Diderot: TIPSTER Program, Automatic Data Extraction from Text Utilizing Semantic Analysis

Y. Wilks, J. Pustejovsky†, J. Cowie

Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003
&
Computer Science†, Brandeis University, Waltham, MA 02254

PROJECT GOALS

The Computing Research Laboratory at New Mexico State University and the Computer Science Department at Brandeis University, are developing an original method for the large scale extraction of information from Japanese and English texts. This method is general and extendible; its techniques are not explicitly tied to these two particular languages, nor to the finance and electronics domains which are the initial targets of the Tipster project.

We are designing and testing a set of procedures for the automatic extraction of information from Japanese and English texts, and for placing this information in prespecified templates. Our approach is based on "partial parsing" to analyze selected text parts, in combination with statistical techniques and the incorporation of large-scale data-bases.

Our method makes use of minimalist AI techniques based on formal lexical structures that are largely automatically derived and tuned against corpora. In addition, we are constructing a range of small-scale demons for such items as dates, company names and place names. Statistical techniques are being used to identify relevant vocabulary and to identify significant sections of text. All our proposed techniques have been tested at one of our two research centers; what we are proposing overall is a state of the art system based on a novel, yet theoretically defensible combination of those techniques.

RECENT RESULTS

The project is in its early stages and much preliminary investigative and infrastructure work has been undertaken. A template filling tool (Hume) was developed at CRL, as much to learn about the structure of the templates as to provide a usable environment for human analysts on SUN workstations. Brandeis undertook the analysis of time as related to the Tipster task and have produced a set of definitions for the time words in the Tipster vocabulary. A survey of the use of metaphor in the texts has been made. Brandeis are co-ordinating the distribution of a Tipster news letter (tipster@cs.brandeis.edu). CRL has produced a list of Japanese resources for the Tipster Japanese groups.

An outline design for the system has been produced and work on the various components was started. Lexical definitions for the words of a single Tipster text were automatically derived from the Longman Dictionary of Contemporary English. These were used as source material for defining Generative Lexical Structures for the words. The task of manually creating GLS structures for a large number of words has proved to be difficult and we have decided to use a subset of these structures which is sufficient for the Tipster task.

A statistically based method for detection of texts has been developed which is accurate in distinguishing two text types (eg Tipster and MUC). Given an appropriate set of words from each text the exact probability of detection can be calculated for text samples of any size. We intend to extend this method to the problem of detecting relevant paragraphs.

PLANS FOR THE COMING YEAR

Our intention is to produce a series of systems of increasing degrees of sophistication. These will be tested in the appropriate Tipster and MUC-4 evaluations. We intend to use several data-intensive pre-processing steps to select relevant paragraphs, using our statistical method, and to mark large lexical units; such as proper names and industrial products. Paragraph selection will be based on data derived from the filled templates. The lexical units will be identified using a variety of U.S. government data-bases. The tagged text will serve as input to the partial parsing stage of the system. At present we are experimenting with several parsing systems and associated grammars. An initial trade-off is being made in favor of robustness over depth, but later versions of the system will incorporate more accurate syntactic and semantic parsing. Inter-lingual based domain models will be constructed and used as templates in the derivation of relevant information from the parsed text. A final post-processing stage (Bruce) will incorporate the specific rules related to the individual fields of a template.