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, in collaboration with Brandeis University, was one of four sites selected to develop systems to extract relevant information automatically from English and Japanese texts. When we started, neither site had been involved in message understanding or information extraction. CRL had extensive experience in multilingual natural language processing and in the use of machine readable dictionaries for system building, Brandeis had developed a theory of lexical semantics and preliminary methods for deriving this lexical information from corpora. Thus, our approach focused on applying new techniques to the information extraction task. In the last two years we have developed information extraction software for 5 five different subject area/language pairs.

The system, *Diderot*, was to be extendible and the techniques used not explicitly tied to the two particular languages, nor to the finance and electronics domains which are the initial targets of the Tipster project. To achieve this objective the project had as a primary goal the exploration of the usefulness of machine readable dictionaries and corpora as source for the semi-automatic creation of data extraction systems.

RECENT RESULTS

The first version of the system was developed in five months and was evaluated in the 4th Message Understanding Conference (MUC-4) where it extracted information from 200 texts on South American terrorism. At this point the system depended very heavily on statistical recognition of relevant sections of text and on the ability to recognize semantically significant phrases (e.g. a car bomb) and proper names. Much of this information was derived from the keys.

The next version of the system used a semantically based parser to structure the information found in relevant sentences in the text. The parsing program was derived automatically from semantic patterns. For English these were derived from the Longman Dictionary of Contemporary English, augmented by corpus information and these were then hand translated to equivalent Japanese patterns. The Japanese patterns were confirmed using a phrasal concordance tool. A simple reference resolving module was also written. The system contained large lists of company names and human names derived from a variety of online sources. This system handled a subset of the joint venture template definition and was evaluated at twelve months into the project.

Attention was then focused on the micro-electronics domain. Much of the semantic information here was derived from the extraction rules for the domain. A single phrase in micro-electronics can contribute to several different parts of the template, to allow for this a new semantic unit the *factoid* was produced by the parser. This produced multiple copies of a piece of text, each marked with a key showing how the copy should be routed and processed in subsequent stages of processing. This routing was performed by a new processing module, which transformed the output from the parser. The statistical based recognition of text relevance was used for micro-electronics only, as a much higher percentage of articles in the corpus were irrelevant. This system was evaluated at 18 months.

Finally the improvements from micro-electronics were fed back to the joint venture system. An improved semantic unit recognizer was added to the parser. This handles conjunctions of names, possessives and bracketing. An information retrieval style interface to the Standard Industrial Classification Manual was linked into the English system. The reference resolving mechanism was extended to handle a richer set of phenomenon (e.g. plural references). This version was evaluated at 24 months.

PLANS FOR THE COMING YEAR

CRL is participating in Tipster Phase 2. This will involve participation in the development of the architecture for the Phase 2 system, user interfaces to the system, software to handle document markup and multilingual information retrieval. Brandeis are continuing work on tuning lexical entries using information from corpora.