Using Bilingual Materials to Develop Word Sense Disambiguation Methods

William A. Gale
Kenneth W. Church
David Yarowsky

AT&T Bell Laboratories
600 Mountain Avenue
P. O. Box 636
Murray Hill NJ, 07974-0636

Abstract

Word sense disambiguation has been recognized as a major problem in natural language processing research for over forty years. Much of this work has been stymied by difficulties in acquiring appropriate lexical resources, such as semantic networks and annotated corpora. Following the suggestion in Brown *et al.* (1991a) and Dagan *et al.* (1991), we have achieved considerable progress recently by taking advantage of a new source of testing and training materials. Rather than depending on small amounts of hand-labeled text, we have been making use of relatively large amounts of parallel text, text such as the Canadian Hansards (parliamentary debates), which are available in two (or more) languages. The translation can often be used in lieu of hand-labeling. For example, consider the polysemous word *sentence*, which has two major senses: (1) a judicial sentence, and (2), a syntactic sentence. We can collect a number of sense (1) examples by extracting instances that are translated as *peine*, and we can collect a number of sense (2) examples by extracting instances that are translated as *phrase*. In this way, we have been able to acquire a considerable amount of testing and training material for developing and testing our disambiguation algorithms.

The availability of this testing and training material has enabled us to develop quantitative disambiguation methods that achieve 90% accuracy in discriminating between two very distinct senses of a noun such as *sentence*. In the training phase, we collect a number of instances of each sense of the polysemous noun. Then in the testing phase, we are given a new instance of the noun, and are asked to assign the instance to one of the senses. We attempt to answer this question by comparing the context of the unknown instance with contexts of known instances using a Bayesian argument that has been applied successfully in related applications such as author identification and information retrieval.

The final section of the paper will describe a number of methodological studies which show that the training set need not be large and that it need not be free from errors. Perhaps most surprisingly, we find that the context should extend ±50 words, an order of magnitude larger than one typically finds in the literature.

1. Word-Sense Disambiguation

Consider, for example, the word *duty* which has at least two quite distinct senses: (1) a tax and (2) an obligation. Three examples of each sense are given in Table 1 below. The classic disambiguation problem is to construct a means for discriminating between two or more sets of examples such as those shown in Table 1. This paper will focus on the methodology required to address the classic problem, and will have less to say about the details required for practical application of this methodology. Consequently, the reader should exercise some caution in interpreting the 90% figure reported here; this figure could easily be swamped out in a practical system by any number of factors that go beyond the scope of this paper. In particular, the Canadian Hansards, one of just the few currently available sources of parallel text, is extremely unbalanced, and is therefore severely limited as a basis for a practical disambiguation system.

Table 1: Sample Concordances of *duty* (split into two senses)

Sense	Examples (from Canadian Hansards)
tax	fewer cases of companies paying *duty* and then claiming a refund
	and impose a countervailing *duty* of 29,1 per cent on candian exports of
	the united states imposed a *duty* on canadian saltfish last year
obligation	it is my honour and *duty* to present a petition duly approved
	working well beyond the call of *duty* ? SENT i know what time they start
	in addition , it is my *duty* to present the government 's comments

Moreover, it is important to distinguish the monolingual word-sense disambiguation problem from the translation issue. It is not always necessary to resolve the word-sense ambiguity in order to translate a polysemous word. Especially in related languages like English and French, it is common for word-sense ambiguity to be preserved in both languages. For example, both the English noun *interest* and the French equivalent *interêt* are multiply ambiguous in both languages in more or less the same ways. Thus, one cannot turn to the French to resolve the ambiguity in the English, since the word is equally ambiguous in both languages.

Furthermore, when one word does translate to two (e.g., *sentence* → *peine* and *phrase*), the choice of target translation need not indicate a sense split in the source. Consider, for example, the group of Japanese words translated by ''wearing clothes'' in English. While the Japanese have five different words for ''wear'' depending on which part of the body is involved, we doubt that English speakers would ever sort ''wearing shoes'' and ''wearing shirt'' into separate categories.

