Word Sense Disambiguation, Lexical Semantics, and NLP Applications

Robert Krovetz
Lexical Research
Hillsborough, NJ 08844
rkrovetz@lexicalresearch.com

Ide and Wilks argue that a fine-grained division of senses may not be an appropriate goal for a computational WSD task (Ide and Wilks, 2006). They propose that NLP needs correspond roughly to homograph-level distinctions, although they acknowledge some evidence of broad level distinctions within a homograph. They further argue that machine translation requires finer-grained distinctions than information retrieval.

These arguments are problematic for the following reasons:

1. They do not provide a clear way of distinguishing the intra-homograph level distinctions that are coarse-grained from those that are not.

2. There is evidence that fine-grained distinctions **are** needed for information retrieval (IR) (Schutze and Pedersen 95). There is no evidence of a greater need for such distinctions for machine translation.

I agree that there should be a greater focus on broad-level distinctions within the community. I also agree that the organization of senses that we find in a standard dictionary is not appropriate for Computational Linguistics. But rather than couch this in terms of fine-grained or coarse-grained distinctions, I propose that we cast it in terms of lexical semantic relations. That is, how does the relationship between senses interact with the needs for NLP applications?

To get a better understanding of this question, I conducted a manual analysis of over 5700 noun and verb homographs that had exactly two non-idiomatic senses (as found in the Longman Dictionary of Contemporary English (Procter 78).[1] If I could not determine any relationship between the senses, I labeled the sense pair 'Homonymous'. If I was able to identify a relationship, I labeled the homograph accordingly (e.g., Metaphor, Process/Result, General/Specific). I did not assign a label to 6% of the homographs because the distinction was unclear. I then supplemented this classification with discourse analysis and with opinions from ten other judges about whether or not the senses were related (I did not ask the other judges to label the relationship, only to provide a judgement about whether the senses were related or not). I also elicited such judgements for the three-sense nouns and verbs.

The primary reason for the analysis is to help address the long tail problem. Of the ambiguous words, which ones should we go to the trouble of tagging, and with which sense distinctions? The two-sense homographs are much easier to deal with than homographs with more sense distinctions, and they constitute a large proportion of the homographs overall (more than 50% of the multi-sense homographs in the Longman dictionary). They are also useful because inter-annotator agreement is likely to be higher for words with a smaller number of sense distinctions. Although highly ambiguous words are more frequent than less ambiguous words, the two and three sense words constitute about 30% of the ambiguous words by token. As mentioned by Ide and Wilks, intra-homograph senses can be very different in meaning. There was a strong consensus

[1] Idioms occur infrequently in corpora (Moon 98), and I assumed there was no systematic semantic relationship between those senses and ones that are non-idiomatic. Idiomatic senses were identified by the use of different fonts compared to the regular senses.

(more than 8 out of the 10 of the judges agreed) that 13% of the intra-homograph 2-sense nouns and 20% of the 2-sense verbs were different in meaning.

Discourse analysis was done partially to test the One Sense per Discourse hypothesis (Gale et al 92), partially to help separate unrelated from related senses, and partially because of the way sense distinctions interact with IR systems. In earlier work, I found significantly more occurrences of multiple senses per discourse than reported in (Gale et al 92) (33% vs. 4%). This has important consequences for semantic acquisition, and for how those senses are used in an NLP application. (Gale et al 92) proposed their hypothesis because of the high cost for semantic annotation. If their hypothesis was correct, we would only have to label one instance of a given word per document. I found that the senses which co-occurred were senses that belonged to particular semantic classes (Krovetz 98). In my current work I am trying to determine how much correlation there is between the semantic classes and co-occurence within a document. My hypothesis is that most of the homonymous population (the intra-homograph sense pairs that were judged to be different senses) will follow the One Sense per Discourse hypothesis. That is, this set will help to reduce the number of word tokens that need to be tagged by hand.

In my talk I will discuss these results in more detail. I propose that it is not necessarily the case that machine translation requires a greater number of fine-grained distinctions than IR, but rather a different set of distinctions (cf. (Krovetz 97)). I will also discuss how my analysis has allowed me to generalize the lexical semantic relations reported in the literature, and how it affected the architecture for a disambiguation system.

References

William Gale, 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

Robert Krovetz, "More than One Sense per Discourse", Proceedigs of the ACL-SIGLEX Workshop (Senseval), 1998

Robert Krovetz, "Homonymy and Polysemy in Information Retrieval" , in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 72-79, 1997

Nancy Ide and Yorick Wilks, "Making Sense About Senses", Word Sense Disambiguation: Algorithms and Applications, Agirre and Edmonds (eds), Springer-Verlag, pp. 47-73, 2006

Moon Rosamund, *Fixed Expressions and Idioms in English: A Corpus-Based Approach*, Clarendon Press, 1998.

Procter Paul, *Longman Dictionary of Contemporary English*, Longman, 1978.

Hinrich Schutze and Jan Pedersen, "Information Retrieval Based on Word Senses", in Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval, pp. 161-175, 1995