Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs

William Gale
Kenneth Ward Church
David Yarowsky

AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974
kwc@research.att.com

Abstract

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). After using both the monolingual and bilingual classifiers for a few months, we have convinced ourselves that the performance is remarkably good. Nevertheless, we would really like to be able to make a stronger statement, and therefore, we decided to try to develop some more objective evaluation measures. Although there has been a fair amount of literature on sense-disambiguation, the literature does not offer much guidance in how we might establish the success or failure of a proposed solution such as the two systems mentioned in the previous paragraph. Many papers avoid quantitative evaluations altogether, because it is so difficult to come up with credible estimates of performance.

This paper will attempt to establish upper and lower bounds on the level of performance that can be expected in an evaluation. An estimate of the lower bound of 75% (averaged over ambiguous types) is obtained by measuring the performance produced by a baseline system that ignores context and simply assigns the most likely sense in all cases. An estimate of the upper bound is obtained by assuming that our ability to measure performance is largely limited by our ability obtain reliable judgments from human informants. Not surprisingly, the upper bound is very dependent on the instructions given to the judges. Jorgensen, for example, suspected that lexicographers tend to depend too much on judgments by a single informant and found considerable variation over judgments (only 68% agreement), as she had suspected. In our own experiments, we have set out to find word-sense disambiguation tasks where the judges can agree often enough so that we could show that they were outperforming the baseline system. Under quite different conditions, we have found 96.8% agreement over judges.

1. Introduction: Using Massive Lexicographic Resources

Word-sense disambiguation is a long-standing problem in computational linguistics (e.g., Kaplan (1950), Yngve (1955), Bar-Hillel (1960), Masterson (1967)), with important implications for a number of practical applications including text-to-speech (TTS), machine translation (MT), information retrieval (IR), and many others. The recent interest in computational lexicography has fueled a large body of recent work on this 40-year-old problem, e.g., Black (1988), Brown *et al.* (1991), Choueka and Lusignan (1985), Clear (1989), Dagan *et al.* (1991), Gale *et al.* (to appear), Hearst (1991), Lesk (1986), Smadja and McKeown (1990), Walker (1987), Veronis and Ide (1990), Yarowsky (1992), Zernik (1990, 1991). Much of this work offers the prospect that a disambiguation system might be able to input unrestricted text and tag each word with the most likely sense with fairly reasonable accuracy and efficiency, just as part of speech taggers (e.g., Church (1988)) can now input unrestricted text and assign each word with the most likely part of speech with fairly reasonable accuracy and efficiency.

The availability of massive lexicographic databases offers a promising route to overcoming the knowledge acquisition bottleneck. More than thirty years ago, Bar-Hillel (1960) predicted that it would be ''futile'' to write expert-system-like rules by-hand (as they had been doing at Georgetown at the time) because there would be no way to scale up such rules to cope with unrestricted input. Indeed, it is now well-known that expert-system-like rules can be notoriously difficult to scale up, as Small and Reiger (1982) and many others have observed:

> ''The expert for THROW is currently six pages long... but it should be 10 times that size.''

Bar-Hillel was very early in realizing the scope of the problem; he observed that people have a large set of facts at their disposal, and it is not obvious how a computer could ever hope to gain access to this wealth of knowledge.

> '' 'But why not envisage a system which will put this knowledge at the disposal of the translation machine?' Understandable as this reaction is, it is very easy to show its futility. What such a suggestion amounts to, if taken seriously, is the requirement that a translation machine should not only be supplied with a dictionary but also with a universal encyclopedia. This is surely utterly chimerical and hardly deserves any further discussion. Since, however, the idea of a machine with encyclopedic knowledge has popped up also on other occasions, let me add a few words on this topic. The number of facts we human beings know is, in a certain very pregnant sense, infinite.'' (Bar-Hillel, 1960)

Ironically, much of the research cited above is taking exactly the approach that Bar-Hillel ridiculed as utterly chimerical and hardly deserving of any further discussion. Back in 1960, it may have been hard to imagine how it would be possible to supply a machine with both a dictionary and an encyclopedia. But much of the recent work cited above goes much further; not only does it supply a machine with a dictionary and an encyclopedia, but many other extensive references works as well, including Roget's Thesaurus and numerous large corpora. Of course, we are using these reference works in a very superficial way; we are certainly not suggesting that the machine should attempt to solve the ''AI Complete'' problem of ''understanding'' these reference works.

