BUILDING SEMANTIC CONCORDANCES: DISAMBIGUATION VS. ANNOTATION

George A. Miller
Cognitive Science Laboratory
Princeton University

1 Abstract

A semantic concordance has been defined by Miller, Leacock, Tengi, and Bunker (1993) as "a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon." According to this definition, a semantic concordance can be viewed either as a corpus of disambiguated text or as a lexicon in which example sentences are available for many definitions.

At Princeton we have now had more than two years' experience trying to build such interconnected databases. We have used WordNet (Miller, 1990; Miller and Fellbaum, 1991) as the lexical component, and the Brown Corpus (Kucera and Francis, 1967; Francis and Kucera, 1982) as the text. WordNet is a lexical database in which sets of synonyms represent lexicalized concepts and semantic relations between words and concepts are represented by bidirectional pointers; the Brown Corpus is a collection of 500 passages (each 2,000 words long) that are representative of published American writing in the 1960s. This paper reports some of our successes and explores some of the problems that we have encountered.

2 Successes

First, the successes. We have created an interface, ConText, that displays text with a word highlighted and, below, gives all the WordNet senses for that word (Leacock, 1993). A reader selects the sense that is judged appropriate in the given context, indicates the choice with the cursor, and the highlight moves on to the next substantive word. The details of this interface have evolved considerably as we have gained experience with the task, but the important fact is that Princeton students are able to use ConText to insert pointers from the substantive words in a text to their appropriate senses in WordNet. The work goes slowly and quality control is a persistent problem, but with intelligent taggers and patient checking for errors it is possible to create a corpus of semantically tagged text.

A second success is closely related to the first. Initially, we undertook the task of semantic tagging in order to check on WordNet's coverage. That is to say, we viewed the semantic concordance as a lexical database with many instances of usage, and hoped to use the instances to improve the lexical database. It is easy, of course, to discover overlooked words by checking WordNet's word list against other word lists. However, it is much more difficult to discover what word senses have been overlooked. Tagging text semantically is one way to uncover those sense omissions. As taggers discover omissions, they are reported back to the lexicographers, who add the missing information to WordNet. Then a tagger goes through the passage again, this time finding the information needed to complete the semantic tagging. As we had hoped, this procedure has progressively improved the completeness of WordNet's vocabulary. And it has also insured that WordNet would contain all of the words and word senses required to tag the text semantically.

3 Problems

It is our ambition, of course, to develop software that will be able to take over the task of human taggers. A corpus of disambiguated text should provide valuable guidance for that project–at least it should provide a way to compare any proposed system with the judgments of human readers. So far we have had little success, but we attribute that to the still small size of our semantic concordance (Miller, Chodorow, Landes, Leacock, and Thomas, 1994). A very large corpus of disambiguated text is required in order to succeed in training any practical learning system.

Of greater relevance for this Symposium, however, are the problems we have encountered in semantic tagging. The first problem is reasonably simple: you cannot define a 'word' as a string of contiguous characters surrounded by spaces. If you try to attach semantic tags to every simple word, you will quickly discover it is impossible. To take an obvious example, the word 'fountain' has two senses in WordNet and 'pen' has five; combining them gives ten possible readings for 'fountain pen,' none of which is correct. The easy solution, of course, is to enter 'fountain pen' as a compound word, and that is what we have done. We do have trouble at times deciding whether or not some familiar collocation is semantically decomposable, but we have usually leaned over backward to accept compound words. As a consequence, when we select a passage to tag semantically we must first tokenize it–scan through it for all the WordNet compounds it contains. It also means that standard word counts, which count only simple words, are not very useful to us.

A second problem is more serious, but perhaps not as serious as we feared. It is sometimes the case that the adjacent context is not sufficient to support a choice of any particular sense of a polysemous word. The author may have had a particular sense in mind, but not provided adequate cues for a reader to determine what it was. Or an author may have been deliberately ambiguous for reasons we can only guess at. In such cases, we decided to respect its ambiguity: to attach two (or more) different semantic tags to the same polysemous word. But this was not an easy decision to enforce. For one thing, taggers like to pick the one sense that they consider most plausible in the context. To overcome their resistance to multiple tags, we created an alternative method for using ConText, the tagging interface: instead of using the cursor only to designate a right sense, the tagger can use it to eliminate senses that are clearly wrong. This strategy at least produces some candidates for multiple tags. But sometimes we find that the reason a tagger cannot settle on a single tag is that the alternatives offered by WordNet are practically (sometimes actually) indistinguishable and must be edited. In short, the fact seems to be that in everyday prose writing there really are NOT many words that cannot be disambiguated.

Which is, after all, another way of saying that there is a difference between everyday prose writing and creative writing. When humanists or literary scholars hear that we are producing disambiguated text, they are sure we must have made some serious mistake. They devote serious study to alternative interpretations of ambiguous phrases in great works of literature. What would it mean, they ask, to disambiguate Joyce's "Finnegan's Wake"?