These examples indicate that word-sense disambiguation and translation are somewhat different problems. It would have been nice if the translation could always be used in lieu of hand-tagging to resolve the word-sense ambiguity but unfortunately, this is not the case. Nevertheless, the translation is often helpful for resolving the ambiguity. It seems to us to make sense to continue to use the Hansard translations to develop the discrimination methodology, while we continue to seek more appropriate sources of testing and training materials. See Yarowsky (1992) for an application of the methods developed here to a somewhat more appropriate source, a combination of the Roget's Thesaurus (Chapman, 1977)[1] and the Grolier's Encyclopedia (1991).

2. Knowledge Acquisition Bottleneck

In our view, the crux of the problem in developing methods for word sense disambiguation is to find a strategy for acquiring a sufficiently large set of training material. We think that we have found such a strategy by turning to parallel text as a source of testing and training materials. Most of the previous work falls into one of three camps: (1) Qualitative Methods, e.g., Hirst (1987), (2) Dictionary-based Methods, e.g., Lesk (1986), and (3) Hand Annotated Corpora, e.g., Kelly and Stone (1975). In each case, the work has been limited by knowledge acquisition bottleneck.

2.1 Qualitative Methods

For example, there has been a tradition in parts of the AI community of building large experts by hand, e.g., Granger (1977), Rieger (1977), Small and Rieger (1982), Hirst (1987). Unfortunately, this approach is not very easy to scale up, as many researchers have observed: ''The expert for THROW is currently six pages long, ... but it should be 10 times that size'' (Small and Reiger, 1982). Since this approach is so difficult to scale up, much of the work has had to focus on ''toy'' domains (e.g., Winograd's Blocks World) or sublanguages (e.g., Isabelle (1984), Hirschman (1986)). Currently, it is not possible to find a semantic network with the kind of broad coverage that would be required for unrestricted text.

1. This thesaurus should not be confused with the much smaller and less up-to-date 1911 edition of Roget's.

From an AI point of view, it appears that the word-sense disambiguation problem is ‘‘AI-Complete,’’ meaning that you can’t solve this problem until you’ve solved all of the other hard problems in AI. Since this is unlikely to happen any time soon (if at all), it would seem to suggest that word-sense disambiguation is just too hard a problem, and we should spend our time working on a simpler problem where we have a good chance of making progress. Rather than accept this rather pessimistic conclusion, we prefer to reject the premise and search for an alternative point of view.

2.2 Machine-Readable Dictionaries (MRDs)

Others such as Lesk (1986), Walker (1987), Ide and Veronis (1990) have turned to machine-readable dictionaries (MRD) such as Oxford’s Advanced Learner’s Dictionary of Current English (OALDCE) in the hope that MRDs might provide a way out of the knowledge acquisition bottleneck. These researchers seek to develop a program that could read an arbitrary text and tag each word in the text with a pointer to a particular sense number in a particular dictionary.

Unfortunately, the approach doesn’t seem to work as well as one might hope. Lesk (1986) reports accuracies of 50-70% on short samples of *Pride and Prejudice*. Part of the problem may be that dictionary definitions are too short to mention all of the collocations (words that are often found in the context of a particular sense of a polysemous word). In addition, dictionaries have much less coverage than one might have expected. Walker (1987) reports that perhaps half of the words occurring in a new text cannot be related to a dictionary entry.

Thus, like the AI approach, the dictionary-based approach is also limited by the knowledge acquisition bottleneck; dictionaries simply don’t record enough of the relevant information, and much of the information that is stored in the dictionary is not in a format that computers can easily digest, at least at present.

2.3 Approaches Based on Hand-Annotated Corpora

A third line of research makes use of hand-annotated corpora. Most of these studies are limited by the availability of hand-annotated text. Since it is unlikely that such text will be available in large quantities for most of the polysemous words in the vocabulary, there are serious questions about how such an approach could be scaled up to handle unrestricted text. Nevertheless, we are extremely sympathetic with the basic approach, and will adopt a very similar strategy ourselves. However, we will introduce one important difference, the use of parallel text in lieu of hand-annotated text, as suggested by Brown *et al.* (1991a), Dagan *et al.* (1991) and others.

