Work on Statistical Methods for Word Sense Disambiguation

William A. Gale
Kenneth W. Church
David Yarowsky

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

1. Introduction: the Task

Our task is to tag nouns in a corpus with a "sense," one of a small list of tags for each noun. 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.

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

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) Discrimination Methods, e.g., Kelly and Stone (1975). In each case, the work has been limited by an inability to get sufficiently large sets of training material. See Gale, Church, and Yarowsky (1992) for a detailed review of the difficulties in extending each of these approaches.

Our work falls in the third group. These methods take sets of training examples as input and develop some means for discriminating the sets based on this training data. The bottleneck for this approach has been the acquisition of training material. Most previous work in this line has used hand tagged sets of examples. In our view, the crux of the problem in developing discrimination methods for word sense disambiguation has been to find a strategy for acquiring a sufficiently large sets of training material. We think that we have found two such strategies for acquiring testing and training materials, one of which we have discussed previously, but will review here, and another discussed for the first time.

Beyond the classical discrimination problem lie various problems in building a practical system, the most pressing of which is to limit the number of parameters: If there are about 10^5 senses to be discriminated, a strategy based on the direct approach of a model per sense will allow only a few parameters for each sense. Much of this paper addresses this problem.

2. A Bayesian Discrimination Approach to Sense Disambiguation

We have been experimenting with Bayesian Discrimination approaches to sense disambiguation. In the training phase, we collect a number of instances of *duty* from each of the two given classes and derive parameters from their contexts. Then in the testing phase, we are given a new instance of *duty*, 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. This approach, long familiar to

statisticians, may be best known to the linguistic community through its use in Information Retrieval. 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

$$score(d) = \prod_{token \text{ in } d} \frac{Pr(token|rel)}{Pr(token|irrel)} \quad [1]$$

we will sort contexts c by

$$score(c) = \prod_{token \text{ in } c} \frac{Pr(token|sense_1)}{Pr(token|sense_2)} \quad [2]$$

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 the next section.

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. Gale, Church, and Yarowsky (1992) describes the details of how we accomplish this.

3. Review of Our Previous Results

3.1 Bilingual training and testing material

Following the suggestion in Brown, Lai, and Mercer (1991) and Dagan, Itai, and Schwall (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 (e.g. Kelly and Stone, 1975), 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 key point is that the translation can often be used in lieu of hand-labeling. As such, it is suitable for both training and testing models. The examples from the two classes for *duty* given in Table 1 were selected by taking words translated as *droit* for the "tax" class and words translated as *devoir* for the "obligation" class. The following table shows six nouns we have studied in some detail.

Table 2: 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

The ability to test automatically on a substantial number of examples was useful for establishing a number

of methodological points. We refer the reader to Gale, Church, and Yarowsky (1992) for graphs and tables substantiating the following findings.

- Information is *measurable* out to 10,000 words away from the polysemous word
- This information is *useful* out to 50 words. This is the basis for our selection of context width.
- Very small training sets perform remarkably well; just 3 exemplars are sufficient to achieve 75% accuracy. Nevertheless, it helps to use larger training sets, up to about 50 or 60 exemplars when performance reaches asymptote.
- The method is robust to errors in the training set. Ten percent errors only degrades performance by 2%. When there are thirty percent errors in the training set, selecting the highest scoring half of all examples results in a set with about half as many errors as input, so iteration is possible.
- There is 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. At 75% coverage, the error rate was reduced by a factor of four.

3.2 Monolingual training material

At first, we thought that the method was completely dependent on the availability of parallel corpora for training. This has been a problem since parallel text remains somewhat difficult to obtain in large quantity, and what little is available is often fairly unbalanced and unrepresentative of general language. Also, the assumption that differences in translation correspond to differences in word-sense has always been somewhat suspect.

Furthermore, the straight-forward use of models built for each sense according to the methods described in the previous section would require about V models. The models in the previous section would require V parameters each, for a total of $V^2 \approx 10^{10}$ parameters. This would be unmanageable for a practical system.

