ROBUST TEXT PROCESSING AND INFORMATION RETRIEVAL

Tomek Strzalkowski, Principal Investigator

Department of Computer Science
New York University
New York, New York, 10003

PROJECT GOALS

The general objective of this research has been the enhancement of traditional key-word based statistical methods of document retrieval with advanced natural language processing techniques. In the work to date the focus has been on obtaining a better representation of document contents by extracting representative phrases from syntactically preprocessed text. In addition, statistical clustering methods have been developed that generate domain-specific term correlations which can be used to obtain better search queries via expansion.

RECENT RESULTS

A prototype text retrieval system has been developed in which a robust natural language processing module is integrated with a traditional statistical engine (NIST's PRISE). Natural language processing is used to (1) preprocess the documents in order to extract contents-carrying terms, (2) discover inter-term dependencies and build a conceptual hierarchy specific to the database domain, and (3) process user's natural language requests into effective search queries. The statistical engine builds inverted index files from pre-processed documents, and then searches and ranks the documents in response to user queries. The feasibility of this approach has been demonstrated in various experiments with 'standard' IR collections such as CACM-3204 and Cranfield, as well as in the large-scale evaluation with TIPSTER database.

The centerpiece of the natural language processing module is the TTP parser, a fast and robust syntactic analyzer which produces 'regularized' parse structures out of running text. The parser, presently the fastest of this type, is designed to produce full analyses, but is capable of generating approximate 'best-fit' structures if under a time pressure or when faced with unexpected input.

We participated in the first Text Retrieval Conference (TREC-1), during which the total of 500 MBytes of Wall Street Journal articles have been parsed. An enhanced version of TTP parser has been developed for this purpose with the average speed ranging from 0.3 to 0.5 seconds per sentence. We also developed and improved the morphological word stemmer, syntactic dependencies extractor, and tested several clustering formulas. A close co-operation with BBN has produced a better part-of-speech tagger which is an essential pre-processor before parsing.

We also took part in the continuing parser/grammar evaluation workshop. In an informal test runs with 100 sentence sample of WSJ material, TTP has come suprisingly strong among 'regular' parsers which are hundreds times slower and far less robust. During the latest meeting the focus of evaluation effort has shifted toward 'deeper' representations, including operator-argument structures which is the standard form of output from TTP. During last year TTP licenses have been issued to several sites for research purposes.

In another effort, in co-operation with the Canadian Institute of Robotics and Intelligent Systems (IRIS), a number of qualitative methods for predicting semantic correctness of word associations are being tested. When finished, these results will be used to further improve the accuracy of document representation with compound terms.

Research on reversible grammars continued last year with some more important results including a formal evaluation system for generation algorithms, and a generalized notion of guides for controling the order of evaluation.

PLANS FOR THE COMING YEAR

The major effort in the coming months is the participation in TREC-2 evaluation. For this purpose we aquired a new version of PRISE system, which is currently being adapted to work with language processing module. New methods of document ranking are also considered, including local scores for most relevant fragments within a document. New clustering methods are tested for generating term similarities, as well as more effective filters to subcategorize similarities into semantic classes.