Very Large Annotated Database of American English

Mitch Marcus, Principal Investigator

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

PROJECT GOALS

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 of naturally spoken English.

RECENT RESULTS

Treebank

We have provided up to 4.5 million words of preliminary Treebank material tagged for part of speech (POS) to 10 different sites in the U.S., including DARPA contractors, industrial research labs and universities. We have finished POS tagging of 351 MUC-3 messages (about 100K words) and grammatical annotation of a subset of 88, which are available by anonymous FTP.

After early concerns about productivity, we investigated a range of methods for syntactic annotation (henceforth, tree banking) with respect to annotator speed, for annotators posteditting the output of Hindle's Fidditch parser. Key results:

1. Annotators take substantially longer to learn tree banking than the POS annotation task, with substantial increases in speed occuring after 2 months of training.
2. Annotators can postedit the full output of Hindle's parser at an average speed of 100-200 words per hour after three weeks, and 400-500 words per hour after two months.
3. Reducing the output to a far more skeletal representation increases average speed to 700-750 words per hour. At this speed, we can maintain an output of 2.5 million words a year of "treebanked" sentences, with each sentence posteditted once.

Since December 1, we have annotated about 250K words of text, with 1/3 bracketted by more than one annotator.

Learning

We have developed, in collaboration with UNISYS, a parsing algorithm for unrestricted text which uses a probability-based scoring function to select the "best" parse. The parser, Pearl, is a time-asynchronous bottom-up chart parser with Earley-type top-down prediction which pursues the highest-scoring theory in the chart, where the score of a theory represents the extent to which the context of the sentence predicts that interpretation. This parser differs from previous stochastic parsers in that it uses a richer form of context to predict likelihood. Trained on a corpus of 1100 sentences from MIT's Voyager direction-finding system and using the string grammar from UNISYS' PUNDIT Language Understanding System, Pearl correctly parsed 35 out of 40 or 88% of test sentences from previously unseen Voyager sentences.

We have also adapted Church's PARTS tagger, from AT&T Bell Labs, to operate independently of particular tag sets, and have developed software for retraining the tagger which is available to those who have obtained a license to use the version of PARTS which AT&T distributes.

PLANS FOR THE COMING YEAR

- Automatic methods for consistency checking between annotators need to be developed.
- We expect to use our retrained version of PARTS to retag the preliminary corpus and then adjudicate between it and the output of the retrained tagger. We expect that the error rate will drop to well under 1% on correct tags, at an additional cost of 5 minutes per 1000 words.
- We intend to begin work on automatic grammar extraction from POS tagged corpora, combining earlier work in bracketting using information theoretic measures with notions of phrase structure (so-called "X-bar theory") from competence linguistics.
- We expect to tree bank lots and lots of sentences.