Kelly and Stone (1975) constructed 1815 disambiguation models by hand, selecting words with a frequency of at least 20 in a half million word corpus. Most subsequent work has sought automatic methods because it is quite labor intensive to construct these rules by hand. Weiss (1973) first built rule sets by hand for five words, then developed automatic procedures for building similar rule sets, which he applied to additional three words. Unfortunately, the system was tested on the training set, so it is difficult to know how well it actually worked.

Black (1987, 1988) studied five 4-way polysemous words using about 2000 hand tagged concordance lines for each word. Using 1500 training examples for each word, his program constructed decision trees based on the presence or absence of 81 ‘‘contextual categories’’ within the context[2] of the ambiguous word. He used three different types of contextual categories: (1) subject categories from LDOCE, the Longman Dictionary of Contemporary English (Longman, 1978), (2) the 41 vocabulary items occurring most frequently within two words of the ambiguous word, and (3) the 40 vocabulary items excluding function words occurring most frequently in the concordance line. Black found that the dictionary categories

2. The context was defined to be the concordance line, which we estimate to be about ± 6 words from the ambiguous word, given that his 2000 concordance lines contained about 26,000 words.

produced the weakest performance (47 percent correct), while the other two were quite close at 72 and 75 percent correct, respectively.

There has recently been a flurry of interest in approaches based on hand-annotated corpora. Hearst (1991) is a very recent example of an approach somewhat like Black (1987, 1988), Weiss (1973) and Kelly and Stone (1975), in this respect, though she makes use of considerably more syntactic information than the others. Her performance also seems to be somewhat better than the others', though it is difficult to compare performance across systems.

3. An Information Retrieval (IR) Approach to Sense Disambiguation

We have been experimenting with an Information Retrieval approach to sense disambiguation. In the training phase, we collect a number of instances of *sentence* that are translated as *peine*, and a number of instances of *sentence* uses that are translated as *phrase*. Then in the testing phase, we are given a new instance of *sentence*, and are asked to assign the instance to one of the two senses. We attempt to answer this question by comparing the context of the unknown instance with contexts of known instances.

Basically we are treating contexts as analogous to documents in an information retrieval setting. Just as the probabilistic retrieval model (van Rijsbergen, 1979, chapter 6; Salton, 1989, section 10.3) sorts documents d by

$$score(d) = \prod_{token \text{ in } d} \frac{Pr(token|rel)}{Pr(token|irrel)}$$

we will sort contexts c by

$$score(c) = \prod_{token \text{ in } c} \frac{Pr(token|sense_1)}{Pr(token|sense_2)}$$

where $Pr(token|sense)$ is an estimate of the probability that *token* appears in the context of $sense_1$ or $sense_2$. Contexts are defined to extend 50 words to the left and 50 words to the right of the polysemous word in question for reasons that will be discussed in section 5. This model ignores a number of important linguistic factors such as word order and collocations (correlations among words in the context). Nevertheless, there are $2V \approx 200{,}000$ parameters in the model. It is a non-trivial task to estimate such a large number of parameters, especially given the sparseness of the training data. The training material typically consists of approximately 12,000 words of text (100 words words of context for 60 instances of each of two senses). Thus, there are more than 15 parameters to be estimated from for each data point. Clearly, we need to be fairly careful given that we have so many parameters and so little evidence.

3.1 Using Global Probabilities to Smooth the Local Probabilities

In principle, the conditional probabilities, $Pr(tok|sense)$, can be estimated by selecting those parts of the entire corpus which satisfy the required conditions (e.g., 100-word contexts surrounding instances of one sense of *duty*), counting the frequency of each word, and dividing the counts by the total number of words satisfying the conditions. However, this estimate, which is known as the maximum likelihood estimate (MLE), has a number of well-known problems. In particular, it will assign zero probability to words that do not happen to appear in the sample. Zero is not only a biased estimate of their true probability, but it is also unusable for the sense disambiguation task.

