Natural Language Research

PIs: Aravind Joshi, Mitch Marcus, Mark Steedman, and Bonnie Webber

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
email: joshi@cis.upenn.edu

OBJECTIVE

The main objective is basic research and system development leading to (1) characterization of information carried by (a) syntax, semantics, and discourse structure, (b) their relation to information carried by intonation, and (c) development of methods for using this information for generation and understanding; (2) development of architectures for integration of utterance planning with lexical, syntactic and intonational choice; (3) development of incremental strategies for using syntactic, semantic, and pragmatic knowledge in understanding and generating language.

RECENT ACCOMPLISHMENTS

- Developed a new weakly supervised learning algorithm that can bracket text using a simple distributional error-correcting technique, which performs as well as recent applications of the I-O algorithm while using an order of magnitude less training data. A similar technique has been applied successfully to the problem of prepositional phrase attachment.
- Developed techniques that combine WordNet and corpus-based lexical statistics acquired from the Penn Treebank. These techniques are being applied to the resolution of syntactic ambiguity.
- The Penn Treebank project has released the results of its first three year phase as a CDROM through the Linguistic Data Consortium, consisting of 4.5 million words of part-of-speech tagged text and 3 million words of skeletally parsed text, including a parsed version of the Brown corpus.
- A categorial grammar based theory of intonation structure and its discourse meaning has been developed and implemented in a database query system which takes as input an orthographic representation of spoken questions including intonational annotations, and yields as output a synthesized spoken response as a speech wave bearing an intonation contour that is appropriate to the context established by the question.
- Developed an environment, Design World, for simulating interactive task-oriented dialogue between two agents, that allows us to explore a number of key issues in inter-agent coordination.
- Investigated the way in which dialogue processing is cued by patterns of spoken language in task-oriented interactions between multiple agents. Results show that the redundancy that makes communication more robust is typically marked by prosodic destressing or broad focus.

PLANS FOR THE COMING YEAR

- Explore statistical morphology induction, lexical disambiguation, and language modeling with stochastic dependency grammars.
- Extend the use of WordNet and lexical statistics to the resolution of a broader set of syntactic ambiguities, and to apply these techniques to the construction of stochastic language models.
- Contribute to a model of limited processing for discourse, using corpora collected by the Linguistic Data Consortium as the basis for a corpus-based 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.
- Extend the part-of-speech disambiguation strategies to the disambiguation of lexical tree assignments to words in a lexicalized tree-adjoining grammar.
- Develop a minimal-response part-of-speech tagger for conversational German without the use of online dictionaries and with minimal human resources.
- Investigate the acquisition of lexical information about novel verbs by combining information about syntactic contexts with information about semantic relationships acquired using WordNet.
- Develop the 'strategic' or discourse-planning component of the spoken reply system.