Extracting Constraints on Word Usage from Large Text Corpora

Kathleen McKeown and Rebecca Passonneau

Department of Computer Science
450 Computer Science Building
Columbia University

PROJECT GOALS

Our research focuses on the identification of word usage constraints from large text corpora. Such constraints are useful both for the problem of selecting vocabulary for language generation and for disambiguating lexical meaning in interpretation. We are developing systems that can automatically extract such constraints from corpora and empirical methods for analyzing text. Identified constraints will be represented in a lexicon that will be tested computationally as part of a natural language system. We are also identifying lexical constraints for machine translation using the aligned Hansard corpus as training data and are identifying many-to-many word alignments.

One primary class of constraints we are examining is lexical; that is, constraints on word usage arriving from collocations (word pairs or phrases that commonly appear together). We are also looking at constraints deriving from domain scales which influence use of scalar adjectives and determiners, constraints on temporal markers and tense, constraints on reference over text, and constraints on cue words and phrases that may be used to convey explicit information about discourse structure.

RECENT RESULTS

- Packaged Xtract, a collocation extraction system, with windows interface for use by other sites and have licensed to several sites.
- Implemented a prototype system to compile candidate translations for English collocations by identifying collocations in the source language using Xtract, and incrementally building the target collocation from highly correlated words in the target corpus. The system has been evaluated on a small number of collocations, yielding 80% accuracy.
- Implemented a system for retrieving semantically related adjectives from a parsed text corpus, using a similarity metric and clustering techniques. Evaluation and comparison with human judges shows that system performance is comparable to human performance.
- Experimented with a genetic programming algorithm to identify statistically significant links between cue words and other words or part of speech in a large text corpus. Early results are promising, predicting, for example, that sentence initial cue words are used in their discourse sense.
- Implemented semantic and syntactic constraints on historical information in statistical reports as revision rules in a report generation system.
- Developed 3 simple algorithms for identifying segment boundaries using features of the text, and evaluated their success at identifying the segment boundaries that humans identify. The algorithms each use different linguistic information: speech pauses, cue words, and referring expressions.
- Developed method for extracting tense sequences across adjacent sentences from corpora and evaluated behavior of semantically under-constrained past and past perfect tenses in the Brown corpus. Developed a semantic representation for past and for perfect, and an algorithm for understanding tense in discourse.

PLANS FOR THE COMING YEAR

In the area of machine translation, we are improving the implementation to prepare for large scale experimentation, including indexing the corpus to speed up testing and automating subcomponents. We will begin large scale testing within the month, beginning with 1 month's worth of data (about 100 collocations) and moving to 1 year's worth of data or 1000 collocations. We will continue to improve the accuracy of our method for retrieving scalar adjectives; we are investigating the use of other sources of linguistic data such as conjunction and negation. We will add tests that exploit gradability in order to identify scalar and non-scalar groups. We will refine methods for extracting tense sequences from corpora that fit certain criteria by adding new tenses, or adverbs, or aspectual type to the criteria and will identify additional constraints on tense understanding. We will further refine our algorithms for identifying discourse structure. We are developing a generation system to test constraints on historical information and a bidirectional semantic engine to test constraints on tense, aspect, and discourse cues.