More than One Sense Per Discourse

Robert Krovetz
NEC Research Institute
Princeton, NJ 08540
krovetz@research.nj.nec.com

Abstract

Previous research has indicated that when a polysemous word appears two or more times in a discourse, it is extremely likely that they will all share the same sense [Gale et al. 92]. However, those results were based on a coarse-grained distinction between senses (e.g, *sentence* in the sense of a 'prison sentence' vs. a 'grammatical sentence'). We report on an analysis of multiple senses within two sense-tagged corpora, Semcor and DSO. These corpora used WordNet for their sense inventory. We found significantly more occurrences of multiple-senses per discourse than reported in [Gale et al. 92] (33% instead of 4%). We also found classes of ambiguous words in which as many as 45% of the senses in the class co-occur within a document. We discuss the implications of these results for the task of word-sense tagging and for the way in which senses should be represented.

1 Introduction

When a word appears more than once in a discourse, how often does it appear with a different meaning? This question is important for several reasons. First, the interaction between lexical semantics and discourse provides information about how word meanings relate to a larger context. In particular, the interaction provides a better understanding of the types of inferences involved. Second, by looking at word senses that systematically co-occur within a discourse we get a better understanding of the distinction between homonymy and polysemy (unrelated vs. related word senses).[1] Word senses that co-occur are more likely to be related than those that are not. Finally,

[1] For example, *race* is homonymous in the sense of 'human race' vs. 'horse race'. *Door* is polysemous in the contexts 'paint the door' vs. 'go through the door'.

the question is important for word sense tagging. If a word appears with only one meaning in a discourse then we can disambiguate only one occurrence and tag the rest of the instances with that sense.

Prior work on the number of senses per discourse was reported in [Gale et al. 92]. Their work was motivated by their experiments with word sense disambiguation. They noticed a strong relationship between discourse and meaning and they proposed the following hypothesis: *When a word occurs more than once in a discourse, the occurrences of that word will share the same meaning.*

To test this hypothesis they conducted an experiment with five subjects. Each subject was given a set of definitions for 9 ambiguous words and a total of 82 pairs of concordance lines for those words. The subjects were asked to determine for each pair whether they corresponded to the same sense or not. The researchers selected 54 pairs from the same discourse and 28 were used as a control to force the judges to say they were different. The control pairs were selected from different discourses and were checked by hand to assure that they did not use the same sense. The result was that 51 of the 54 pairs were judged to be the same sense (by a majority opinion). Of the 28 control pairs, 27 were judged to be different senses. This gave a probability of 94% (51/54) that two ambiguous words drawn from the same discourse will have the same sense. [Gale et al. 92] then assumed that there is a 60/40 split between unambiguous/ambiguous words, so there is a 98% probability that two word occurrences in the same discourse will have the same sense.

[Gale et al. 92] suggested that these results could be used to provide an added constraint for improving the performance of word-sense disambiguation algorithms. They also proposed that it be used to help evaluate sense tagging. Only one instance of the word in a discourse would need to be tagged and the remaining instances could be tagged automatically with the same sense. This would provide a much larger set of training instances, which is a central problem for disambiguation.

In our own experiments with disambiguation we found a number of instances where words appeared in the same document with *more* than one meaning [Krovetz and Croft 92]. These observations were based on experiments with two corpora used in information retrieval. One corpus consisted of titles and abstracts from Communications of the ACM (a Computer Science journal). The other corpus consisted of short articles from TIME magazine. In the CACM corpus a word rarely appeared more than once in a document (since the documents were so short). However, in the TIME corpus we found a number of cases where words appeared in the same document with more than one meaning. A sample of these words is given below:

party dinner party / political party

headed headed upriver / headed by

great great grandson / Great Britain
great Irishmen / Great Britain

park Industrial park / Dublin's park
Industrial park / parking meter

line a line drawn by the U.S. / hot line

We even found one instance in which five different senses of a word occurred within the same document: 'mile long cliff *face*', 'difficulties ... is *facing* because', 'in the *face* of temptations', 'about *face*', and 'his pavilion *facing* lovely west lake'[2]

[Gale et al. 92]'s hypothesis raises the question: What is a *sense*? Most of the work on sense-disambiguation has focused on meanings that are unrelated, the so-called 'Bank model' (river bank vs. savings bank). But in practice word senses are often related. Unrelated senses of a word are *homonymous* and related senses are termed *polysemous*.[3] In [Gale et al. 92]'s experiments they asked the subjects to determine whether the pairs of concordance lines exhibited the same sense or not. But human judgement will vary depending on whether the senses are homonymous or polysemous [Panman 82]. People will often agree about the sense of a word in context when the senses are unrelated (e.g., we expect that people will reliably tag 'race' in the sense of a horse race vs. human race), but people will disagree when the senses are related.