2. A Brief Summary of Our Previous Work

Our own work has made use of many of these lexical resources. In particular, (Gale *et al.*, to appear) achieved considerable progress by using well-understood statistical methods and very large datasets of tens of millions of words of parallel English and French text (e.g., the Canadian Hansards). By aligning the text as we have, we were able to collect a large set of examples of polysemous words (e.g., *sentence*) in each sense (e.g., *judicial sentence* vs. *syntactic sentence*), by extracting instances from the corpus that were translated one way or the other (e.g, *peine* or *phrase*). These data sets were then analyzed using well-understood Bayesian discrimination methods, which have been used very successfully in many other applications, especially author identification (Mosteller and Wallace, 1964, section 3.1) and information retrieval (IR) (van Rijsbergen, 1979, chapter 6; Salton, 1989, section 10.3), though their application to word-sense disambiguation is novel.

In author identification and information retrieval, it is customary to split the discrimination process up into a testing phase and a training phase. During the training phase, we are given two (or more) sets of documents and are asked to construct a discriminator which can distinguish between the two (or more) classes of documents. These discriminators are then applied to new documents during the testing phase. In the author identification task, for example, the training set consists of several documents written by each of the two (or more) authors. The resulting discriminator is then tested on documents whose authorship is disputed. In the information retrieval application, the training set consists of a set of one or more relevant documents and a set of zero or more irrelevant documents. The resulting discriminator is then applied to all documents in the library in order to separate the more relevant ones from the less relevant ones.

There is an embarrassing wealth of information in the collection of documents that could be used as the basis for discrimination. It is common practice to treat documents as ''merely'' a bag of words, and to ignore much of the linguistic structure, especially dependencies on word order and correlations between pairs of words. In other words, one assumes that there are two (or more) sources of word probabilities, *rel* and *irrel*, in the IR application, and $author_1$ and $author_2$ in the author identification application. During the training phase, we attempt to estimate $Pr(w|source)$ for all words w in the vocabulary and all sources. Then during the testing phase, we score all documents as follows and select high scoring documents as being relatively likely to have been generated by the source of interest.

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

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

In the sense disambiguation application, the 100-word context surrounding instances of a polysemous word (e.g., *sentence*) are treated very much like a document.[1]

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

That is, during the testing phase, we are given a new instance of a polysemous word, e.g., *sentence*, and asked to assign it to one or more senses. We score the words in the 100-word context using the formula given above, and assign the instance to $sense_1$ if the score is large.

1. It is common to use very small contexts (e.g., 5-words) based on the observation that people seem to be able to disambiguate word-senses based on very little context. We have taken a different approach. Since we have been able to find useful information out to 100 words (and measurable information out to 10,000 words), we feel we might as well make use of the the larger contexts. This task is very difficult for the machine; it needs all the help it can get.

The conditional probabilities, $Pr(w|sense)$, are determined during the training phase by counting the number of times that each word in the vocabulary was found near each sense of the polysemous word (and then smoothing these estimates in order to deal with the sparse-data problems). See Gale *et al.* (to appear) for further details.

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 (1992) has found a way to extend our use of the Bayesian techniques by training on the Roget's Thesaurus (Chapman, 1977)[2] and Grolier's Encyclopedia (1991) instead of the Canadian Hansards, thus circumventing many of the objections to our use of the Hansards. Yarowsky (1992) inputs a 100-word context surrounding a polysemous word and scores each of the 1042 Roget Categories by:

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

The program can also be run in a mode where it takes unrestricted text as input and tags each word with its most likely Roget Category. Some results for the word *crane* are presented below, showing that the program can be used to sort a concordance by sense.