Recently, Yarowsky (1991) has found a way to train on the Roget's Thesaurus and Grolier's Encyclopedia instead of the Hansards, thus circumventing many of the objections to our use of the Hansards. It also reduces the number of models needed from about V to about 10^3. Yarowsky's method inputs a 100-word context surrounding a polysemous word and scores each of the 1042 Roget Categories by:

$$\prod_{w \text{ in context}} Pr(w|Roget\ Category_i)$$

This may appear to be a particularly crude approximation, but as shown in the example below, it is surprisingly successful.

Input	Output
Treadmills attached to *cranes* were used to lift heavy obje	TOOLS/MACHINERY (348)
and for supplying power for *cranes* , hoists , and lifts .SB The	TOOLS/MACHINERY (348)
Above this height , a tower *crane* is often used .SB This compri	TOOLS/MACHINERY (348)
elaborate courtship rituals *cranes* build a nest of vegetation o	ANIMAL,INSECT (414)
are more closely related to *cranes* and rails .SB They range in	ANIMAL,INSECT (414)
low trees .PP At least five *crane* species are in danger of exti	ANIMAL,INSECT (414)

The method rests on the same Bayesian foundations as the bilingually trained material. Each of the 1042 models, $Pr(w|RogetCategory_i)$, $i = 1, \cdots, 1042$, is trained by interpolating between local probabilities and global probabilities, just as before. However, we do not have a corpus tagged with Roget categories, so it may not be obvious how to extract subsections of the corpus meeting local conditions. Consider the Roget Category TOOLS/MACHINERY (348). Ideally, we would extract 100-word contexts in the 10 million

word Grolier Encyclopedia surrounding words in category 348 and use them to compute the local probabilities. Yarowsky suggested extracting contexts around all words in category 348 and weighting appropriately to compensate for the fact that some of these contexts should not have been included in the training set. The following table shows a sample of the 30,924 concordances for the category 348.

OOD CARVING .SB The gutter	*adz* has a concave blade for forming
le equipment such as a hydraulic	*shovel* capable of lifting 26 cubic me
ommon .SB Resembling a power	*shovel* mounted on a floating hull , th
ic equipment , valves for nuclear	*generators* , oil-refinery turbines , and
to 8000 BC , flint-edged wooden	*sickles* were used to gather wild grain
steel-penetrating carbide-tipped	*drills* forced manufacturers to find str
itement heightens the colors .SB	*Drills* live in the forests of equatorial
the traditional ABC method and	*drill* were unchanged , and dissatisfac
s center of rotation .PP A tower	*crane* is an assembly of fabricated ste
n marshy areas .SB The crowned	*crane* , however , occasionally nests i

Note that some of the words in category 348 are polysemous (e.g. *drill* and *crane*), and consequently not all of their contexts should have been included in the training set for category 348. There are at least two factors that ameliorate this problem. First, as noted in the previous section, the Bayesian models can tolerate some noise in their inputs even when there are few categories.\cond, the signal is concentrated in category 348, while noise is distributed through the remaining 1041 categories. In addition, an attempt is made to weight the concordance data to minimize this effect and to make the sample representative of all tools and machinery, not just the more common ones. If a word such as *drill* occurs k times in the corpus, all words in the context of *drill* contribute weight $1/k$ to frequency sums. Yarowsky (1992) reports 93% correct disambiguation, averaged over the following words selected from the word-sense disambiguation literature: *bow, bass, galley, mole, sentence, slug, star, duty, issue, taste, cone, interest.* The results were judged by hand, since the corpus does not provide Roget Categories.

In summary, this method provides a way to train on monolingual materials and a way to reduce the required number of models substantially. It does not provide material for extensive testing.

4. Monolingual testing material

The work with bilingual materials provided us with basic discrimination methods and approximate parameters such as window size. The category based models give remarkable accuracy with far fewer parameters, and can be trained on monolingual material suggesting the possibility of a practical system for sense tagging unrestricted text. This possibility requires answers to several questions beyond the classical discrimination question, however, especially how to reduce the number of parameters required. To address these questions we again ran into the problem of acquiring sufficient testing material to develop methods. We have devised another kind of testing material for these studies, called "pseudo-words." This notion arose as we were working with the bilingual material and observed that the level of sense disambiguation we were aiming for was fairly gross: one might regard it as accidental that the two English senses were not in fact two separate words as they were in French.