In order to avoid these problems, we have decided to use information from the entire corpus in addition to information from the conditional sample in order. We will estimate $Pr(tok|sense)$ by interpolating between local probabilities computed within the 100-word context and global probabilities computed over the entire corpus $Pr(tok)$. The local probabilities are more relevant and the global probabilities are better measured. We seek a trade-off between random measurement errors and bias errors. This is accomplished by estimating the relevance of the larger corpus to the conditional sample in order to find the optimal trade off between random error and bias. See Gale *et al.* (to appear) for further details.

3.2 An Example

Table 2 (below) gives a sense of what the interpolation procedure does for some of the words that play an important role in disambiguating between the two senses of *duty* in the Canadian Hansards. Table 2 lists the 15 words with the largest product (shown as the first column) of the model score (the second column) and the frequency in the 6000 word training corpus (the third column). The conditioned samples are obtained by extracting a 100-word window surrounding each of the 60 training examples. The training sets were selected by randomly sampling instances of *duty* in the Hansards until 60 instances were found that were translated as *droit* and 60 instances were found that were translated as *devoir*. The first set of 60 are used to construct the model for the tax sense of *duty* and the second set of 60 are used to construct the model for the obligation sense of *duty*.

The column labeled ''freq'' shows the number of times that each word appeared in the conditioned sample. For example, the count of 50 for the word *countervailing* indicates that *countervailing* appeared 50 times within the conditioned sample. This is a remarkable fact, given that *countervailing* is a fairly unusual word. It is much less surprising to find a common word like *to* appearing quite often (228 times) in the other conditioned sample. The second column (labeled ''weight'') models the fact that 50 instances of *countervailing* are more surprising than 228 instances of *to*. The weights for a word are its log likelihood in the conditioned sample compared with its likelihood in the global corpus. The first column, the product of these log likelihoods and the frequencies, is a measure of the importance, in the training set, of the word for determining which sense the training examples belong to. Note that words with large scores do seem to intuitively distinguish the two senses, at least in the Canadian Hansards.

There are obviously some biases introduced by the unusual nature of this corpus, which is hardly a balanced sample of general language. For example, the set of words listed in Table 2 under the obligation sense of *duty* is heavily influenced by the fact that the Hansards contain a fair amount of boilerplate of the form: ''Mr. speaker, pursuant to standing order..., I have the honour and duty to present petitions duly signed by... of my electors....''

Table 2: Selected Portions of the Two Models for the Two Senses of *duty*

tax sense of *duty*				obligation sense of *duty*			
weight*freq	weight	freq	word	weight*freq	weight	freq	word
285	5.7	50	countervailing	64	3.2	20	petitions
111.8	4.3	26	duties	59.28	0.26	228	to
99.9	2.7	37	u.s	56.28	0.42	134	|
73.1	1.7	43	trade	51	3	17	petition
70.2	1.8	39	states	47.6	2.8	17	pursuant
69.3	3.3	21	duty	46.28	0.52	89	mr
68.4	3.6	19	softwood	37.8	2.7	14	honour
68.4	1.9	36	united	37.8	1.4	27	order
58.8	8.4	7	rescinds	36	2	18	present
54	3	18	lumber	33.6	2.8	12	proceedings
50.4	4.2	12	shingles	31.5	3.5	9	prescription
50.4	4.2	12	shakes	31.32	0.87	36	house
46.8	3.6	13	35	29.7	3.3	9	reject
46.2	2.1	22	against	29.4	4.2	7	boundaries
41.8	1.1	38	canadian	28.7	4.1	7	electoral

4. Materials

4.1 Six Polysemous Words

We will focus on six polysemous words: *duty, drug, land, language, position* and *sentence*. Table 3 (below) shows the six English nouns, along with two French translations. The penultimate column shows the number of times that each English noun was found with the particular French translation in the corpus, while the final column shows the accuracy of the system in identifying the appropriate sense based on the

Table 3: Six Polysemous Words				
English	French	sense	N	% correct
duty	droit	tax	1114	97
	devoir	obligation	691	84
drug	médicament	medical	2992	84
	drogue	illicit	855	97
land	terre	property	1022	86
	pays	country	386	89
language	langue	medium	3710	90
	langage	style	170	91
position	position	place	5177	82
	poste	job	577	86
sentence	peine	judicial	296	97
	phrase	grammatical	148	100

context of use. We selected these nouns because they could be disambiguated by looking at their French translation in the Canadian Hansards. As mentioned above, the polysemous noun *interest*, for example, would not meet this constraint because the French target *interêt* is just as ambiguous as the English source.