We have tried to handle some of these literary problems by creating a tag labeled 'metaphor,' and puzzled taggers do resort to it when none of the familiar senses of a word are quite right. (We probably should have called it 'trope,' but we feared that that term would not be immediately familiar.) For example, the word 'roadblock' in such a sentence as "The major roadblock to their finding jobs is...," does not refer to an obstruction set up across a road by police officers, and a tagger might well ask whether this is a metaphor. But if it is a metaphor at all, it is a frozen metaphor, and it is simpler just to add another sense for 'roadblock.' Most of the uses that taggers regarded as metaphorical have been resolved in this manner. In general, we have found it simpler to add new words and new senses as needed, rather than trying to generate them from morphological or semantic principles, and it is merely a generaliza- tion of this strategy to add new senses for frozen metaphors.

4 Annotations

There are problems, however, that cannot be disposed of so easily. In the summer of 1994 we decided to create a semantic concordance for Stephen Crane's novel, "The Red Badge of Courage." That is to say, we decided to develop a semantic concordance viewed primarily as a text with extensive semantic footnotes. We are developing a reading interface where someone can read the novel and, using a cursor or touch screen, obtain the meaning of any substantive word. I say 'the meaning' deliberately to contrast with a machine readable dictionary that would simply list alternative meanings of the word and leave it to the reader to decide which were appropriate in the given context. We undertook this project because we hope that such a semantic concordance might have educational value.

But we encountered problems of a sort that our work on the Brown Corpus had not prepared us for. The best way to describe these problems is by an example. The first chapter describes the youth's decision to enlist in the Union Army. His mother tries to dissuade him. Then Crane writes, "At last, however, he had made firm rebellion against this yellow light thrown upon the color of his ambitions." It is clear that 'this yellow light thrown upon the color of his ambitions' refers to the objections raised by his mother, but consider the problems it poses for word-by-word tagging. 'Yellow light' is defined in WordNet as a traffic signal to proceed with caution, but The Red Badge of Courage was published in 1895, many years before traffic lights were needed. Could 'yellow' mean cowardly? Or old? Perhaps 'yellow light' is sunlight. And what color is ambition? We assume that these are the kinds of questions that literary scholars like to consider, but they frustrate us.

Clearly, such problems cannot be solved by adding ad hoc senses to WordNet. Nor can we simply omit them because we cannot disambiguate them on a word-by-word basis–that is precisely where puzzled readers are most likely to request help. We have decided, therefore, to make 'annotations' available for such phrases instead of (or in addition to) WordNet definitions. The anomalous phrase can be treated as a unique collocation, and the comment on it as a unique interpretation. That is to say, if any word of the phrase 'this yellow light thrown upon the color of his ambitions' is questioned, a note discussing the meaning of the whole phrase will be made available to the reader. This solution complicates the reading interface considerably, and places on us the unwelcome burden of composing literary footnotes.

It may prove to be the case, therefore, that this work will most useful pedagogically in teaching science and mathematics, where the use of figurative language is limited. In technical prose there are likely to be many unfamiliar terms, so a disambiguated text may be especially helpful. But even in technical prose there may be places where a reader needs to be reminded of the structure of the argument. So annotations as well as definitions may be necessary even there. In short, annotations as well as definitions may be unavoidable if we hope to achieve the goal of explicating text for readers.

In sum, we have found that word-by-word disambiguation can go much further than we expected, but it has its limits. We assume that these limits would also exist for any system of machine translation.

Acknowledgments:

This research has been supported in part by Grant No. N00014-91-1634 from the Advanced Research Projects Agency and the Office of Naval Research, and by grants from the Linguistic Data Consortium and the James S. McDonnell Foundation. I am indebted to Martin Chodorow, Christiane Fellbaum, Patricia Gildea-Long, Shari Landes, Claudia Leacock, Daniel Markham, Katherine Miller, Randee Tengi, and Pamela Wakefield for their collaboration in the projects mentioned here.

5 References

Francis, W. N., and Kucera, H. Frequency Analysis of English Usage: Lexicon and Grammar. Boston, Mass.: Houghton Mifflin, 1982.

Kucera, H., and Francis, W. N. Computational Analysis of Present-Day American English. Providence, R.I.: Brown University Press, 1967.

Leacock, C. ConText: A tool for semantic tagging of text: User's guide. Cognitive Science Laboratory, Princeton University: CSL Report No. 54, February 1993.

Miller, G. A. (ed.), WordNet: An on-line lexical database. International Journal of Lexicography (special issue), 3, 235-312, 1990.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. Using a semantic concordance for sense identification. ARPA Workshop on Human Language Technology, March, 1994.

Miller, G. A., and Fellbaum, C. Semantic networks of English. Cognition (special issue), 41, 197-229, 1991.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. A semantic concordance. ARPA Workshop on Human Language Technology, March, 1993.