4.1 Pseudo-Words

Consider *ability* and *mining*. We construct the "pseudo-word" *ability/mining* by supposing that each use of either word is replaced by the pseudo-word so that we cannot tell the meanings apart by looking at the word. We then have a pseudo-word with an *ability* pseudo-sense and a *mining* pseudo-sense. To construct the pseudo-words, we use (nearly) unambiguous words, since ambiguous words would introduce an uncontrolled number of senses. The words used were chosen as follows. First, we collected a set of 160 words which had but one noun definition in COBUILD learner's dictionary, (Sinclair, 1987). Examples of the words collected include *ability, mining, employer, airport, colonist, and nutrient.* One word (*physical*)

was rejected due to its sense and part of speech ambiguities.

We then constructed two initial sets of 100 pseudo-words, one set having two senses each, and one set having three senses each. These sets were constructed as random selections of pairs or triplets of the chosen unambiguous words subject to a constraint of using each chosen word as nearly equally often as possible. Finally, since ambiguous words have non-overlapping Roget categories for their senses (as defined by the Roget categories), we rejected any pairs or triplets that contained words with overlapping categories. These procedures resulted in 77 pseudo-words with three pseudo-senses, and 92 pseudo-words with two pseudo-senses.

The pseudo-words can be used for both automatic training and automatic testing, and can be produced in large numbers from monolingual material. We have used them primarily for testing extensions of Yarowsky's category based models.

We achieve comparable levels of performance on this set of materials as we do with bilingual materials. For the two 2-sense pseudo-words, models for the particular pair of contexts, with the full context of 100 words, achieve an accuracy of .92. Note that this number is based on context alone, and can be combined with other information (especially a prior) to improve performance. Furthermore, it is based on no changes in the Roget categories associated with a given pseudo-sense, changes that Yarowsky found quite important. If we delete the closest ten words, leaving 90 words of context, the accuracy becomes .89. This removes close collocational evidence and seems to us to be a better estimate of the accuracy that class based models using context alone and no category changes might aspire to. The methods reported in Yarowsky (1992), and not optimized for the various findings below, nor for changes in categories gave a .79 accuracy on the 2-sense pseudo-words.

In the following sections we report on some preliminary studies of problems for making a practical system. The results from the various studies are each compared to a reference model, for which the parameters will be specified as each parameter is introduced. The results should not be taken as indicative of the accuracy that a practical system could achieve, because we know other factors that need to be studied, especially the addition of categories to identify a sense, and bigram evidence from the immediately preceding word.

4.2 An attempt to restrict the number of models

As we have said, for the Bayesian discrimination approach, a major problem for a practical system is to reduce the number of parameters. Our experience with systems that people will actually use suggested a limit of about one megabyte of data. The data required is proportional to the number of models times the factor per model. Our first attempt considered using a smaller number of models.

Roget's Thesaurus provides a five level hierarchy of categories, and Yarowsky had chosen the fourth level, the lowest with names. The reference model uses the fourth level hierarchy. We tried the third level, which would cut the number of models by a factor of about six. This resulted in a substantial decrease in performance. The accuracy for the 2-sense pseudo-words fell from .82 with fourth level categories to .72 for third level categories.

4.3 Restricting words per model

A more fruitful approach to parameter reduction has been to reduce the number of words per model. The score when not all the words are retained is a more general case than that given in section 2. Assuming a Poisson distribution for each word, and n independent words, Mosteller and Wallace (p. 55) give the equation

$$\log score(c) = \sum_{1}^{n} [x_i \log(p_{li}/p_{gi}) - w(p_{li} - p_{gi})] \qquad [3]$$

where c is the context, n is the number of retained words, x_i is the observed frequency in the context of the i^{th} retained word, p_{li} is the probability of the i^{th} word in a desired context, p_{gi} is the global probability of the i^{th} word, and w is the length of the context. Note that when all words are included, the second term vanishes and the equation reduces to Equation [2]. The individual terms of the sum are called the scores for the words.