In addition, for studying methodological questions, it is important that there be an adequate number of instances of both translations in corpus, though this constraint can be relaxed in a practical application as we will see. Unless stated otherwise, the studies to be reported here all use 60 instances of the six polysemous words in each of the two main senses for training $Pr(tok|sense)$. An additional 90 instances of each word in each sense are used for testing. Consequently, we require a total of 150 (60 + 90) instances of each word in each sense in order to investigate the methodological issues.

4.2 Sentence Alignment and Word Correspondence

In order to collect the testing and training sets, we need to know the ‘‘truth.’’ As mentioned above, we approximate the ‘‘truth’’ by assuming that the French translation in the Hansards is adequate for our purposes. The process of identifying the French translation is a two step procedure. As in Brown, Lai, and Mercer (1991b), we begin by aligning the parallel texts at the sentence level (Gale and Church, 1991a). In our experience, 90% of the English sentences match exactly one French sentence, but other possibilities, especially two sentences matching one (2-1) or one matching two (1-2), are not uncommon. The method correctly aligned all but 4% of the regions. Moreover, by selecting the best scoring 80% of the corpus, the error rate dropped to 0.7%. See (Gale and Church, 1991a) for more details on the method and its evaluation.

After the sentences have been aligned, we can then identify the French correspondences using a very simple set of programs designed by Yarowsky. Gale and Church (1991b) describe a more elaborate program that attempts to find correspondences for most of the words in the English text, not just the polysemous words of interest. However, for our purposes here, the more elaborate methods proved unnecessary.

5. Studies of Methodological Questions

5.1 How Much Context Should We Use?

As mentioned above, we use a very wide context, 100-words surrounding the polysemous word in question. Most previous studies have limited themselves to a much narrower notion of context, perhaps only 5 words to the left and 5 words to the right of the polysemous word, based on the observation that people don’t seem to need very much context (Kaplan, 1950; Choueka and Lusignan, 1985). Although people may be able to get by without the additional context, we find that there are often very useful clues even quite far away from the polysemous word in question. Figure 1 shows that information is *measurable* out to 10,000 words away from the polysemous word, and Figure 2 shows that this information is *useful* out to 50 words. Since the disambiguation problem is as difficult for the machine as it is, we believe that it would be a mistake to ignore this information just because people don’t seem to need it. As in computer chess, it is not always

Contextual Clues are Measurable Out to 10,000 Words

Figure 1. The horizontal axis *d* shows the size of the context as a distance (in words) from the polysemous word in question. The vertical scale shows disambiguation performance (percent correct), computed over a context of ten words at the distance specified by the horizontal axis: [–d–5,–d] and [d, d+5],. Vertical lines indicate means and standard deviations computed over a group of 90 instances times six polysemous words times two senses. Note that performance remains significantly above chance (50%) out to 10,000 words away from the polysemous word.

best for the computer to try to copy human strategies.

Figure 1 demonstrates that the information is measurable at very large distances from the polysemous words. In order to show this, we selected a very unusual context ([–d–5,–d] and [d, d+5]) and measured performance as a function of *d*. This experiment thus asks, if you did not know any of the intervening words, would ten words at distance *d* be sufficient for disambiguation? The answer is ''yes'' for for $d < 10,000$, at least in the Hansards. We found this result surprising, given that almost all previous disambiguation studies have concentrated so heavily on very narrow contexts. The result almost certainly has something to do with discourse structure, and may depend fairly strongly on the average length of an average debate. In addition, the result may depend on other factors such as part of speech.