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

After using both the monolingual and bilingual classifiers for a few months, we have convinced ourselves that the performance is remarkably good. Nevertheless, we would really like to be able to make a stronger statement, and therefore, we decided to try to develop some more objective evaluation measures.

2. Note that this edition of the Roget's Thesaurus is much more extensive than the 1911 version, though somewhat more difficult to obtain in electronic form.

3. The Literature on Evaluation

Although there has been a fair amount of literature on sense-disambiguation, the literature does not offer much guidance in how we might establish the success or failure of a proposed solution such as the two described above. Most papers tend to avoid quantitative evaluations. Lesk (1986), an extremely innovative and commonly cited reference on the subject, provides a short discussion of evaluation, but fails to offer any very satisfying solutions that we might adopt to quantify the performance of our two disambiguation algorithms.[3]

Perhaps the most common evaluation technique is to select a small sample of words and compare the results of the machine with those of a human judge. This method has been used very effectively by Kelly and Stone (1975), Black (1988), Hearst (1991), and many others. Nevertheless, this technique is not without its problems, perhaps the worst of which is that the sample may not be very representative of the general vocabulary. Zernik (1990, p. 27), for example, reports 70% performance for the word *interest*, and then acknowledges that this level of performance may not generalize very well to other words.[4]

Although we agree with Zernik's prediction that *interest* is not very representative of other words, we suspect that *interest* is actually more difficult than most other words, not less difficult. Table 1 shows the performance of Yarowsky (1992) on twelve words which have been previously discussed in the literature. Note that *interest* is at the bottom of the list.

The reader should exercise some caution in interpreting the numbers in Table 1. It is natural to try to use these numbers to predict performance on new words, but the study was not designed for that purpose. The test words were selected from the literature in order to make comparisons over systems. If the study had been intended to support predictions on new words, then the study should have used a random sample of such words, rather than a sample of words from the literature.

3. ''What is the current performance of this program? Some very brief experimentation with my program has yielded accuracies of 50-70% on short samples of *Pride and Prejudice* and an Associated Press news story. Considerably more work is needed both to improve the program and to do more thorough evaluation... There is too much subjectivity in these measurements.'' (Lesk, 1986, p. 6)

4. ''For all 4 senses of INTEREST, both recall and precision are over 70%... However, not for all words are the obtained results that positive... The fact is that almost any English word possesses multiple senses. (Zernik, 1990, p. 27)

Table 1: Comparison over Systems		
Word	Yarowsky (1992)	Previous Systems
bow	91%	< 67% (Clear, 1989)
bass	99%	100% (Hearst, 1991)
galley	99%	50-70% (Lesk, 1986)
mole	99%	N/A (Hirst, 1987)
sentence	98%	90% (Gale *et al.*)
slug	97%	N/A (Hirst, 1987)
star	96%	N/A (Hirst, 1987)
duty	96%	96% (Gale *et al.*)
issue	94%	< 70% (Zernik, 1990)
taste	93%	< 65% (Clear, 1989)
cone	77%	50-70% (Lesk, 1986)
interest	72%	72% (Black, 1988); 70% (Zernik, 1990)
AVERAGE	92%	N/A

In addition to the sampling questions, one feels uncomfortable about comparing results across experiments, since there are many potentially important differences including different corpora, different words, different judges, differences in treatment of precision and recall, and differences in the use of tools such as parsers and part of speech taggers, etc. In short, there seem to be a number of serious questions regarding the commonly used technique of reporting percent correct on a few words chosen by hand. Apparently, the literature on evaluation of word-sense disambiguation algorithms fails to offer a clear role model that we might follow in order to quantify the performance of our disambiguation algorithms.

4. What is the State-of-the-Art, and How Good Does It Need To Be?

Moreover, there doesn't seem to be a very clear sense of what is possible. Is *interest* a relatively easy word or is it a relatively hard word? Zernik says it is relatively easy; we say it is relatively hard.[5] Should we expect the next word to be easier than *interest* or harder than *interest*?

