RESEARCH IN NATURAL LANGUAGE PROCESSING

A. Joshi, M. Marcus, M. Steedman, B. Webber

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

PROJECT GOALS

The main objective is to develop robust methods for the understanding and generation of both written and spoken human language, including but not limited to English. Penn is pursuing development of: (1) New mathematical and computational frameworks which are highly constrained, yet adequate to allow a simple, concise description of complex linguistic phenomena. These new frameworks are tested by the explicit encoding within each framework of a wide range of phenomena across a diverse set of human languages. (2) Both statistical and symbolic learning methods which automatically extract and effectively utilize the implicit linguistic knowledge in the Penn Treebank and the corpora of the Linguistic Data Consortium. These techniques have been tested against the performance of the best current methods.

RECENT RESULTS

- In a lexicalized grammar such as the lexicalized tree-adjoining grammar (LTAG), each lexical item is associated with one or more elementary trees (structures), called *supertags*. We have developed techniques to eliminate or substantially reduce the supertag assignment ambiguity by using local lexical dependencies and their distribution, prior to parsing. After this step only explicit indication of substitutions and adjoinings must be indicated to complete parsing. Preliminary experiments on short fragments show a success rate of 88%, with experiments continuing on full sentences from WSJ material.

- The Information-Based Intonation Synthesis (IBIS) spoken reply system has been extended by a richer semantics for the assignment of stress on the basis of contrast in the domain of discourse. Synthesis of spoken responses as speech waves bearing an intonation contour appropriate to the context of utterance has thereby been considerably improved.

- A weakly supervised symbolic learning algorithm called Error Based Transformation Learning has been developed that matches or beats the performance of the best standard methods for a range of key language analysis tasks. This method has also been used for part of speech tagging for several languages other than English with very good results.

- A new algorithm for word-sense determination performs as least as well as existing algorithms, while only using only a window of five words around the target word, as opposed to 100 words for these existing methods.

PLANS

- Apply part-of-speech disambiguation strategies to the disambiguation of lexical category assignmments words in a combinatory categorial parser.

- Port the IBIS spoken response generator to the larger domain involved in the task of critiquing of Medical Diagnosis by an expert system.

- Explore statistical morphology induction, lexical disambiguation, and language modeling with stochastic dependency grammars.

- Test the XTAG system on a corpus and build a TAG parsed corpus to serve as the basis for statistical experiments with the TAG grammar and parser.

- Contribute to a model of limited processing for discourse, using LDC corpora as the basis for an empirical analysis of bottom-up cues to discourse structure, such as variation in the forms of referring expressions, and prosodic marking by topline and baseline variation.

- Develop part-of-speech taggers and morphological learners for a range of languages other than English.

- Develop the 'strategic' or discourse-planning component of the spoken reply system.