Although Figure 1 shows that contextual clues are measurable at surprisingly large distances, much of this information might not be very useful. In particular, it might have been possible to find the same information at smaller distances. Figure 2 attempts to address this concern by examining the *marginal* contribution of context as a function of distance. Figure 2 is computed just like Figure 1, except that Figure 2 uses a $2d-1$ word context ([–d,–1] and [1,d]), rather than the 10 word context in Figure 1. This experiment thus asks, given that you know all the words out to *d*, what is the value of a few additional words further out? The contribution is largest, not surprisingly, for smaller *d*, but nevertheless, the contribution continues to grow out to at least twenty words, perhaps fifty words, well beyond the ±6 word contexts typically found in many disambiguation studies. Increasing the context from ±6 words to ±50 words improves performance from 86% to 90%.

Contextual Clues are Useful Out to 50 Words

Figure 2. The horizontal axis shows the size of the context, as in Figure 1. The vertical axis shows performance (percent correct), computed over a context of $2d-1$ words, from $-d$ to d (but excluding 0). Note that performance rises very rapidly at first and reaches an asymptote at about 50 words.

5.2 Quantity and Quality of Training Material

In a practical application, we might be concerned that the Bayesian discrimination methods would be too demanding on the quantity and quality of the training material. This section will consider the quantity question first and then return to the quality question.

As mentioned above, the method would have limited applicability if it requires unreasonably large training sets. We expect performance to degrade with the size of the training set, and we would like to control this source of variability as we study other factors. In addition, we would also like to be able to predict performance when the size of the training set is severely constrained, because this is usually the case for most senses of most words. Figure 3 shows performance as a function of the size of the training set. Note that very small training sets perform remarkably well; just 3 exemplars are sufficient to achieve 75%. Nevertheless, it helps to use larger training sets, up to about 50 or 60 exemplars when performance reaches asymptote.

The quality of the training set is another potential source of concern. If training materials are to be collected on a large scale, then we will need to accept a certain number of errors. Moreover, if the method is robust to errors, then it will be possible to consider bootstrapping methods that might be able to speed up the data collection effort.

Small Training Sets Perform Surprisingly Well

Figure 3. The horizontal axis shows the size of the training set, while the vertical scale shows performance (percent correct, averaged over 90 instances times two senses times six polysemous words). Note that very small training sets perform remarkably well; just 3 exemplars are sufficient to achieve 75%. Nevertheless, it helps to use larger training sets, up to about 50 or 60 exemplars when performance reaches asymptote.

In order to study the quality issue, we deliberately introduced a variable number of errors into the training set. Table 4 shows the mean performance (percent correct), averaged over 90 instances times two senses times six polysemous words, as a function of the quality of the training set (the fraction of errors deliberately introduced into the training set) and coverage (the fraction of the test set with the largest discrimination score).

Two observations on this table are important. First, at 10 percent errors input, the output errors have only increased from 10 percent to 12 percent. Thus we can accommodate up to about ten percent errors with little degradation of performance. Second, at fifty percent coverage, input errors of twenty to thirty percent result in about half as many errors on output. Therefore if one had obtained a set of examples with no more than twenty to thirty percent, one could iterate example selection just once or twice and have example sets that had less than ten percent errors.

Table 4: Robust to Errors in Training Set

coverage	quality			
	0%	10%	20%	30%
50%	98.6	96	92	84
80%	93	92	88	81
100%	90	88	82	78

5.3 Recall vs. Precision

One can make a tradeoff between coverage (fraction of words for which a disambiguation is attempted) and accuracy (fraction of attempted words which are correctly disambiguated), analogous to the recall versus precision tradeoff in Information Retrieval. Figure 4 shows the error rate depends very strongly on coverage, and consequently, it is important to state coverage carefully when reporting performance. At 66 percent coverage, the error rate is about a quarter of its value at 100 percent coverage. (Some of the literature has not been as careful as it might be in this respect.)

Error Rate Depends on Coverage

Figure 4. The horizontal axis shows coverage while the vertical scale shows error rate. Note that the error rate increases very quickly with small increases in coverage, because most of the errors result when there is very little information to go on.

