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

- An algorithm was designed based on Earley's parser for estimating the parameters of a stochastic context-free grammar. Contrary to other approaches, this algorithm does not require that the grammar is in a normal form.
- A new predictive left-to-right parser for TAG was designed and included in a software package (XTAG).
- An X-based Graphical Interface for Tree-Adjoining Grammars (XTAG) has been released for distribution. This software package includes: (1) a graphical editor for trees; (2) a parser for unification-based tree-adjoining grammars; (3) utilities for defining grammars and lexicon for tree-adjoining grammars; and (4) a user manual.
- The notion of stochastic tree-adjoining grammars was defined and an algorithm for estimating from a corpus the probabilities of a stochastic TAG was designed. Lexicalized tree adjoining grammar (LTAG) provides a stochastic model that is both hierarchical and sensitive to lexical information.
- Developed a new notion of derivation for the tree adjoining grammars, which is sensitive to the distinction between modifier and predicational auxiliary trees. This distinction is relevant to the design of probabilistic LTAGs.
- Developed a new formalism, structure unification grammar, that allows many of the key insights of a variety of grammatical formalisms to be brought to together in one framework, although at a cost of some increased computational complexity.
- The Pereira-Pollack approach to *incremental interpretation* was extended to support a discourse-based algorithm for resolving verb phrase ellipsis.

PLANS FOR THE COMING YEAR

- Continue work on automatic extraction of linguistic structure, extending work on determination of part-of-speech tag sets and adding morphophonemic rules to the morphology algorithm, focusing on automatically discovering high-level grammatical structure.
- Extend the techniques used for the design of polynomial time and space shift-reduce parsers for arbitrary context-free grammars to tree adjoining grammars.
- Complete the work on stochastic tree-adjoining grammars, implement an algorithm for estimating from a corpus the probabilities of a stochastic TAG, and investigate the design of algorithms for using parsed corpora such as the Penn Treebank as the basis for the estimation of stochastic tree-adjoining grammars.
- Complete the work on the new derivation for LTAGs based on the distinction between modifier and predicational auxiliary trees and integrate this formulation in the framework of stochastic TAGs.
- Complete the integration of coordination in the tree adjoining grammar framework.
- Begin work on the problem of word-order variation, which is more common in languages such as German, Korean, Japanese, among others.