SESSION 6: LEXICON AND LEXICAL SEMANTICS

Paul S. Jacobs, Chair

Artificial Intelligence Laboratory
GE Research and Development Center
Schenectady, NY 12301

MOTIVATION AND BACKGROUND

While other word-level marking tasks such as morphology and part-of-speech tagging have arrived recently at a well-developed methodology and a basis for comparing results across systems, the robust discrimination of word senses in text is a less mature discipline. Yet, word sense discrimination is central to many natural language processing tasks, such as data extraction and machine translation.

The three papers in this session all describe work at distinguishing word senses in broad classes of naturally-occuring text, albeit using different approaches and focusing on different aspects of the problem. They all use statistical methods to a degree, and attempt to produce quantitative measures of accuracy for comparison. They differ substantially in the degree to which knowledge-based methods are considered as well as in the applications for which the work is aimed.

NOTEWORTHY PROGRESS

The paper, "One Sense per Discourse" by Gale, Church, and Yarowsky, reports that words used repeatedly in the same section of text tend to be used in the same sense each time. The research reported uses the large, bilingual Hansard corpus, aligning each word in English, for example, with its French translation as a source of information about the sense of the English word. The English word *sentence* for example, will align with the French word *peine* if it refers to a penalty, and with the word *phrase* if it describes a piece of text.

Aside from the main result, reflected in the title, and a thorough analysis of a huge volume of textual data, the Gale et. al. paper seems to provide some hope that testing on a relatively straightforward task with a readily available source of data might carry over into other tasks. In other words, if one can train a system to distinguish word senses based on context using the Hansard corpus, perhaps this training will help to distinguish word senses for other translation tasks or even for data extraction or information retrieval. This hypothesis remains to be tested, but any carry-over would mean that these large quantities of training material would be useful without any special hand annotation.

"Lexical Disambiguation using Simulated Annealing" by Cowie, Guthrie, and Guthrie, uses input text and data from the Longman Dictionary of Contemporary English (LDOCE), guessing that word senses can be distinguished using the possible subject fields of the words in the surrounding context. For example, the word *interest* when surrounded by words that can have a financial subject field is much more likely to have a financial sense. Annealing comes into play as an algorithm for maximizing entropy in the combination of interpretations, in other words, finding the set of word senses with the greatest degree of overlap in their possible subject field encodings.

One of the interesting aspects of the Cowie et. al. work is that it raises the possibility that dictionary definitions themselves could be made more useful by some kind of automatic sense disambiguation. Also, training on the lexicographer's choice of examples in illustrating different senses might reduce the level of noise in these examples, perhaps meaning that less data is required to produce good contextual discriminators. Unlike the Gale work, this research assumes that sense discrimination takes place with respect to a lexicon rather than with respect to a corpus, perhaps a more realistic assumption in the practice of current NLP.

"The Acquisition of Lexical Semantic Knowledge from Large Corpora", by James Pustejovsky, places corpus analysis in a subservient role to lexical representation, using the statistical analysis of a corpus as a way of determining, for example, the degree to which a word sense can be extended metonymically. This sort of corpus-based evidence can be compelling. A word like *announced*, for example, normally demands an animate subject, but in a particular corpus it might occur most frequently with an organization as the subject, offering evidence that these occurrences are either extensions of the word sense or examples of metonymy, i.e. that the name of the organization represents an individual or group of individuals.

Unlike the other two papers, the Pustejovsky work embraces the role of statistics as part of knowledge-based processing, not as a replacement for the development of lexicons and lexical theories. Furthermore, using the TIPSTER data extraction task as an application, this research is representative of work trying to use corpus analysis as an aid to the knowledge acquisition problems that burden current text interpretation systems.

CURRENT PROBLEMS AND ISSUES

There are some important points to consider in comparing and weighing these preliminary results. Statistical methods have a special appeal: these systems robustly process large volumes of text, and produce interesting, quantitative results. Yet, are these results meaningful? Are they comparable? How can one extrapolate from the results to the effects of automated knowledge acquisition on NLP tasks?

Statistics is not a replacement for knowledge-based processing or knowledge representation theories, rather, it is one of many tools that can help to produce a robust, functional NLP system. This observation isn't obvious from reading the Gale paper, but it does help to fit the papers together.

Another non-obvious question is how to view the different examples that have been selected for analysis, as well as the results that are produced using the sample examples in different corpora. For example, the task of discriminating the senses of *slug* (a worm-like creature or a piece of metal) seems fundamentally different from a word like *concern* (a business or something to think about). Not only is *concern* harder, but we can see how this would make a big difference in a task like TIPSTER, where a word like *concern* might refer to a key player in a transaction only if it takes the business sense. Thus, the successful discrimination of the hard, relevant words seems ultimately to be the test of these methods.

Even where the same words are used for testing (like *interest*), the numbers appear to reflect drastically different results. In Gale's work, *interest* almost always seems to come out right (96%), while it is 70% in Cowie's (and others') reports. This might mean that *interest* is less ambiguous in Hansard than in the other corpora, or it might be lucky that the different senses happen to translate into *interêt* most of the time, anyway. Thus, while it's true that 70% is better on a given task than 50%, there's no way now to compare one set of numbers to another or to know whether even 96% is any good. This doesn't mean we shouldn't report numbers, but means we have to find a meaningful way to compare.

Finally, for all the reporting that's been done on statistical analysis of corpora, there still hasn't been much use of automated training in text interpretation. It seems that it is only a matter of time before statistical training becomes part of all knowledge-based NLP, but until this happens, we don't have a measure of the degree to which training actually helps, for example, in data extraction or machine translation. We expect that this will be a topic for the lexical semantics session at one of the future workshops.