The disagreement between individuals about polysemous senses might be considered an impediment, but we prefer to view it as a source of data. We can use the judgements to help distinguish homonymous from polysemous senses. When the judgements are *systematically* inconsistent, we predict that the senses will be polysemous. In other words, the inconsistency in human judgement (with respect to determining the meaning of a word in context) can be viewed as a feature rather than a bug.

In addition, there are a variety of tests to help establish word sense identity. For example, we can conjoin two senses and note the anomaly (zeugma): "The newspaper fired its employees and fell off the table" [Cruse 86]. We can also determine whether a word is a member of a class that is systematically ambiguous (e.g., language/people or object/color - see [Krovetz 93]).

[Gale et al. 92]'s hypothesis also raises the question: What is a *discourse*? Is it a paragraph, a newspaper article, a document that is about a given topic, or something else? How do the concepts of discourse and topic relate to each other? Research on topic segmentation [Hearst 97] and work on text coherence [Morris and Hirst 91] addresses this question. We can't provide an answer to how this work affects [Gale et al. 92]'s hypothesis, but the question of what constitutes a discourse is central to its testability.

[2]These examples illustrate a difference from other work on word meanings. Most of that work has not considered any morphological variants for a word or differences across part of speech.

[3]The word *polysemy* is itself polysemous. In general usage it is a synonym for lexical ambiguity, but in linguistics it refers to senses that are related.

This paper is concerned with the first question we raised - how does word sense identity affect [Gale et al. 92]'s results? In particular, what happens if we consider the distinction between homonymy and polysemy? We conducted experiments to determine whether [Gale et al. 92]'s hypothesis would hold when applied to finer grained sense distinctions. These experiments are described in the following section.

2 Experiments

Our experiments used two sense-tagged corpora, *Semcor* [Miller et al. 94] and *DSO* [Ng and Lee 96]. Both of these corpora used WordNet as a basis for the sense inventory [Miller 1990]. WordNet contains a large number of words and senses, and is comparable to a good collegiate dictionary in its coverage and sense distinctions. Semcor is a semantic concordance in which all of the open class words[4] for a subset of the Brown corpus[5] were tagged with the sense in WordNet. The DSO corpus is organized differently from Semcor. Rather than tag all open-class words, it consists of a tagging of 191 highly ambiguous words in English within a number of files. These files are drawn from the Brown corpus and the Wall Street Journal. The 191 words are made up of 121 nouns and 70 verbs.

We conducted experiments to determine how often words have more than one meaning per discourse in the two sense-tagged corpora. This was defined as more than one WordNet sense tag in a file from the Brown corpus (for Semcor) and in a file from either the Brown Corpus or the Wall Street Journal for DSO.

For Semcor we wrote a program to identify all instances in which a tagged word occurred in a file from the Brown corpus with more than one sense. The program determined the potential ambiguity of these words (the number of senses they had in WordNet) as well as the actual ambiguity (the number of senses for those words in Semcor). We then computed the proportion of the ambiguous words within the corpus that had more than one sense in a document.

For the DSO corpus we determined how many of the tagged words had more than one sense in a document. We also determined how many documents contained an instance of the tagged word with more than one sense.

3 Results

The statistics for the experiment are given in Table 1. We indicate the number of unique words with a breakdown according to part of speech. We also show the number of words that have more

[4] Nouns, verbs, adjectives, and adverbs.

[5] The Brown corpus consists of 500 discourse fragments of 2000 words, each.

	Nouns	Verbs	Adj
Word Types	8451	3296	1521
Potential ambiguity	4016	2161	962
Actual ambiguity	1659	1089	169
Multiple Sense/Discourse	517	365	55

Table 1: Statistics on multiple-senses within a discourse for Semcor. *Potential ambiguity* refers to the number of unique words that have more than one sense in WordNet. *Actual ambiguity* is the number of those words that were found to have more than one sense within the tagged corpus.

than one sense in WordNet (potential ambiguity) and the number that have more than one sense in the corpus (actual ambiguity). Finally, we indicate the number of words that have more than one sense per discourse.

The statistics provide a strong contrast with the results from [Gale et al. 92]. About 33% of the ambiguous words in the corpus had multiple senses within a discourse. There was no difference in this respect for the different parts of speech.

However, the statistics do show significant differences between the different parts of speech with regard to potential vs. actual ambiguity. The proportion of ambiguous words in WordNet [potential ambiguity] was 47% for nouns, 66% for verbs, and 63% for adjectives. The proportion of potentially ambiguous words that were found to be ambiguous in the corpus was 41%, 50% and 18% for nouns, verbs, and adjectives (respectively). We do not have any explanation for why the actual ambiguity for adjectives is so low.

