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 have been working for the past 18 months on the development of a system to perform automatic data extraction from texts on restricted subject areas (for business - joint ventures and for micro-electronics - improvements in semiconductor technology) in two languages (English and Japanese).

The eventual aim is to be able to automatically populate data-bases containing information relevant to the work of government analysts.

The system, *Diderot*, is 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 has 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

In the past year we have implemented 5 different text extraction systems three for English and two for Japanese. Experiments have been carried out on the derivation of syntactic and semantic information automatically from machine readable dictionaries and text corpora. Statistical methods have been developed for recognizing relevant texts or sections of texts. Methods have been developed which tag text using finite state automata which mark organizations, human names, places, products and dates. A parser generator has been produced which converts Generative Lexical Semantic structures into Definite Clause Grammar rules.

Japanese and English systems for a subset of the joint venture domain were evaluated on their performance against unseen texts for the Tipster 12 month evaluation. Full coverage Japanese and English systems for micro electronics were tested for the 18 month evaluation. Full coverage systems for the business domain were not completed in time for this evaluation and enhanced versions of the 12 month systems were used. The detailed results of the evaluations are reported in the Tipster 12 and 18 month session notebooks. The performance of the English systems is still very poor. The performance of the Japanese systems is significantly better, with the most recently developed micro-electronics system performing with a precision of 48% and a recall of 26%. It would appear that a combination of factors has influenced this result including more specific hand tuning and a single system developer.

Tools have been created to support human text extraction for Japanese and English for both domains. These have been used by IDA to produce all the training data (key structures extracted from over 4,000 texts) for the Tipster extraction project. Work has commenced on a tool (Tabula Rasa) which allows the creation of template extraction tools for human analysts and which also supports the definition of an extraction task.

PLANS FOR THE COMING YEAR

The principal objective for the next six months is to produce systems for English and Japanese which perform uniformly well (around 50% precision and 40% recall). Initially we plan to hand tune our English systems to provide a 'core' system of good quality. We then intend to extend the coverage of this system by automatically extending the lexicon using machine readable dictionaries and information extracted automatically from corpora. At present most parts of the Diderot system are easily configurable for new languages and/or domains. The modules which perform reference resolution were specifically written for each domain (topic) and language. We intend to modify these and separate as far as possible domain information from the general purpose reference resolution mechanism.

Subsequently, we hope to be involved in the next phase of Tipster, both in the provision of modules for information extraction and in the provision of tools to assist the design process and to support the integration of automatic and human information extraction.