USING A SEMANTIC CONCORDANCE FOR SENSE IDENTIFICATION

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas*

Cognitive Science Laboratory
Princeton University
Princeton, NJ 08542

ABSTRACT

This paper proposes benchmarks for systems of automatic sense identification. A textual corpus in which open-class words had been tagged both syntactically and semantically was used to explore three statistical strategies for sense identification: a guessing heuristic, a most-frequent heuristic, and a co-occurrence heuristic. When no information about sense-frequencies was available, the guessing heuristic using the numbers of alternative senses in WordNet was correct 45% of the time. When statistics for sense-frequencies were derived from a semantic concordance, the assumption that each word is used in its most frequently occurring sense was correct 69% of the time; when that figure was calculated for polysemous words alone, it dropped to 58%. And when a co-occurrence heuristic took advantage of prior occurrences of words together in the same sentences, little improvement was observed. The semantic concordance is still too small to estimate the potential limits of a co-occurrence heuristic.

1. INTRODUCTION

It is generally recognized that systems for automatic sense identification should be evaluated against a null hypothesis. Gale, Church, and Yarowsky [1] suggest that the appropriate basis for comparison would be a system that assumes that each word is being used in its most frequently occurring sense. They review the literature on how well word-disambiguation programs perform; as a lower bound, they estimate that the most frequent sense of polysemous words would be correct 75% of the time, and they propose that any sense-identification system that does not give the correct sense of polysemous words more than 75% of the time would not be worth serious consideration.

The value of setting such a lower bound is obvious. However, Gale, Church, and Yarowsky [1] do not make clear how they determined what the most frequently occurring senses are. In the absence of such information, a case can be made that the lower bound should be given by the proportion of monosemous words in the textual corpus.

Although most words in a dictionary have only a single sense, it is the polysemous words that occur most frequently in speech and writing. This is true even when we ignore the small set of highly polysemous closed-class words (pronouns, prepositions, auxiliary verbs, etc.) that play such an important structural role. For example, 82.3% of the open-class words in WordNet [2] are monosemous, but only 27.2% of the open-class words in a sample of 103 passages from the Brown Corpus [3] were monosemous. That is to say, 27% of the time no decision would be needed, but for the remaining 73% of the open-class words, the response would have to be "don't know." This is probably the lowest lower bound anyone would propose, although if the highly polysemous, very frequently used closed-class words were included, it would be even lower.

A better performance figure would result, of course, if, instead of responding "don't know," the system were to guess. What is the percentage correct that you could expect to obtain by guessing?

2. THE GUESSING HEURISTIC

A guessing strategy presumes the existence of a standard list of words and their senses, but it does not assume any knowledge of the relative frequencies of different senses of polysemous words. We adopted the lexical database WordNet [2] as a convenient on-line list of open-class words and their senses. Whenever a word is encountered that has more than one sense in WordNet, a system with no other information could do no better than to select a sense at random.

The guessing heuristic that we evaluated was defined as follows: on encountering a noun (other than a proper noun), verb, adjective, or adverb in the test material, look it up in WordNet. If the word is monosemous (has a single sense in WordNet), assign that sense to it. If the word is polysemous (has more than one sense in WordNet), choose a sense at random with a probability of 1/n, where n is the number of different senses of that word.

This guessing heuristic was then used with the sample of 103 passages from the Brown Corpus. Given the distribution of open-class words in those passages and the number of senses of each word in WordNet, estimating the probability of a correct sense identification is a straightforward calculation. The result was that 45.0% of the 101,284 guesses would be correct. When the percent correct was calculated for just the 76,067 polysemous word tokens, it was 26.8%.

3. THE MOST-FREQUENT HEURISTIC

Data on sense frequencies do exist. During the 1930s, Lorge [4] hired students at Columbia University to count how often each of the senses in the *Oxford English Dictionary* occurred in some 4,500,000 running words of prose taken from magazines of the day. These and other word counts were used by Thorndike in writing the *Thorndike-Barnhart Junior Dictionary* [5], a dictionary for children that first appeared in 1935 and that was widely used in the public schools for many years. Not only was Thorndike able to limit his dictionary to words in common use, but he was also able to list senses in the order of their frequency, thus insuring that the senses

* Hunter College and Graduate School of the City University of New York

he included would be the ones that children were most likely to encounter in their reading. The Lorge-Thorndike data, however, do not seem to be available today in a computer-readable form.

