Very Large Annotated Database of American English

PI: Mitch Marcus

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389
email:mitch@cis.upenn.edu

Objective

To construct a data base (the "Penn Treebank") of written and transcribed spoken American English annotated with detailed grammatical structure. This data base will serve as a national resource, providing training material for a wide variety of approaches to automatic language acquisition, a reference standard for the rigorous evaluation of some components of natural language understanding systems, and a research tool for the investigation of the grammar and prosodic structure of naturally spoken English.

Summary of Accomplishments

Treebank

- We have developed the Penn Treebank Annotator's Manual which gives about 40 pages of specification for the use of our part-of-speech tag set.
- We have manually corrected over 4 million words by part of speech, with an error rate of about 2.5%. This output is produced at a consistent rate of about 3,000-3,500 words per hour.
- To provide highly accurate bigram frequency estimates, two subcorpora were tagged twice and then adjudicated, yielding ~160,000 words tagged with an accuracy estimated to exceed 99.5%. This procedure has an output of about 1,000 words per hour overall.
- We have developed a mouse-based annotator's workstation for syntactic bracketting, and a preliminary tag set and annotator's guide for syntactic bracketing. Annotators have begun learning to edit bracketed structures, posteditting the output of Don Hindle's Fidditch parser, which we have acquired from AT&T Bell Labs.

Learning

- We have developed a new information theoretic parsing algorithm which derives an (unlabelled) bracketing of free text without the use of an explicit grammar. Our current implementation determines recursively nested sentence structure, with an error rate of roughly 2 misplaced boundaries for test sentences of 10-15 words, and roughly five for sentences of 15-30 words.
- To see whether a purely distributional analysis can discover part of speech information, we have developed a similarity measure which accurately clusters closed-class lexical items of the same grammatical category, failing only on those words which are ambiguous between multiple parts of speech.

Plans

- We expect to produce our first skeletal parse trees within the next few weeks.
- We intend to retrain Church's tagger, with error rate expected to be less than 3%. By adjudicating between the output of this new tagger and our current output, we expect well below 1% error, at an additional cost of between 5 and 10 minutes per 1000 words.
- We intend to develop a strategy for collecting naturally occurring task-oriented dialogues in a variety of settings.
- We intend to aggressively extend our results in learning in a variety of settings. We also intend to investigate symbolic learning procedures based on a study of the process of creolization of pidgin languages; this work is already underway.