The importance of using the correct formula is substantial. Inclusion of the constant improved the accuracy of the model for two way ambiguous pseudo-words from .74 to .82.

4.3.1 maximum number of words per model We discard words for which p_l is not significantly different from p_g, which reduces the number of words to consider to a few hundred per model on average. However, some models have less than 100 words to consider and some have almost 1000. The following table shows the increase in accuracy for a set of models which differed only in the maximum number of words allowed per model.

max words	10	20	50	100	150	200
accuracy	.768	.798	.817	.819	.818	.817

The reference model uses a maximum of 200 words per model, which had been the best in previous models. Here, it appears that 100 words per model would do as well or better, and that only very few words per model (10 or 20) substantially decreases performance. We expect to continue using up to 200 words per model until other factors have been explored.

4.3.2 selection of words The previous subsection showed that immense savings in data space could be obtained with little loss in accuracy. Therefore we condsidered just how to select the words to keep. We have always excluded a finite list of about 1000 function words. Letting f denote the frequency of a word in a training document and s its score (as defined above), we had previously found heuristically that sf tended to select interesting words. We thought there might be some merit to $s\sqrt{f}$ or to s, so we tested them also. Later, we found the suggestion of Mosteller and Wallace, (p. 56), interpreted here as $s(f - p_g n)$, where p_g is the global probability of the word, and n is the length of the test document, usually 100 in this case. We also considered s. The reference model uses the Mosteller and Wallace criterion, the best motivated theoretically. The results are shown in the following table.

Effects of Selection Criteria

MW	sf	$s\sqrt{f}$	s
.818	.817	.825	.700

Several reasonable methods give about the same results, although one can find poor criteria (s).

4.4 Weighting for context size

Recall that a class model is built by combining the evidence from all of the contexts of a set of seed lemmas defining the class. As explained above, weighting the evidence from the context of each seed lemma by the total length of the context gives models representative of the most frequent words in the class, and not of the class. Yarowsky used equal weights for the context of each seed lemma to overcome this problem. Following preliminary results, our reference model uses the square root of the length of the class as a weight, instead of either equal weighting, or length weighting. Use of equal weighting reduced the accuracy of the reference model from .82 to .80.

5. Conclusions

We have been studying statistical methods for word sense discrimination. Previous work has given us the hope that a practical system might be constructed. The recent work reported here emphasizes the following points.

Pseudo-words give us a large amount of testing material. While this source of material will only serve to develop methods that make the most gross distinctions, we believe that (a) it is good practice to work the easiest cases first, and (b) this method truly breaks the testing material bottleneck that has stymied progress in word sense disambiguation for so long.

Practical systems will need to be built with only about a million bytes worth of parameters. We have found that restricting the number of class models is a poor way to do this. With many class models having at most a few hundred words each, there are issues of how best to select the words to use, and of how many words per model are useful. We have begun to address these issues.

References

1. Brown, Peter, Jennifer Lai, and Robert Mercer (1991) "Aligning Sentences in Parallel Corpora," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp 169-176.
2. Chapman, Robert (1977). *Roget's International Thesaurus (Fourth Edition)*, Harper and Row, NY.
3. 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.
4. Gale, W., K. Church, and D. Yarowsky, (1992) "A Method for Disambiguating Word Senses in a Large Corpus," *to appear in Computers and the Humanities.*
5. Grolier's Inc. (1991) *New Grolier's Electronic Encyclopedia.*
6. Hirst, G. (1987), *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge.
7. Kelly, Edward, and Phillip Stone (1975), *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.
8. 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.
9. Mosteller, Frederick, and David Wallace (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts.
10. Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley.
11. Sinclair, J., Hanks, P., Fox, G., Moon, R., Stock, P. et al. (eds.) (1987) *Collins Cobuild English Language Dictionary*, Collins, London and Glasgow.
12. van Rijsbergen, C. (1979) *Information Retrieval*, Second Editional, Butterworths, London.
13. Yarowsky, David (1992), "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," *Proceedings COLING-92.*