More recently, the editors of *Collins COBUILD Dictionary of the English Language* [6] made use of the 20,000,000-word COBUILD corpus of written English to insure that the most commonly used words were included. Entries in this dictionary are organized in such a way that, whenever possible, the first sense of a polysemous word is both common and central to the meaning of the word. Again, however, sense-frequencies do not seem to be generally available in a computer-readable form.

At the ARPA Human Language Technology Workshop in March 1993, Miller, Leacock, Tengi, and Bunker [7] described a semantic concordance that combines passages from the Brown Corpus [3] with the WordNet lexical database [2] in such a way that every open-class word in the text (every noun, verb, adjective, or adverb) carries both a syntactic tag and a semantic tag pointing to the appropriate sense of that word in WordNet. The version of this semantic concordance that existed in August 1993, incorporating 103 of the 500 passages in the Brown Corpus, was made publicly available, along with version 1.4 of WordNet to which the passages were tagged.[1] Passages in the Brown Corpus are approximately 2,000 words long, and average approximately 1,000 open-class words each. Although this sample is much smaller than one would like, this semantic concordance does provide a basis for estimating sense frequencies for open-class words broken down by part of speech (word/pos). For example, there are seven senses of the word "board" as a noun (board/n1, board/n2, . . . , board/n7), and four senses as a verb (board/v1, board/v2, . . . , board/v4); the frequencies of all eleven senses in the semantic concordance can be tabulated separately to determine the most frequent board/n and the most frequent board/v.

The fact that the words that occur most frequently in standard English tend to be the words that are most polysemous creates a bad news, good news situation. The bad news is that most of the content words in textual corpora require disambiguation. The good news is that polysemous words occur frequently enough that statistical estimates are possible on the basis of relatively small samples. It is possible, therefore, to pose the question: on the basis of the available sample, how often would the most frequent sense be correct? A larger semantic concordance would undoubtedly yield a more precise lower bound, but at least an approximate estimate can be obtained.

The most-frequent heuristic was defined as follows: on encountering a noun, verb, adjective, or adverb in the test material, look it up in WordNet. If the word is monosemous, assign that sense to it. If the syntactically tagged word (word/pos) has more than one sense in WordNet, consult the semantic concordance to determine which sense occurred most often in that corpus and assign that sense to it; if there is a tie, select one of the equally frequent senses at random. If the word is polysemous but does not occur in the semantic concordance, choose a sense at random with a probability of 1/n, where n is the number of different senses of that word in WordNet.

In short, when there are data indicating the most frequent sense of a polysemous word, use it; otherwise, guess.

[1] Via anonymous ftp from clarity.princeton.edu.

3.1 A Preliminary Experiment

In order to obtain a preliminary estimate of the accuracy of the most-frequent heuristic, a new passage from the Brown Corpus (passage P7, an excerpt from a novel that was classified by Francis and Kučera [3] as "Imaginative Prose: Romance and Love Story") was semantically tagged to use as the test material. The training material was the 103 other passages from the Brown Corpus (not including P7) that made up the semantic concordance. The semantic tags assigned by a human reader were then compared, one word at a time, with the sense assigned by the most-frequent heuristic.

For this particular passage, only 62.5% of the open-class words were correctly tagged by the most-frequent heuristic. This estimate is generous, however, since 24% of the open-class words were monosemous. When the average is taken solely over polysemous words, the most frequent sense was right only 50.8% of the time.

These results were lower than expected, so we asked whether passage P7 might be unusual in some way. For example, the sentences were relatively short and there were fewer monosemous words than in an average passage in the training material. However, an inspection of these data did not reveal any trend as a function of sentence length; short sentences were no harder than long ones. And the lower frequency of monosemous words is consistent with the non-technical nature of the passage; there is no obvious reason why that should influence the results for polysemous words. Without comparable data for other passages, there is no way to know whether these results for P7 are representative or not.

