One Sense Per Discourse

William A. Gale
Kenneth W. Church
David Yarowsky

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974-0636

1. ABSTRACT

It is well-known that there are polysemous words like *sentence* whose "meaning" or "sense" depends on the context of use. We have recently reported on two new word-sense disambiguation systems, one trained on bilingual material (the Canadian Hansards) and the other trained on monolingual material (Roget's Thesaurus and Grolier's Encyclopedia). As this work was nearing completion, we observed a very strong discourse effect. That is, if a polysemous word such as *sentence* appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense. This paper describes an experiment which confirmed this hypothesis and found that the tendency to share sense in the same discourse is extremely strong (98%). This result can be used as an additional source of constraint for improving the performance of the word-sense disambiguation algorithm. In addition, it could also be used to help evaluate disambiguation algorithms that did not make use of the discourse constraint.

2. OUR PREVIOUS WORK ON WORD-SENSE DISAMBIGUATION

2.1. Data Deprivation

Although there has been a long history of work on word-sense disambiguation, much of the work has been stymied by difficulties in acquiring appropriate testing and training materials. AI approaches have tended to focus on "toy" domains because of the difficulty in acquiring large lexicons. So too, statistical approaches, e.g., Kelly and Stone (1975), Black (1988), have tended to focus on a relatively small set of polysemous words because they have depended on extremely scarce hand-tagged materials for use in testing and training.

We have achieved considerable progress recently by using a new source of testing and training materials and the application of Bayesian discrimination methods. Rather than depending on small amounts of hand-tagged 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. 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 use of bilingual materials for discrimination decisions in machine tranlation has been discussed by Brown and others (1991), and by Dagan, Itai, and Schwall (1991). The use of bilingual materials for an essentially monolingual purpose, sense disambiguation, is similar in method, but differs in purpose.

2.2. Bayesian Discrimination

Surprisingly good results can be achieved using Bayesian discrimination methods which have been used very successfully in many other applications, especially author identification (Mosteller and Wallace, 1964) and information retrieval (IR) (Salton, 1989, section 10.3). Our word-sense disambiguation algorithm uses the words in a 100-word context[1] surrounding the polysemous word very much like the other two applications use the words in a test document.

Information Retreival (IR):

$$\prod_{w \text{ in } doc} \frac{Pr(w|rel)}{Pr(w|irrel)}$$

[1] It is common to use very small contexts (e.g., 5-words) based on the observation that people do not need very much context in order to performance the disambiguation task. In contrast, we use much larger contexts (e.g., 100 words). Although people may be able to make do with much less context, we believe the machine needs all the help it can get, and we have found that the larger context makes the task much easier. In fact, we have been able to measure information at extremely large distances (10,000 words away from the polysemous word in question), though obviously most of the useful information appears relatively near the polysemous word (e.g., within the first 100 words or so). Needless to say, our 100-word contexts are considerably larger than the smaller 5-word windows that one normally finds in the literature.

Author Identification:

$$\prod_{w \text{ in } doc} \frac{Pr(w|author_1)}{Pr(w|author_2)}$$

Word-Sense Disambiguation:

$$\prod_{w \text{ in } context} \frac{Pr(w|sense_1)}{Pr(w|sense_2)}$$

This model treats the context as a bag of words and ignores a number of important linguistic factors such as word order and collocations (correlations among words in the context). Nevertheless, even with these oversimplifications, the model still contains an extremely large number of parameters: $2V \approx 200,000$. 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 each data point. Clearly, we need to be fairly careful given that we have so many parameters and so little evidence.

2.3. Parameter Estimation

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 *sentence*, 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, which is obviously unacceptable.

We will estimate $Pr(tok|sense)$ by interpolating between local probabilities, probabilities computed over the 100-word contexts, and global probabilities, probabilities computed over the entire corpus. There is a trade-off between measurement error and bias error. The local probabilities tend to be more prone to measurement error whereas the global probabilities tend to be more prone to bias error. We seek to determine the relevance of the larger corpus to the conditional sample in order to find the optimal trade-off between bias and measurement error.

The interpolation procedure makes use of a prior expectation of how much we expect the local probabilities to differ from the global probabilities. In their author identification work Mosteller and Wallace "expect[ed] both authors to have nearly identical rates for almost any word" (p. 61). In fact, just as they had anticipated, we have found that only 2% of the vocabulary in the Federalist corpus has significantly different probabilities depending on the author. In contrast, we expect fairly large differences in the sense disambiguation application. Approximately 20% of the vocabulary in the Hansards has a local probability that is significantly different from its global probability. Since the prior expectation depends so much on the application, we set the prior for a particular application by estimating the fraction of the vocabulary whose local probabilities differ significantly from the global probabilities.

2.4. A Small Study

We have looked at six polysemous nouns in some detail: *duty, drug, land, language, position* and *sentence*, as shown in Table 1. The final column shows that performance is quite encouraging.

