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 and devising suitable weighting schemes for different types of terms. 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 second Text Retrieval Conference (TREC-2), during which the total of 850 MBytes of Wall Street Journal and San Jose Mercury News articles have been parsed. An enhanced version of TTP parser has been developed for this purpose with the average speed generally below 0.2 seconds per sentence (approx. 85 words per second). We also developed and improved the morphological word stemmer, syntactic dependencies extractor, and tested several clustering formulas. In a close co-operation with BBN, we used their POST part-of-speech tagger which is an essential pre-processor before parsing.

During last year TTP licenses have been issued to several sites for research purposes and one commercial license is pending.

The main effort during TREC-2 was on generating appropriate term-based representation of documents for improved search performance. Documents were indexed with a mixture of statistical terms (words and stems) and linguistically derived phrases. Subsequently a new method for weighting mixed sets of terms has been developed which produced a significant increase in retrieval precision.

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.

PLANS FOR THE COMING YEAR

A major effort in the coming months is the participation in TREC-3 evaluation. Work will continue on improved term weighting methods, term derivation and disambiguation techniques, and automatic feedback for routing. More clustering methods will be tested for generating term similarities, as well as more effective filters to subcategorize similarities into semantic classes. We plan to move from category B participation (exploratory systems) to category A (full data set).

This site has also been selected as a TIPSTER Phase 2 contractor. As a part of this contract an integrated system for document retrieval and information extraction will be built.