One might ask if 70% is good or bad. In fact, both Black (1988) and Yarowsky (1992) report 72% performance on this very same word. Although it is dangerous to compare such results since there are many potentially important differences (e.g., corpora, judges, etc.), it appears that Zernik's 70% figure is fairly representative of the state of the art.[6]

Should we be happy with 70% performance? In fact, 70% really isn't very good. Recall that Bar-Hillel (1960, p. 159) abandoned the machine translation field when he couldn't see how a machine could possibly do a decent job in translating text if it couldn't do better than this in disambiguating word senses. Bar-Hillel's real objection was an empirical one. Using his numbers,[7] it appears that programs, at the time, could disambiguate only about 75% of the words in a sentence (e.g., 15 out of 20). If *interest* is a relatively easy word, as Zernik (1990) suggests, then it would seem that Bar-Hillel's argument remains as true today as it was in 1960, and we ought to follow his lead and find something more productive to do with our time. On the other hand, if we are correct and *interest* is a relatively difficult word, then it is possible that we have made some progress over the past thirty years...

5. Upper and Lower Bounds

5.1 Lower Bounds

We could be in a better position to address the question of the relative difficulty of *interest* if we could establish a rough estimate of the upper and lower bounds on the level of performance that can be expected. We will estimate the lower bound by evaluating the performance of a straw man system, which ignores context and simply assigns the most likely sense in all cases. One might hope that reasonable systems should generally

5. As evidence that *interest* is relatively difficult, we note that both the Oxford Advanced Learner's Dictionary (OALD) (Crowie *et al.*, 1989, p. 654) and COBUILD (Sinclair *et al.*, 1987), for example, devote more than a full column to this word, indicating that it is an extremely complex word, at least by their standards.

6. In fact, Zernik's 70% figure is probably significantly inferior to the 72% reported by Black and Yarowsky, because Zernik reports precision and recall separately, whereas the others report a single figure of merit which combines both Type I (false rejection) and Type II (false acceptance) errors by reporting precision at 100% recall. Gale *et al.* show that error rates for 70% recall were half of those for 100% recall, on their test sample.

7. ''Let me state rather dogmatically that there exists at this moment no method of reducing the polysemy of the, say, twenty words of an average Russian sentence in a scientific article below a remainder of, I would estimate, at least five or six words with multiple English renderings, which would not seriously endanger the quality of the machine output. Many tend to believe that by reducing the number of initially possible renderings of a twenty word Russian sentence from a few tens of thousands (which is the approximate number resulting from the assumption that each of the twenty Russian words has two renderings on the average, while seven or eight of them have only one rendering) to some eighty (which would be the number of renderings on the assumption that sixteen words are uniquely rendered and four have three renderings apiece, forgetting now about all the other aspects such as change of word order, etc.) the main bulk of this kind of work has been achieved, the remainder requiring only some slight additional effort.'' (Bar-Hillel, 1960, p. 163)

outperform this baseline system, though not all such systems actually do. In fact, Yarowsky (1992) falls below the baseline for one of the twelve words (*issue*), although perhaps, we needn't be too concerned about this one deviation.[8]

There are, of course, a number of problems with this estimate of the baseline. First, the baseline system is not operational, at least as we have defined it. Ideally, the baseline system ought to try to estimate the most likely sense for each word in the vocabulary and then assign that sense to each instance of the word in the test set. Unfortunately, since it isn't clear just how this estimation should be accomplished, we decided to ''cheat'' and let the baseline system peek at the test set and ''estimate'' the most likely sense for each word as the more frequent sense in the test set. Consequently, the performance of the baseline cannot fall below chance ($100/k$% for a particular word with k senses).[9]

In addition, the baseline system assumes that Type I (false rejection) errors are just as bad as Type II (false acceptance) errors. If one desires extremely high recall and is willing to sacrifice precision in order to obtain this level of recall, then it might be sensible to tune a system to produce behavior which might appear to fall below the baseline. We have run into such situations when we have attempted to help lexicographers find extremely unusual events. In such a case, a lexicographer might be quite happy receiving a long list of potential candidates, only a small fraction of which are actually the case of interest. One can come up with quite a number of other scenarios where the baseline performance could be somewhat misleading, especially when there is an unusual trade-off between the cost of a Type I error and the cost of a Type II error.