3.2 A Larger Sample

Rather than tag other new passages to use as test material, we decided to use passages that were already tagged semantically. That is to say, any tagged passage in the semantic concordance can be made to serve as a test passage by simply eliminating it from the training material. For example, in order to use passage X as a test passage, we can delete it from the semantic concordance; then, using this diminished training material, the most-frequent heuristic is evaluated for passage X. Next, X is restored, Y is deleted, and the procedure repeats. Since there are 103 tagged passages in the semantic concordance, this produces 103 data points in addition to the one we already have for P7.

Using this procedure, the average number of correct sense identifications produced by the most-frequent heuristic is 66.9% (standard deviation, σ = 3.7%) when all of the open-class words, both monosemous and polysemous, are included. When only polysemous words are considered, the average drops to 56.4% (σ = 4.3%). This larger sample shows that the results obtained from the preliminary experiment with passage P7 were indeed low, more than a standard deviation below the mean.

The scores obtained when the most-frequent heuristic is applied to these 2,000-word passages appear to be normally distributed. Cumulative distributions of the scores for all 104 passages are shown in Figure 1. Separate distributions are shown for all open-class words (both monosemous and polysemous) and for the polysemous open-class words alone.

No doubt some of this variation is attributable to differences in genre between passages. Table 1 lists the 15 categories of prose sampled by Francis and Kučera [5], along with the number of passages of each type in the semantic concordance and the average

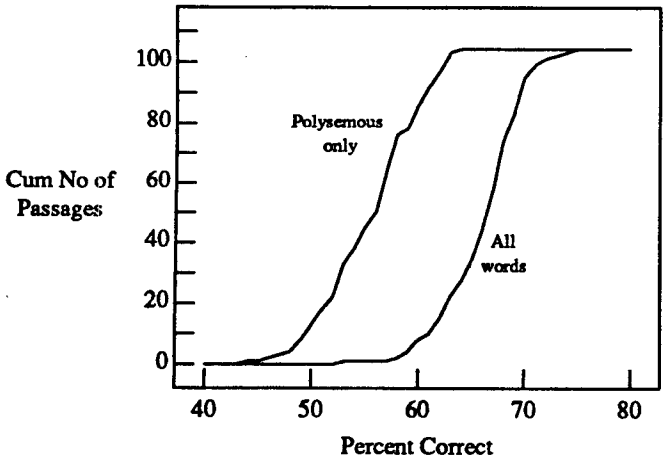

Fig. 1. Cumulative distributions of percent correct when the most-frequent heuristic is applied to 104 passages from the Brown Corpus.

percentage correct according to the most-frequent heuristic. The passages of "Informative Prose" (A through J) tend to give lower scores than the passages of "Imaginative Prose" (K through R), suggesting that fiction writers are slightly more likely to use words in their commonest senses. But the differences are small.

Table 1
Mean percent correct for genres recognized by Francis and Kučera.

Genre	N	All Words	Polysemous
A. Press: Reportage	7	69	60
B. Press: Editorial	2	63	51
C. Press: Reviews	3	64	54
D. Religion	4	62	52
E. Skills and Hobbies	6	63	53
F. Popular Lore	4	66	54
G. Belles Lettres	3	64	52
H. Miscellaneous (reports)	1	62	50
J. Learned (science)	33	66	55
K. General Fiction	29	69	59
L. Detective Fiction	2	68	58
M. Science Fiction	2	68	57
N. Western Fiction	1	68	59
P. Romance and Love Story	2	67	55
R. Humor	5	69	58

3.3 Effects of Guessing

As the most-frequent heuristic is defined above, when a polysemous open-class word is encountered in the test material that has not occurred anywhere in the training material, a random guess at its sense is used. Such cases, which lower the average scores, are a necessary but unfortunate consequence of the relatively small sample of tagged text that is available; with a large sample we should have sense frequencies for all of the polysemous words. However, we can get some idea of how significant this effect is by simply omitting all instances of guessing, i.e., by basing the percentage correct only on those words for which there are data available in the training material.