6. Conclusions

Difficulties in acquiring suitable testing and training materials have deterred progress on word-sense disambiguation over the past forty years. We have achieved considerable progress recently by taking advantage of a new source of testing and training materials. Rather than depending on small amounts of hand-labeled text, we have been making use of relatively large amounts of parallel text, text such as the Canadian Hansards, which are available in multiple languages. Consider, for example, the polysemous word *drug*, which has two major senses: (1) a medical drug, and (2), an illicit drug. We can collect a number of examples of each sense by using the French translation (*méd*i*cament* vs. *drogue*) as an indicator of the sense of the English word. In this way, we have been able to acquire considerable amounts of testing and training material for study of quantitative methods.

The availability of this testing and training material has enabled us to develop quantitative disambiguation methods that achieve 90% accuracy in discriminating between two senses corresponding to different topics, based on context alone. In addition, an perhaps more importantly, the availability of this testing and

training materials has allowed us to carry out a number of methodological studies. In particular, we find that a training set of as few as ten exemplars seems to be quite useful, and that the training set can tolerate a fair number of errors. The most surprising result, perhaps, is that the width of context should be ± 50 words, an order of magnitude larger than one normally finds in the literature.

References

1. Bar-Hillel (1960), ‘‘Automatic Translation of Languages,’’ in *Advances in Computers*, Donald Booth and R. E. Meagher, eds., Academic, New York.
2. Black, Ezra (1987), *Towards Computational Discrimination of English Word Senses*, Ph. D. thesis, City University of New York.
3. Black, Ezra (1988), ‘‘An Experiment in Computational Discrimination of English Word Senses,’’ *IBM Journal of Research and Development*, v 32, pp 185-194.
4. Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer (1991a), ‘‘Word Sense Disambiguation using Statistical Methods,’’ *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp 264-270.
5. Brown, Peter, Jennifer Lai, and Robert Mercer (1991b) ‘‘Aligning Sentences in Parallel Corpora,’’ *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp 169-176.
6. Chapman, Robert (1977). *Roget’s International Thesaurus (Fourth Edition)*, Harper and Row, NY.
7. Choueka, Yaacov, and Serge Lusignan (1985), ‘‘Disambiguation by Short Contexts,’’ *Computers and the Humanities*, v 19. pp. 147-158.
8. Church, Kenneth (1989), ‘‘A Stochastic Parts Program an Noun Phrase Parser for Unrestricted Text,’’ *Proceeding, IEEE International Conference on Acoustics, Speech and Signal Processing*, Glasgow.
9. Cruse, D. A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge, England.
10. Dagan, Ido, Alon Itai, and Ulrike Schwall (1991), ‘‘Two Languages are more Informative than One,’’ *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp 130-137.
11. Fillmore, Charles, and Sue Atkins, (1991) ‘‘Word Meaning: Starting where MRD’s Stop,’’ invited talk at the 29th Annual Meeting of the Association for Computational Linguistics.
12. Gale, W., and Church, K. (1991a) ‘‘A Program for Aligning Sentences in Bilingual Corpora,’’ *Association for Computational Linguistics*.
13. Gale, W., and Church, K. (1991b) ‘‘Identifying Word Correspondences in Parallel Text,’’ *Fourth Darpa Workshop on Speech and Natural Language*, Asilomar.
14. Gale, W., K. Church, and D. Yarowsky, (to appear) ‘‘A Method for Disambiguating Word Senses in a Large Corpus,’’ *Computers and the Humanities*.
15. Granger, Richard (1977), ‘‘*FOUL-UP* A program that figures out meanings of words from context,’’ *IJCAII-77*, pp. 172-178.
16. Grolier’s Inc. (1991) *New Grolier’s Electronic Encyclopedia*.
17. Hearst, Marti (1991) ‘‘Toward Noun Homonym 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*, available from the UW Centre for the New OED and Text Research, University of Waterloo, Waterloo, Ontario, Canada.
18. Hirschman, Lynette (1986), ‘‘Discovering Sublanguage Discovery,’’ in *Analyzing Language in Restricted Domains*, Ralph Grishman and Richard Kittredge, eds., Lawrence Erlbaum, Hillsdale, New Jersey.
19. Hirst, G. (1987), *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge.
20. Ide, N. and Veronis, J. (1990) ‘‘Mapping Dictionaries: A Spreading Activation Approach,’’ in *Proceedings of the Sixth Annual Conference of the UW Centre for the OED and Text Research*, available from the UW Centre for the New OED and Text Research, University of Waterloo, Waterloo, Ontario, Canada.
21. Isabelle, P. (1984) ‘‘Machine Translation at the TAUM Group,’’ in King, M. (ed.) *Machine Translation Today: The State of the Art*, Edinburgh University Press.