Nevertheless, the proposed baseline does seem to provide a usable rough estimate of the lower bound on performance. Table 2 shows the baseline performance for each of the twelve words in Table 1. Note that performance is generally above the baseline as we would hope.

Table 2: The Baseline

Word	Baseline	Yarowsky (1992)
issue	96%	94%
duty	87%	96%
galley	83%	99%
star	83%	96%
taste	74%	93%
bass	70%	99%
slug	62%	97%
sentence	62%	98%
interest	60%	72%
mole	59%	99%
cone	51%	77%
bow	48%	91%
AVERAGE	70%	92%

As mentioned previously, the test words in Tables 1 and 2 were selected from the literature on polysemy, and therefore, tend to focus on the more difficult cases. In another experiment, we selected a random sample of 97 words; 67 of them were unambiguous and therefore had a baseline performance of 100%.[10] The remaining thirty words are listed along with the number of senses and baseline performance: *virus* (2, 98%), *device* (3, 97%), *direction* (2, 96%), *reader* (2, 96%), *core* (3, 94%), *hull* (2, 94%), *right* (5, 94%), *proposition* (2, 89%), *deposit* (2, 88%), *hour* (4, 87%), *path* (2, 86%), *view* (3, 86%), *pyramid* (3, 82%), *antenna* (2, 81%), *trough* (3, 77%), *tyranny* (2, 75%), *figure* (6, 73%), *institution* (4, 71%), *crown* (4, 64%), *drum* (2, 63%), *pipe* (4, 60%), *processing* (2, 59%), *coverage* (2, 58%), *execution* (2, 57%), *min* (2, 57%), *interior* (4, 56%), *campaign* (2, 51%), *output* (2, 51%), *gin* (3, 50%), *drive* (3, 49%). In studying these 97 words, we found that the average baseline performance is much higher than we might have guessed (93% averaged over tokens, 92% averaged over types). In particular, note that this baseline is well above the 75% figure that we associated with Bar-Hillel above. Of course, the large number of unambiguous words contributes greatly to the baseline. If we exclude the unambiguous words, then the average baseline

8. Many of the systems mentioned in Table 2 including Yarowsky (1992) do not currently take advantage of the prior probabilities of the senses, so they would be at a disadvantage relative to the baseline if one of the senses had a very high prior, as is the case for the test word *issue*.

9. In addition, the baseline doesn't deal as well as it could with skewed distributions. One could almost certainly improve the model of the baseline by making use of a notion like entropy that could deal more effectively with skewed distributions. Nevertheless, we will stick with our simpler notion of the baseline for expository convenience.

10. The 67 unambiguous words were: *acid, annexation, benzene, berry, capacity, cereal, clock, coke, colon, commander, consort, contract, cruise, cultivation, delegate, designation, dialogue, disaster, equation, esophagus, fact, fear;, fertility, flesh, fox, gold, interface, interruption, intrigue, journey, knife, label, landscape, laurel, lb, liberty, lily, locomotion, lynx, marine, memorial, menstruation, miracle, monasticism, mountain, nitrate, orthodoxy, pest, planning, possibility, pottery, projector, regiment, relaxation, reunification, shore, sodium, specialty, stretch, summer, testing, tungsten, universe, variant, vigor, wire, worship*.

performance falls to 81% averaged over tokens and 75% averaged over types.

5.2 Upper Bounds

We will attempt to estimate an upper bound on performance by estimating the ability for human judges to agree with one another (or themselves). We will find, not surprisingly, that the estimate varies widely depending on a number of factors, especially the definition of the task. Jorgensen (1990) has collected some interesting data that may be relevant for estimating the agreement among judges. As part of her dissertation under George Miller at Princeton, she was interested in assessing ''the extent of psychologically real polysemy in the mental lexicon for nouns.'' Her experiment was designed to study one of the more commonly employed methods in lexicography for writing dictionary definitions, namely the use of citation indexes. She was concerned that lexicographers and computational linguists have tended to depend too much on the intuitions of a single informant. Not surprisingly, she found considerable variation across judgements, just as she had suspected. This finding could have serious implications for evaluation. How do we measure performance if we can't depend on the judges?