We also examined words that were ambiguous with regard to part-of-speech. There were 752 words in Semcor that were ambiguous between noun and verb. Of these words, 267 (36%) appeared in a document in both forms. There were 182 words that were ambiguous between noun and adjective. Of these words, 82 (45%) appeared in a document in both forms.

The results with the DSO corpus support the findings with Semcor. *All* of the 191 words were found to occur in a discourse with more than one sense. On average, 39% of the files containing the tagged word had occurrences of the word with different senses.

4 Analysis

When two senses co-occur in a discourse it is possible that the co-occurrence is accidental. We therefore examined those senses that co-occured in four or more files (for nouns) and three or more

files (for verbs and adjectives).

For nouns, the systematic sense co-occurrences were primarily due to logical polysemy [Apresjan 75], [Pustejovsky 95] or to general/specific sense distinctions. A sample of these co-occurrences is given below[6]:

Logical Polysemy

agent/entity (city, school, church)
meal/event (dinner)
language/people (Assyrian, English)
figure/ground (door)
result/process (measurement)
metonymy (sun, diameter)

General/Specific

day (solar *day*/mother's *day*)
question (the *question* at hand/ask a *question*)
man (race of *man*/bearded *man*)

The figure/ground ambiguity refers to *door* as a physical object or to the space occupied by the door. The metonymic ambiguity for *sun* refers to the physical object as opposed to the rays of the sun. For *diameter* we can refer to the line or to the length of the line.

For verbs, the sense co-occurrences were more difficult to characterize. They generally seemed like active/passive distinctions. For example:

see 'We saw a number of problems' (recognize)
'We saw the boat' (perceive)

know 'know a fact' (be-convinced-of)
'know the time' (be-aware-of)

remember 'remember to bring the books' (keep-in-mind)
'remember when we bought the books' (recollect)

For adjectives the different senses reflect either differing dimensions, or absolute/relative distinctions:

[6]Some of the examples occurred in less than four files, but we mention them because they help to illustrate the members of the class.

old not young vs. not new

long spatial vs. temporal

little not big vs. not much

same identical vs. similar

The noun/verb ambiguities often reflected a process/result difference (e.g., *smile*, *laugh*, or *name*). The noun/adjective ambiguities represent a number of systematic classes:

nationality or religion British, German, Catholic, American, Martian (!)

belief humanist, liberal, positivist

made-of chemical, liquid, metal

gradable-scale quiet, young, cold

We note that there are some cases where multiple senses *might* have been identified, but WordNet was not consistent in the distinctions in meaning. For example, *dinner* has the meal vs. event distinction, but the same ambiguity was not represented for *lunch* or *breakfast*. *Assyrian*, and *English* have the language/people distinction, but these senses were not provided for *Dutch* or *Korean*. These omissions are not a criticism against WordNet per se - dictionaries are not designed to contain systematic sense distinctions whenever we have logical polysemy. In our work with the Longman Dictionary [Procter 78] we noticed a number of cases where sense distinctions were not made systematically. These inconsistencies are a reflection of human judgement with regard to polysemy.

The polysemous relations we found for isolated words were also found for lexical phrases. Although phrases usually have only one meaning,[7] we found instances in which they occurred with more than one sense within a discourse. Out of eight ambiguous lexical phrases in Semcor,[8] three occurred with more than one sense in a discourse. These phrases were: *United States* (country vs. government), *interior design* (branch of architecture vs. occupation), and *New York* (city vs. state). The first two instances are similar to other classes of logical polysemy that have been reported in the literature. The country vs. government distinction is akin to the difference between *white*

[7] This generalization is not true for phrasal verbs (verb-particle constructions).

[8] These phrases are all nouns. We also noticed senses of verbs that co-occurred. However, it is especially difficult to analyze phrasal lexemes because they occur less frequently than isolated words. Co-occurrences for particular senses are even more infrequent.

house as a physical entity vs. as an agent ('He entered the White House' vs. 'The White House dismissed the chief prosecutor'). The ambiguity between fields of knowledge and occupations is also common. Although lexical phrases have less ambiguity than isolated words, we observe that the different senses can still co-occur.

