Extracting Constraints on Word Usage from Large Text Corpora

Kathleen McKeown, Diane Litman, 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 important for natural language systems, both for the problem of selecting vocabulary for language generation and for disambiguating lexical meaning in interpretation. The first stage of our research involves the development of 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 will examine is lexical; that is, constraints on word usage arriving from collocations (word pairs or phrases that commonly appear together). We will also look 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. We also plan to examine corpora of prosodically labeled transcribed speech in order to identify intonational constraints on word usage.

RECENT RESULTS

- Added syntactic parser to Xtract, a collocation extraction system, to further filter collocations produced, eliminating those that are not consistently used in the same syntactic relation. This increased precision from 40% to 80%. Recall of this stage was evaluated at 94%.
- Developed and implemented a method for retrieving the elements of adjectival scales, using mutual information between adjective-noun collocations and clustering techniques to group them.
- Designed a system to compile a list of candidate translations between English and French words using an evaluation of mutual information between words in the aligned Hansard corpus.
- Performed empirical analysis of advising transcripts identifying a class of adjectives used *evaluatively* (as opposed to adjectives conveying objective information) and constraints on their use. Developed and implemented new control tools in FUF to use lexical constraints in text generation.
- Identified semantic and syntactic constraints on historical information in statistical reports through partially automated analysis using Xtract tools.
- Completed an empirical study of discourse segmentation, assessing the ability of naive subjects to assign segment boundaries based on the notion of intention used in Grosz/Sidner's definition of discourse segment.
- Selected and obtained several corpora for analysis: the AP news wire (for finding scalar and other relations); the Brown corpus (for anaylzing temporal adverbs and morphological units); and the Pear stories (for investigating the role of tense, cue phrases, and intonation).

PLANS FOR THE COMING YEAR

In the area of machine translation, we are extending our system to identify collocations using Xtract in both the French and English corpora and then produce a translation score based on the mutual information of the individual words they contain. We will complete implementation of this technique and evaluate it through large scale experimentation. We will improve the accuracy of our method for retrieving scalar adjectives by experimenting with other clustering techniques and will begin looking at methods for ordering the scales retrieved. We are also applying the technique used for adjectives to identify topically related groups of nouns to aid identification of discourse segments. In addition, using the segment boundaries found empirically as a baseline, we will develop automatic methods for identifying such boundaries, based on the analysis of usage constraints on referring expressions, tense, and lexical cohesion. We will evaluate precision and recall of automatic segmentation methods. We are extending Xtract to find collocations between tenses and temporal adverbials. Finally, we are also testing constraints on evaluative adjectives and historical information in a generation system.