Jorgensen selected twelve high frequency nouns at random from the Brown Corpus, six were highly polysemous (*head, life, world, way, side, hand*) and six were less so (*fact, group, night, development, something, war*). Sentences containing each of these words were drawn from the Brown Corpus and typed on filing cards. Nine subjects where then asked to cluster a packet of these filing cards by sense. A week or two later, the same nine subjects were asked to repeat the experiment, but this time they were given access to the dictionary definitions.

Jorgensen reported performance in terms of the ''Agreement-Disagreement'' (A-D) ratio (Shipstone, 1960) for each subject and each of the twelve test words. We have found it convenient to transform the A-D ratio into a quantity which we call the percent agreement, the number of observed agreements over the total number of possible agreements. The grand mean percent agreement over all subjects and words is only 68%. In other words, at least under these conditions, there is considerable variation across judgements, perhaps so much so that it would be hard to show that a proposed system was outperforming the baseline system (75%, averaged over ambiguous types). Moreover, if we accept Bar-Hillel's argument that 75% is not-good-enough, then it would be hard to show that a system was doing well-enough.

6. A Discrimination Experiment

For evaluation purposes, it is important to find a task that is somewhat easier for the judges. If the task is too hard (as Jorgensen's classification task may be), then there will be almost no room between the limits of the measurement and the baseline. In other words, there won't be enough dynamic range to measure differences between better systems and worse systems. In contrast, if we focus on easier tasks, then we might have enough dynamic range to show some interesting differences. Therefore, unlike Jorgensen who was interested in highlighting differences among judgments, we are much more interested in highlighting agreements. Fortunately, we have found in (Gale *et al.*, 1992) that the agreement rate can be very high (96.8%), which is well above the baseline, under very different experimental conditions.

Of course, it is a fairly major step to redefine the problem from a classification task to a discrimination one, as we are proposing. One might have preferred not to do so, but we simply don't know how one could establish enough dynamic range in that case to show any interesting differences. It has been our experience that it is very hard to design an experiment of any kind which will produce the desired agreement among judges. We are very happy with the 96.8% agreement that we were able to show, even if it is limited to a much easier task than the one that Jorgensen was interested in.

We originally designed the experiment in Gale *et al.* (1992) to test the hypothesis that multiple uses of a polysemous word tend to have the same sense within a common discourse. A simple (but non-blind) pilot experiment provided some suggestive evidence confirming the hypothesis. A random sample of 108 nouns (which included the 97 words previously mentioned) was extracted for further study. A panel of three judges (the three authors of this paper) were given 100 sets of concordance lines containing one of the test words selected from a single article in Grolier's. 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 and
The Brachycera have short *antennae* and include the more evolved

silk moths passes over the *antennae* .SB Only males that detect
relatively simple form of *antenna* is the dipole , or doublet

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*. The results of the experiment are shown below in Table 3. 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. In either case, the agreement rate is well above the previously described ceiling.

Table 3

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%

Incidentally, the experiment did, in fact, confirm the hypothesis that multiple uses of a polysemous word will generally take on the same sense within a discourse. Of the 82 judgments, 54 were selected from the same discourse and were judged to have the same sense by the majority in 96.9% of the cases. (The remaining 28 of the 82 judgments were used as a control to force the judges to say that some pairs were different.)