The co-occurrence of multiple senses within a discourse can be used as evidence for lexical semantic relations, and to help distinguish homonymy from polysemy. So *quack* as a noun and as a verb are related in the sense of a sound made by a duck, but not in the sense of a bad doctor. This is akin to gravity/gravitation being related in the sense of 'the force of gravity', but not with regard to the 'gravity of the offense'. In our earlier work we established links between senses in the dictionary by looking for words which occurred in their own definition, but with a different part of speech. We in essence treated dictionary definitions as a small "discourse" (we can even find deictic relationships between dictionary definitions - see [Krovetz 93]). The hypothesis is that if senses co-occur within a discourse they will be related even if they differ in part-of-speech. For example, we would predict that *paint* as a noun and as a verb will co-occur in a discourse much more often than *train* as a noun and as a verb.

We can learn about lexical semantic relations by examining dictionary definitions of related senses. For example, the relationship between *dust* as a noun and as a verb can be one of covering or removing. The dictionary tells us that it has both meanings.

The biggest problem we encountered in our analysis was the number of tagged files. We wanted to ensure that the sense co-occurrences were not simply an accident, so we looked for sense pairs that co-occurred in several files. But the existing tagged corpora are not large enough to get reliable statistics. *Dust* as a verb only appears twice out of the 106,000 tagged word forms in Semcor. This is not often enough to get statistics about co-occurrence with a noun, much less co-occurrence with specific senses.

5 Conclusions and Future Work

[Gale et al. 92]'s hypothesis is probably correct for homonymous senses. It is unlikely that a document which mentions *bank* in the sense of a river bank will also use it in the sense of a savings bank. However, even with homonymous senses, we expect there will be certain cases that will predictably co-occur. For example, in legal documents *support* in the sense of *child support* can co-occur with *support* in the sense of supporting an argument. The work reported in this paper shows that the hypothesis is not true with regard to senses that are polysemous.

We do not want to give the impression that the distinction between homonymy and polysemy is straightforward. It is not. In practice the differences in meaning are not always clear. But that does not mean that the distinction between homonymy and polysemy is vacuous. We gain a

better understanding of the difference by looking at systematic classes of ambiguity. Another set of semantically tagged files was just released.[9] These files will allow us to examine a larger number of words in which the multiple senses co-occurrences are systematic.

Our results indicate that we cannot simply adopt [Gale et al. 92]'s suggestion that we disambiguate one occurrence of a word in a discourse and then assign that sense to the other occurrences. However, we *can* leverage the systematic classes of ambiguity. If a word appears in a discourse and there are senses of that word that are systematically polysemous, we can attempt to tag the other occurrences in the discourse in light of this ambiguity. In the future we will examine rules associated with classes of polysemous words that will allow these occurrences to be tagged.

Acknowledgements

I am grateful to David Lebeaux and Christiane Fellbaum for their comments on this paper.

References

[Apresjan 75] Apresjan Ju, "Regular Polysemy", *Linguistics*, Vol. 142, pp. 5-32, 1975.

[Cruse 86] Cruse David, *Lexical Smantics*, Cambridge University Press, 1986.

[Gale et al. 92] Gale William, Kenneth Church, and David Yarowsky, "One Sense Per Discourse", in *Proceedings of the ARPA Workshop on Speech and Natural Language Processing*, pp. 233–237, 1992.

[Hearst 97] Hearst Marti, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", *Computational Linguistics*, Vol. 23(1), pp. 33–64, 1997.

[Krovetz and Croft 92] Krovetz Robert and W. Bruce Croft, "Lexical Ambiguity and Information Retrieval", *ACM Transactions on Information Systems*, pp. 145–161, 1992.

[Krovetz 93] Krovetz Robert, "Sense Linking in a Machine-Readable Dictionary", in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 330–332, 1993.

[Miller 1990] Miller George, "WordNet: An on-line Lexical Database", *International Journal of Lexicography*, Vol. 3(4), pp. 235-312, 1990.

[9] Brown2, which consists of an additional 83 files.

[Miller et al. 94] Miller George, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert Thomas, "Using a Semantic Concordance for Sense Identification", in *Proceedings of the ARPA Human Language Technology Workshop*, 1994.

[Morris and Hirst 91] Morris Jane and Graeme Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", *Computational Linguistics*, Vol. 17(1), pp. 21–48, 1991.

[Ng and Lee 96] Ng Hwee Tou and Hian Beng Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach", in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 40–47, 1996

[Panman 82] Panman Otto, "Homonymy and Polysemy", *Lingua*, Vol. 58, pp. 105–136, 1982

[Procter 78] Procter Paul, *Longman Dictionary of Contemporary English*, Longman, 1978.

[Pustejovsky 95] Pustejovsky James, *The Generative Lexicon*, MIT Press, 1995

[Yarowsky 92] Yarowsky David, "Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in *Proceedings of the 14th Conference on Computational Linguistics, COLING-92*, pp. 454–450, 1992.