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

- Licensed Xtract, a collocation extraction system with windows interface, to 26 sites, including 1 commercial site.
- Completed implementation of a collocation translation system called Champollion, which uses an incremental, statistical algorithm and ported Champollion from a DOS to Unix platform.
- Evaluated Champollion using 3 human judges on 2 year's worth of Hansard's data, yielding an average of 77% accuracy on one year and 61% accuracy on the other with problems on the second set due to reference database corpus size.
- Developed a system that automatically extracts relevant data from text corpora and corpora-based databases for linguistic tests proposed for determining which is the marked element of a pair of antonymous adjectives.
- Assessed the statistical significance of the markedness results, showing that some simple linguistic tests are good indicators of markedness, while others fail to perform well. We combined the tests using a smoothed log-linear regression model yielding a small, but statistically significant improvement.
- Using the results of a corpus analysis of automatically collected pairs of conjoined adjectives, designed a graph model to separate adjectives into sub-scales (positive and negative) according to the sub-graphs formed.
- Implemented and evaluated a genetic programming algorithm to induce decision trees for cue word disambiguation based on lexical and part of speech data in a large text corpus.
- Analyzed distribution of NPs in Pear corpus of oral narratives, leading to new, more comprehensive formulation of the factors correlated with definiteness.
- Analyzed co-occurrence data of 16 aspectual verbs in Brown corpus (e.g., 'begin', 'continue') to rank them by their ability to categorize aspectual class of their verb arguments. Six aspectual verbs account for 90% of the different verbs appearing as arguments.

PLANS FOR THE COMING YEAR

For collocation translation, we are continuing evaluation of Champollion, using several years worth of the Hansard as a reference corpus and testing on separate data. We will also investigate methods for including prepositions in the translation algorithm. For scalar adjectives, we are implementing the graph model so that we can automatically partially order elements of groups of semantically related scalar adjectives that are produced by a program we implemented in earlier years. For analysis of aspectual verbs, we plan to compare aspectual data derived from Brown with similar data from a new corpus, or to integrate a supplementary classification method (e.g., use of independent semantic net) with the current distributional model.