Note that the tendency for multiple uses of a polysemous word to have the same sense is extremely strong; 96.9% is much greater than the baseline, and indeed, it is considerably above the level of performance that might be expected from state-of-the-art word-sense disambiguation systems. Since it is so reliable and so easy to compute, it might be used as a quick-and-dirty measure for testing such systems. Unfortunately, we also need a complementary measure that would penalize a system like the baseline system that simply assigned all instances of a polysemous word to the same sense. At present, we have yet to identify a quick-and-dirty measure that accomplishes this control, and consequently, we are forced to continue to depend on the relatively expensive panel of judges. But, at least, we have been able to establish that it is possible to design a discrimination experiment such that the panel of judges can agree with themselves often enough to be useful. In addition, we have established that the discourse constraint on polysemy is extremely strong, much stronger than our ability to tag word-senses automatically. Consequently, it ought to be possible to use this constraint in our next word-sense tagging algorithm to produce even better performance.

7. Conclusions

We began this discussion with a review of our recent work on word-sense disambiguation, which extends the approach of using massive lexicographic resources (e.g., parallel corpora, dictionaries, thesauruses and encyclopedia) in order to attack the knowledge-acquisition bottleneck that Bar-Hillel identified over thirty years ago. After using both the monolingual and bilingual classifiers for a few months, we have convinced ourselves that the performance is remarkably good. Nevertheless, we would really like to be able to make a stronger statement, and therefore, we decided to try to develop some more objective evaluation measures. A survey of the literature on evaluation failed to identify an attractive role model. In addition, we found it particularly difficult to obtain a clear estimate of the state-of-the-art.

In order to address this state of affairs, we decided to try to establish upper and lower bounds on the level of performance that we could expect to obtain. We estimated the lower bound by positing a simple baseline system which ignored context and simply assigned the most likely sense in all cases. Hopefully, most reasonable systems would outperform this system. The upper bound was approximated by trying to estimate the limit of our ability to measure performance. We assumed that this limit was largely dominated by the ability for the human judges to agree with one another. The estimate depends very much, not surprisingly, on the particular experimental design. Jorgensen, who was interested in highlighting differences among informants, found a very low estimate (68%), well below the baseline (75%), and also well below the level that Bar-Hillel asserted as not-good-enough. In our own work, we have attempted to highlight agreements, so that there would more dynamic range between the baseline and the limit of our ability to measure performance. In so doing, we were able to obtain a much more usable estimate of (96.8%) by redefining the task from a classification task

to a discrimination task. In addition, we also made use of the constraint that multiple instances of a polysemous word in the same discourse have a very strong tendency to take on the same sense. This constraint will probably prove useful for improving the performance of future word-sense disambiguation algorithms.

Similar attempts to establish upper and lower bounds on performance have been made in other areas of computational linguistics, specifically part of speech tagging. For that application, it is generally accepted that the baseline part-of-speech tagging performance is about 90% (as estimated by a similar baseline system that ignores context and simply assigns the most likely part of speech to all instances of a word) and that the upper bound (imposed by the limit for judges to agree with one another) is about 95%. Incidentally, most part of speech algorithms are currently performing at or near the limit of our ability to measure performance, indicating that there may be room for refining the experimental conditions along similar lines to what we have done here, in order to improve the dynamic range of the evaluation.

References

Bar-Hillel (1960), ''Automatic Translation of Languages,'' in *Advances in Computers*, Donald Booth and R. E. Meagher, eds., Academic, NY.

Black, Ezra (1988), ''An Experiment in Computational Discrimination of English Word Senses,'' *IBM Journal of Research and Development*, v 32, pp 185-194.

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer (1991), ''Word Sense Disambiguation using Statistical Methods,'' *ACL*, pp. 264-270.

Chapman, Robert (1977). *Roget's International Thesaurus (Fourth Edition)*, Harper and Row, NY.

Choueka, Yaacov, and Serge Lusignan (1985), ''Disambiguation by Short Contexts,'' *Computers and the Humanities*, v 19. pp. 147-158.

Church, Kenneth (1988), ''A Stochastic Parts Program an Noun Phrase Parser for Unrestricted Text,'' *Applied ACL Conference*, Austin, Texas.

Clear, Jeremy (1989). ''An Experiment in Automatic Word Sense Identification,'' Internal Document, Oxford University Press, Oxford.