When guesses are dropped out, an improvement of approximately 2% is obtained. That is to say, the mean for all substantive words increases from 66.9% to 69.0% (σ = 3.8%), and the mean for polysemous words alone increases from 56.4% to 58.2% (σ = 4.5%).

We take these values to be our current best estimates of the performance of a most-frequent heuristic when a large database is available. Stated differently: any sense identification system that does no better than 69% (or 58% for polysemous words) is no improvement over a most-frequent heuristic.

4. THE CO-OCCURRENCE HEURISTIC

The criterion of correctness in these studies is agreement with the judgment of a human reader, so it should be instructive to consider how readers do it. A reader's judgments are made on the basis of whole phrases or sentences; senses of co-occurring words are allowed to determine one another and are identified together. The general rule is that only senses that suit all of the words in a sentence can co-occur; not only does word W_1 constrain the sense of another word W_2 in the same sentence, but W_2 also constrains the sense of W_1. That is what is meant when we say that context guides a reader in determining the senses of individual words. Given the importance of co-occurring senses, therefore, we undertook to determine whether, on the basis of the available data, co-occurrences could be exploited for sense identification.

In addition to information about the most frequent senses, a semantic concordance also contains information about senses that tend to occur together in the same sentences. It is possible to compile a semantic co-occurrence matrix: a matrix showing how often the senses of each word co-occur in sentences in the semantic concordance. For example, if the test sentence is "The horses and men were saved," we search the semantic co-occurrence matrix for co-occurrences of horse/n and man/n, horse/n and save/v, and man/n and save/v. This search reveals that the fifth sense of the noun horse, horse/n5, co-occurred twice in the same sentence with man/n2 and four times with man/n6, but neither horse/n nor man/n co-occurred in the same sentence with save/v. If we then take the most frequent of the two co-occurring senses of man/n, we select man/n2. But no co-occurrence information is provided as to which one of the 7 senses of save/v should be chosen; for save/v it is necessary to resort to the most frequent sense, as described above.

The co-occurrence heuristic was defined as follows. First, compile a semantic co-occurrence matrix. That is to say, for every word-sense in the semantic concordance, compile a list of all the other word-senses that co-occur with it in any sentence. Then, on encountering a noun, verb, adjective, or adverb in the test material, look it up in WordNet. If the word is monosemous, assign that sense to it. If the word has more than one sense in WordNet, consult the semantic co-occurrence matrix to determine what senses of the word co-occur in the training material with other words in the test sentence. If only one sense of the polysemous word co-occurrs in the training material with other words in the test sentence, assign that sense to it. If more than one sense of the polysemous word co-occurs in the training material with other words in test sentence, select from among the co-occurring senses the sense that is most frequent in the training material; break ties by a random choice. If the polysemous word does not co-occur in the training material with other words in the test sentence, select the sense that is most

frequent in the training material; break ties by a random choice. And if the polysemous word does not occur at all in the training material, choose a sense at random with a probability of 1/n.

In short, where there are data indicating co-occurrences of senses of polysemous words, use them; if not, use the most-frequent heuristic; otherwise, guess.

When this co-occurrence heuristic was applied to the 104 semantically tagged passages, the results were almost identical to those for the most-frequent heuristic. Means using the co-occurrence heuristic were perhaps a half percent lower than those obtained with the most-frequent heuristic. And when the effects of guessing were removed, an improvement of approximately 2% was obtained, as before. This similarity can be attributed to the limited size of the semantic concordance: no co-occurrence data were available for 28% of the polysemous words, so the most-frequent heuristic had to be used; moreover, those words for which co-occurrence data were available tended to occur in their most frequent senses.

On the basis of results obtained with the available sample of semantically tagged text, therefore, there is nothing to be gained by using the more complex co-occurrence heuristic. Since context is so important in sense identification, however, we concluded that our semantic concordance is still too small to estimate the potential limits of a co-occurrence heuristic.