Table 1: Six Polysemous Words

English	French	sense	N	%
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

These nouns were selected because they could be disambiguated by looking at their French translation in the Canadian Hansards (unlike a polysemous word such as *interest* whose French translation *interêt* is just as ambiguous as the English source). In addition, for testing methodology, it is helpful that the corpus contain plenty of instances of each sense. The second condition, for example, would exclude *bank*, perhaps the canonical example of a polysemous noun, since there are very few instances of the "river" sense of *bank* in our corpus of Canadian Hansards. We were somewhat surprised to discover that these two conditions are actually fairly stringent, and that there are a remarkably small number of polysemous words which (1) can be disambiguated by looking at the French translation, and (2) appear 150 or more times in two or more senses.

Table 2: Automatic Sense Tagging Using Roget's Categories

Input		Output
Treadmills attached to	*cranes* were used to lift heavy objects	TOOLS/MACHINERY (348)
and for supplying power for	*cranes*, hoists, and lifts. The centrifug	TOOLS/MACHINERY (348)
Above this height, a tower	*crane* is often used. This comprises	TOOLS/MACHINERY (348)
elaborate courtship rituals	*cranes* build a nest of vegetation on	ANIMALS/INSECTS (414)
are more closely related to	*cranes* and rails. They range in length	ANIMALS/INSECTS (414)
low trees. .PP At least five	*crane* species are in danger of extincti	ANIMALS/INSECTS (414)

2.5. Training on Monolingual 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. Moreover, the assumption that differences in translation correspond to differences in word-sense has always been somewhat suspect.

Recently, Yarowsky (1991) has found a way to train on the Roget's Thesaurus (Chapman, 1977) and Grolier's Encyclopedia (1991) instead of the Hansards, thus circumventing many of the objections to our use of the Hansards. Yarowsky's method inputs a 100-word context surrounding a polysemous word and scores each of the 1042 Roget Categories by:

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

Table 2 shows some results for the polysemous noun *crane*.

Each of the 1042 models, $Pr(w|Roget\ Category_i)$, is trained by interpolating between local probabilities and global probabilities just as before. However, the local probabilities are somewhat more difficult to obtain in this case since we do not have a corpus tagged with Roget Categories, and therefore, it may not be obvious how to extract subsections of the corpus meeting the 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 these to compute the local probabilities. Since we don't have a tagged corpus, Yarowsky suggested extracting contexts around all words in category 348 and weighting appropriately in order to compensate for the fact that some of these contexts should not have been included in the training set. Table 3 below shows a sample of the 30,924 concordances for the words in category 348.

Table 3: Some Concordances for TOOLS/MACHINERY

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

Note that some of the words in category 348 are polysemous (e.g., *drill* and *crane*), and consequently, not all of their contexts should be included in the training set for category 348. In particular, lines 7, 8 and 10 in Table 3 illustrate the problem. If one of these spurious senses was frequent and dominated the set of examples, the situation could be disastrous. 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. Although the training materials still contain a substantial level of noise, we have found that the resulting models work remarkably well, nontheless. Yarowsky (1991) 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.*

3. A HYPOTHESIS: ONE SENSE PER DISCOURSE

As this work was nearing completion, we observed that senses tend to appear in clumps. In particular, it appeared to be extremely unusual to find two or more senses of a polysemous word in the same discourse.[2] A simple (but non-blind) preliminary experiment provided some suggestive evidence confirming the hypothesis. A random sample of 108 nouns was extracted for further

[2]This hypothesis might help to explain some of the long-range effects mentioned in the previous footnote.

study. A panel of three judges (the authors of this paper) were given 100 sets of concordance lines. Each set showed all of the instances of one of the test words in a particular Grolier's article. The judges were asked to indicate if the set of concordance lines used the same sense or not. Only 6 of 300 article-judgements were judged to contain multiple senses of one of the test words. All three judges were convinced after grading 100 articles that there was considerable validity to the hypothesis.

With this promising preliminary verification, the following blind test was devised. Five subjects (the three authors and two of their colleagues) were given a questionnaire starting with a set of definitions selected from OALD (Crowie *et al.*, 1989) and followed by a number of pairs of concordance lines, randomly selected from Grolier's Encyclopedia (1991). The subjects were asked to decide for each pair, whether the two concordance lines corresponded to the same sense or not.

antenna
1. jointed organ found in pairs on the heads of insects and crustaceans, used for feeling, etc. → the illus at insect.
2. radio or TV aerial.

lack eyes, legs, wings,	*antennae* and distinct mouthparts
The Brachycera have short	*antennae* and include the more evol

The questionnaire contained a total of 82 pairs of concordance lines for 9 polysemous words: *antenna, campaign, deposit, drum, hull, interior, knife, landscape,* and *marine.* 54 of the 82 pairs were selected from the same discourse. The remaining 28 pairs were introduced as a control to force the judges to say that some pairs were different; they were selected from different discourses, and were checked by hand as an attempt to assure that they did not happen to use the same sense. The judges found it quite easy to decide whether the pair used the same sense or not. Table 4 shows that there was very high agreement among the judges. With the exception of judge 2, all of the judges agreed with the majority opinion in all but one or two of the 82 cases. The agreement rate was 96.8%, averaged over all judges, or 99.1%, averaged over the four best judges.

