CERVANTES - A SYSTEM SUPPORTING TEXT ANALYSIS

Jim Cowie
Computing Research Laboratory
New Mexico State University, Box 30001/3CRL, Las Cruces, NM, 88003

COTR: Lynne Rich, DoD, larich@afterlife.ncsc.mil
Principal Investigators: Jim Cowie, jcowie@nmsu.edu (505) 646-5181, William Ogden, ogden@nmsu.edu (505) 646 6222, Ted Dunning (Now at HNC)

Summary

CRL is engaged in the development of document management software and user interfaces to support government analysts in their information analysis tasks. It is also continuing to develop language technologies to support document detection and information extraction in a variety of languages. It has also been responsible for the integration and delivery of both the six and twelve month Tipster demonstration systems and the development of the first Tipster document manager.

Approach

CRL has provided general purpose enabling technology for several aspects of the Tipster Phase II program. This work has been carried out in close compliance with the decisions of the Architecture Working Group. The software developed at CRL is based on CRL's extensive experience of user interface support for government analysts developed during Phase I of Tipster. In addition we have developed a large scale document management system, which allows documents used in a Tipster compliant system to be handled in a uniform manner.

CRL has investigated the problem of information retrieval against collections of text written in multiple languages. This multilingual information retrieval capability is being designed so that it can be integrated into any statistical information retrieval system.

Achievements

CRL has been heavily involved in the design of the Tipster Architecture. Prototype document managers supporting the architecture were implemented and used to support the Tipster 6 and 12 month demonstration systems. A mature version of the document manager software has now been developed and distributed. CRL has also developed a Tipster Architecture Validation Suite which allows the testing of Tipster Compliant Document Managers. Both these products will now be subject to an engineering review board.

The CRL has provided multilingual Human-Computer Interface software, which conforms to the Tipster architecture. This includes a sophisticated editor which allows the display and editing of annotations on documents. A user can work with annotations produced by any Tipster compliant language processing software (e.g name recognizers, phrase spotters). The editor supports multiple languages, including Arabic, Japanese, Spanish, Chinese and Russian. All the CRL graphical user interface tools are now available to government agencies and to other research groups (see paper on Graphical User Interfaces).

CRL has used the architecture as the foundation for other DoD programs - Oleada and Temple (see separate summaries).

CRL has also made significant progress in its research in multi-lingual query generation. Methods have been developed to 'translate' English queries into Spanish (see research papers)

CRL has developed systems for recognizing proper names in English, Spanish, Japanese, and Chinese texts (see Multilingual Named Entity Task)

Software Packages

The following packages are available from CRL - a Tipster compliant document manager and user documentations, a Tipster document manager validation suite, a graphical user interface toolkit to support development of multi-lingual Tipster applications, and English name recognition software and data.