5. SUMMARY AND CONCLUSIONS

The considerable improvement that results from having knowledge of sense frequencies is apparent from the results summarized in Table 2, where the guessing heuristic is contrasted with the most-frequent and co-occurrence heuristics (with guessing removed).

Table 2
Percent correct sense identifications for open-class words without and with information on sense frequencies.

Heuristic	Monosemous and Polysemous	Polysemous only
Guessing	45.0	26.8
Most frequent	69.0	58.2
Co-occurrence	68.6	57.7

The similarity of the results obtained with the most-frequent and the co-occurrence heuristics is attributable to the fact that when co-occurrence data were indeterminate or lacking, the most-frequent heuristic was the default. With a large semantic concordance, we would expect the co-occurrence heuristic to do better—it should be able to capture the topical context which, in other work [8], we have found to give scores as high as 70-75% for polysemous words.

How representative are the percentages in Table 2? Obviously, they are specific to the Brown Corpus; in a restricted domain of discourse, polysemous words would not be used in such a wide variety of ways and a most-frequent heuristic would be correct far more frequently. The percentages in Table 2 are "broadly representative of current edited American English" [3]. They are also, of course, specific to WordNet. If WordNet did not draw so many sense distinctions, all of these statistical heuristics would be correct more often. But WordNet does not draw impossibly fine sense distinctions. Dictionaries differ widely in the number of sense distinctions they draw; pocket dictionaries offer few and unabridged dictionaries offer many alternative senses. WordNet is somewhere in the middle; it provides about the same semantic granularity as a good desk dictionary. Anything coarser could not have been used to tag passages from the Brown Corpus.

Finally, can these heuristics provide anything more than benchmarks? Can they play a role in a system that does an acceptable job of sense identification? It should be noted that none of these heuristics takes into account the local context. Even the co-occurrence heuristic is indifferent to word order; imposing word-order constraints would have made sparse data sparser still. Local context—say, ± 2 or 3 words—should contain sufficient information to identify the intended sense of most polysemous words. Given a system capable of exploiting local context, statistical heuristics might still provide a default, as Yarowsky [9] suggests; something to fall back on when local identification fails. Under those conditions, these statistical heuristics could indeed provide a floor on which more intelligent systems could build.

ACKNOWLEDGMENTS

This work has been sponsored in part by Grant No. N00014-91-J-1634 from the Advanced Research Projects Agency, Information and Technology Office, and the Office of Naval Research, and in part by grants from the James S. McDonnell Foundation, from The Pew Charitable Trusts, from the Linguistic Data Consortium, and from Sun Microsystems. We are indebted to Henry Kučera and W. Nelson Francis for permission to use the Brown Corpus in our research. And we are indebted for assistance and advice to Ross T. Bunker, Christiane Fellbaum, Benjamin Johnson-Laird, Katherine Miller, Randee Tengi, Pamela Wakefield, and Scott Wayland.

REFERENCES

1. Gale, W., Church, K. W., and Yarowsky, D. (1992) Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 249-256.
2. Miller, G. A., Ed. (1990) Five Papers on WordNet. *International Journal of Lexicology*, 3, No. 4. (Revised, March 1993)
3. Francis, W. N., and Kučera, H. (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, MA: Houghton Mifflin.
4. Lorge, I. (1937) The English semantic count. *Teachers College Record*, 39, 65-77.
5. Thorndike, E. L., and Barnhart, C. L., Eds. (1935) *Thorndike-Barnhart Junior Dictionary*. Glenview, IL: Scott Foresman.
6. *Collins COBUILD English Language Dictionary*. (1987) London: Collins.
7. Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. (1993) A semantic concordance. *Proceedings of a Human Language Technology Workshop*, pp. 303-308.
8. Leacock, C., Towell, G., and Voorhees, E. (1993) Corpus-based statistical sense resolution. *Proceedings of a Human Language Technology Workshop*, pp. 260-265.
9. Yarowsky, D. (1993) One sense per collocation. *Proceedings of a Human Language Technology Workshop*, pp. 266-271.