22. Jackson, Howard (1988) *Words and their Meaning*, Longman, London.

23. Jacobs, Paul, George Krupka, Susan McRoy, Lisa Rau, Norman Sondheimer, and Uri Zernik (1990), ‘‘Generic Text Processing: A Progress Report,’’ *Proceedings DARPA Speech and Natural Language Workshop*, pp. 359-364.

24. Kaplan, Abraham (1950), ‘‘An Experimental Study of Ambiguity in Context,’’ cited in *Mechanical Translation*, v. 1, nos. 1-3.

25. Kelly, Edward, and Phillip Stone (1975), *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.

26. Kucera, H., and W. Francis (1967), *Computational Analysis of Present-day American English*, Brown University Press, Providence.

27. Lesk, Michael (1986), ‘‘Automatic Sense Disambiguation: How to tell a Pine Cone from an Ice Cream Cone,’’ *Proceeding of the 1986 SIGDOC Conference*, Association for Computing Machinery, New York.

28. Longman Group Limited, eds. (1978), *Longman Dictionary of Contemporary English*, Longman, Burnt Mill, England.

29. Masterson, Margaret (1967), ‘‘Mechanical Pidgin Translation,’’ in *Machine Translation*, Donald Booth, ed., Wiley, 1967.

30. Mosteller, Fredrick, and David Wallace (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts.

31. Quine, W. v. O. (1960), *Word and Object*, MIT Press, Cambridge.

32. Reiger, Charles (1977), ‘‘Viewing Parsing as Word Sense Discrimination,’’ in *A Survey of Linguistic Science*, W. Dingall, ed., Greylock.

33. Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley.

34. Sinclair, J., Hanks, P., Fox, G., Moon, R., Stock, P. et al. (eds.) (1987) *Collins Cobuild English Language Dictionary*, Collins, London and Glasgow.

35. Small, S. and C. Rieger (1982), ‘‘Parsing and Comprehending with Word Experts (A Theory and its Realization),’’ in *Strategies for Natural Language Processing*, W. Lehnert and M. Ringle, eds., Lawrence Erlbaum Associates, Hillsdale, NJ.

36. Stone, Phillip, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie (1966), *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, Cambridge.

37. van Rijsbergen, C. (1979) *Information Retrieval*, Second Editional, Butterworths, London.

38. Walker, Donald (1987), ‘‘Knowledge Resource Tools for Accessing Large Text Files,’’ in *Machine Translation: Theoretical and Methodological Issues*, Sergei Nirenberg, ed., Cambridge University Press, Cambridge, England.

39. Weinreich, U. (1980), *On Semantics*, University of Pennsylvania Press, Philadelphia.

40. Weiss, Stephen (1973), ‘‘Learning to Disambiguate,’’ *Information Storage and Retrival*, v. 9, pp 33-41.

41. Yarowsky, David (1992), ‘‘Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora,’’ *Proceedings COLING-92*.

42. Yngve, Victor (1955), ‘‘Syntax and the Problem of Multiple Meaning,’’ in *Machine Translation of Languages*, William Locke and Donald Booth, eds., Wiley, New York.

43. Zernik, Uri (1990) ‘‘Tagging Word Senses in Corpus: The Needle in the Haystack Revisited,’’ in *Text Based Intelligent Systems: Current Research in Text Analysis, Information Extraction, and Retrival*, P. Jacobs, ed., GE Research and Development Center, pp25-29.