Table 4: Strong Agreement

Judge	n	%
1	82	100.0%
2	72	87.8%
3	81	98.7%
4	82	100.0%
5	80	97.6%
Average		96.8%
Average (without Judge 2)		99.1%

As we had hoped, the experiment did, in fact, confirm the one-sense-per-discourse hypothesis. Of 54 pairs selected from the same article, the majority opinion found that 51 shared the same sense, and 3 did not.[3]

We conclude that with probability about 94% (51/54), two polysemous nouns drawn from the same article will have the same sense. In fact, the experiment tested a particularly difficult case, since it did not include any unambiguous words. If we assume a mixture of 60% unambiguous words and 40% polysemous words, then the probability moves from 94% to $100\% \times .60 + 94\% \times .40 \approx 98\%$. In other words, there is a very strong tendency (98%) for multiple uses of a word to share the same sense in well-written coherent discourse.

One might ask if this result is specific to Grolier's or to good writing or some other factor. The first author looked at the usage of these same nine words in the Brown Corpus, which is believed to be a more balanced sample of general language and which is also more widely available than Grolier's and is therefore more amenable to replication. The Brown Corpus consists of 500 discourse fragments of 2000 words, each. We were able to find 259 concordance lines like the ones above, showing two instances of one of the nine test words selected from the same discourse fragment. However, four of the nine test words are not very interesting in the Brown Corpus *antenna, drum, hull,* and *knife,* since only one sense is observed. There were 106 pairs for the remaining five words: *campaign, deposit, interior, landscape,* and *marine.* The first author found that 102 of the 106 pairs were used in the same sense. Thus, it appears that one-sense-per-discourse tendency is also fairly strong in the Brown Corpus (102/106 ≈ 96%), as well as in the Grolier's Encyclopedia.

4. IMPLICATIONS

There seem to be two applications for the one-sense-per-discourse observation: first it can be used as an additional source of constraint for improving the performance of the word-sense disambiguation algorithm, and

[3] In contrast, of the 28 control pairs, the majority opinion found that only 1 share the same sense, and 27 did not.

secondly, it could be used to help evaluate disambiguation algorithms that did not make use of the discourse constraint. Thus far, we have been more interested in the second use: establishing a group of examples for which we had an approximate ground truth. Rather than tagging each instance of a polysemous word one-by-one, we can select discourses with large numbers of the polysemous word of interest and tag all of the instances in one fell swoop. Admittedly, this procedure will introduce a small error rate since the one-sense-per-discourse tendency is not quite perfect, but thus far, the error rate has not been much of a problem. This procedure has enabled us to tag a much larger test set than we would have been able to do otherwise.

Having tagged as many words as we have (all instances of 97 words in the Grolier's Encyclopedia), we are now in a position to question some widely held assumptions about the distribution of polysemy. In particular, it is commonly believed that most words are highly polysemous, but in fact, most words (both by token and by type) have only one sense, as indicated in Table 5 below. Even for those words that do have more than one possible sense, it rarely takes anywhere near $log_2 senses$ bits to select the appropriate sense since the distribution of senses is generally quite skewed. Perhaps the word-sense disambiguation problem is not as difficult as we might have thought.

Table 5: Distribution of Polysemy

senses	types	tokens	avg. entropy
1	67	7569	0
2	16	2552	0.58
3	7	1313	0.56
4	5	1252	1.2
5	1	1014	0.43
6	1	594	1.3

5. CONCLUSION

In conclusion, it appears that our hypothesis is correct; well-written discourses tend to avoid multiple senses of a polysemous word. This result can be used in two basic ways: (1) as an additional source of constraint for improving the performance of a word-sense disambiguation algorithm, and (2) as an aide in collecting annotated test materials for evaluating disambiguation algorithms.

6. REFERENCES

1. Black, Ezra (1988), "An Experiment in Computational Discrimination of English Word Senses," *IBM Journal of Research and Development*, v 32, pp 185-194.
2. Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer (1991), "Word Sense Disambiguation using Statistical Methods," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp 264-270.
3. Chapman, Robert (1977). *Roget's International Thesaurus (Fourth Edition)*, Harper and Row, New York.
4. 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.
5. Gale, Church, and Yarowsky, 1992, "Discrimination Decisions for 100,000-Dimensional Spaces" AT&T Statistical Research Report No. 103.
6. Grolier's Inc. (1991) *New Grolier's Electronic Encyclopedia.*
7. Hirst, G. (1987), *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge.
8. Kelly, Edward, and Phillip Stone (1975), *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.
9. Mosteller, Fredrick, and David Wallace (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts.
10. Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley Publishing Co.
11. Yarowsky, David (1991) "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", submitted to COLING-92.