Crowie, Anthony *et al.* (eds.) (1989), ''Oxford Advanced Learner's Dictionary,'' Fourth Edition, Oxford University Press.

Dagan, Ido, Alon Itai, and Ulrike Schwall (1991), ''Two Languages are more Informative than One,'' *ACL*, pp. 130-137.

Gale, William, Kenneth Church, and David Yarowsky (to appear) ''A Method for Disambiguating Word Senses in a Large Corpus,'' *Computers and Humanities*.

Gale, William, Kenneth Church, and David Yarowsky (1992) ''One Sense Per Discourse,'' Darpa Speech and Natural Language Workshop.

Gove, Philip *et al.* (eds.) (1975) ''Webster's Seventh New Collegiate Dictionary,'' G. & C. Merriam Company, Springfield, MA.

Grolier's Inc. (1991) *New Grolier's Electronic Encyclopedia*.

Hanks, Patrick (ed.) (1979), *Collins English Dictionary*, Collins, London and Glasgow.

Hearst, Marti (1991), ''Noun Homograph Disambiguation Using Local Context in Large Text Corpora,'' *Using Corpora*, University of Waterloo, Waterloo, Ontario.

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

Jorgensen, Julia (1990) ''The Psychological Reality of Word Senses,'' *Journal of Psycholinguistic Research*, v. 19, pp 167-190.

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

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

Lesk, Michael (1986), ''Automatic Sense Disambiguation: How to tell a Pine Cone from an Ice Cream Cone,'' *Proceeding of the 1986 SIGDOC Conference*, ACM, NY.

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

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

Procter, P., R. Ilson, J. Ayto, et al. (1978), *Longman Dictionary of Contemporary English*, Longman, Harlow and London.

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

Shipstone, E. (1960) ''Some Variables Affecting Pattern Conception,'' *Psychological Monographs, General and Applied*, v. 74, pp. 1-41.

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

Smadja, F. and K. McKeown (1990), ''Automatically Extracting and Representing Collocations for Language Generation,'' *ACL*, pp. 252-259.

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.

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

Veronis, Jean and Nancy Ide (1990), ''Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries,'' in *Proceedings COLING-90*, pp 389-394.

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.

Weiss, Stephen (1973), ''Learning to Disambiguate,'' *Information Storage and Retrieval*, v. 9, pp 33-41.

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

Yngve, Victor (1955), ''Syntax and the Problem of Multiple Meaning,'' in *Machine Translation of Languages*, William Locke and Donald Booth, eds., Wiley, NY.

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 Retrieval*, P.S. Jacobs, ed., GE Research & Development Center, Schenectady, NY.

Zernik, Uri (1991) ''Train1 vs. Train2: Tagging Word Senses in Corpus,'' in Zernik (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum, Hillsdale, NJ.

Footnote 5 had to be deleted...

There are many other questions in addition to the ones mentioned thus far. For example, the numbers in Table 1 are all based on the judgments of a single informant. Moverover, the evaluation methodology did not employ a double-blind procedure, but rather used a correcting procedure which may have permitted several errors to pass undetected. Although there is considerable precedent in the literature for taking such expedients, it is possible that expedients such as these might produce misleading results.

This paper will discuss the first question, agreement among judges, at some length, and will find that it is possible to achieve relatively high agreement rates (96.8%) among multiple judges if one is careful, and therefore it may not be necessary to employ a large panel of judges. This paper will have less to say on the second question, though we have experimented with Strayhorn's method (1990) of adjusting for the bias by introducing a small number of errors as a control. Initial results indicate that the judge caught over 95% of these control errors, indicating that the average performance, reported as 92% in Table 1, should be adjusted slightly, from 92.45% to 92.1%.

Strayhorn, Joseph (1990) ''Estimating the Errors Remaining in a Data Set: Techniques for Quality Control,'' *The American Statistician*, 44:1, pp. 14-18.

Footnote 12 also had to be deleted...

This set of numbers demonstrate that it would be extremely desirable if someone could build an operational algorithm for identifying unambiguous words.