section 5, and use an additional function word collocation model when no other evidence is available, we achieve full coverage at a precision of 92%. This result is comparable to those previously reported in the literature using wider context of up to 50 words away [5,6,7,12]. Due to the large number of variables involved, we shall not attempt to compare these directly. Our results are encouraging, however, and and we plan to conduct a more formal comparison of the "bag of words" approaches relative to our separate modeling of local collocation types. We will also consider additional collocation types covering a wider range of syntactic relationships. In addition, we hope to incorporate class-based techniques, such as the modeling of verb-argument selectional preferences [Resnik, 1992], as a mechanism for achieving improved performance on unfamiliar collocations.

8. CONCLUSION

This paper has examined some of the basic distributional properties of lexical ambiguity in the English language. Our experiments have shown that for several definitions of sense and collocation, an ambiguous word has only one sense in a given collocation with a probability of 90-99%. We showed how this claim is influenced by part-of-speech, distance, and sample frequency. We discussed the implications of these results for data set creation and algorithm design, identifying potential weaknesses in the common "bag of words" approach to disambiguation. Finally, we showed that models of local collocation can be combined in a disambiguation algorithm that achieves overall precision of 92%.

References

1. Bahl, L., P. Brown, P. de Souza, R. Mercer, "A Tree-Based Statistical Language Model for Natural Language Speech Recognition," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37, 1989.
2. Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer, "Word Sense Disambiguation using Statistical Methods," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp 264-270.
3. Gale, W., K. Church, and D. Yarowsky, "One Sense Per Discourse," *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, 1992.
4. Gale, W., K. Church, and D. Yarowsky, "On Evaluation of Word-Sense Disambiguation Systems," in *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, 1992b.
5. Gale, W., K. Church, and D. Yarowsky, "A Method for Disambiguating Word Senses in a Large Corpus," in *Computers and the Humanities*, 1993.
6. Hearst, Marti, "Noun Homograph Disambiguation Using Local Context in Large Text Corpora," in *Using Corpora*, University of Waterloo, Waterloo, Ontario, 1991.
7. Leacock, Claudia, Geoffrey Towell and Ellen Voorhees "Corpus-Based Statistical Sense Resolution," in *Proceedings, ARPA Human Language Technology Workshop*, 1993.
8. Kelly, Edward, and Phillip Stone, *Computer Recognition of English Word Senses*, North-Holland, Amsterdam, 1975.
9. Resnik, Philip, "A Class-based Approach to Lexical Discovery," in *Proceedings of 30th Annual Meeting of the Association for Computational Linguistics*, 1992.
10. Rivest, R. L., "Learning Decision Lists," in *Machine Learning*, 2, 1987, pp 229-246.
11. Sproat, R., J. Hirschberg and D. Yarowsky "A Corpus-based Synthesizer," in *Proceedings, International Conference on Spoken Language Processing*, Banff, Alberta. October 1992.
12. Yarowsky, David "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," in *Proceedings, COLING-92*, Nantes, France, 1992.