Multilingual Informedia: Search & Summarization on Demand from Combined English ...  

Project Title: **Multilingual Informedia: Search & Summarization on Demand from English Language and Foreign Media Sources**

PI: Howard Waclaw

CoPil's: Alex Waibel, Jaime Carbonell, Stephen E. Cross, Alex Hauptmann, Scott Steele

Sponsor: **Defense Advanced Research Project Agency** (DARPA), and **SPAWAR** (Space Warfare Systems) NRaD (Naval Research and Development)

**Project Description**

This project began in 1997 and was completed in July 2000. The purpose of the Multilingual project was to develop automated systems and tools enabling multilingual and multimedia information capture, search, retrieval, summarization and reuse. The system, built on the underlying Informedia system concepts, technology and infrastructure, is designed to access textual, audio and video (TV) information, to index, categorize, retrieve, summarize and analyze it, in one or more languages. We focused primarily on the Serbo-Croatian language to demonstrate viability and extend the system. We implemented and demonstrated a prototype system that was a multi-language browser of text, video and radio material that accepts English queries and returns the most relevant English and English language reports or segments in their original language, in full or summary form. This enables the analyst to compare divergent American and foreign reporting of the same topic. The semantic-expansion translation that we use reconstructs all consistent meanings and phrases in the English query, resulting in an expanded target language query without loss of meaning but at some cost in precision. We also built and delivered a functional broadcast news-focuser multiple network-connected offsite location including DARPA and NSA.

**Project Goals**

- Robust full-content indexing, search and retrieval of text, audio and video documents, connected speech recognition and new statistical natural language processing techniques.
- Multilingual document access via queries in English or the target languages. English queries matched by semantic-expansion translation into each target language (German, Serbo-Croatian and as an option other languages of the coalition forces including French, Italian, Spanish, Japanese and Korean). Semantic-expansion is a new information-preserving query translation method at CMU.
- On-demand summarization of individual documents, or production of synthetic summaries combining information from multiple documents focusing on maximally query-relevant and reducing cross-document redundancy, using novel methods for attaining summary sub-document information metrics, and zoom-in/zoom-out variable grain-size summaries.
- Video segmentation, indexing and summarization into meaningful and indexable semi-comprehensive fast-skin summaries; tools for extraction, annotation and reuse of desirable content.
- New statistical-learning methods for rapid training of indexing/search, categorization and summarization for new document collections and new languages.

**Project Background**

From its inception in 1995, the Informedia project's goal has been to allow search and retrieval medium, similar to what is available today for text only. To enable this access to video content recognition is used to provide a text transcript for the audio track; and image processing detects boundaries, recognizes faces and allows for image similarity comparison. Everything is indexed searchable information video library, where users can ask queries and receive relevant news stories. The Multilingual Informedia Project pursues a seamless extension of the Informedia approach.

http://www.informedia.cs.cmu.edu/mli/index.html

3/13/2006
and discovery across video documents in multiple languages. Previously, we successfully demonstated that current speech recognizers allow accurate information retrieval for automatically processed news TV broadcasts. The multilingual system should perform speech recognition on foreign language news broadcasts, segment it into stories and index the foreign data together with existing English data. This first multilingual prototype should easily be extensible to other languages. There are several components to the Multilingual Informedia system that differ significantly from the original Infoseek system:

- The speech recognizer recognizes a foreign language, specifically Serbo-Croatian.
- A keyword-based translation module transforms English queries into Serbo-Croatian, and searches for equivalent words in a joint corpus of English and Serbo-Croatian news broadcasts.
- English topic labels for the foreign language news stories allow a user to identify a relevant topic in the target language.

We built upon an existing technological base at Carnegie Mellon, integrating several previous areas of investigation, including speaker-independent, connected speech recognition; text-retrieval; demand summarization; machine translation; image processing; automatic capture and digitalization of multimedia information (text, audio, and video); and intelligent aids to the creation and reuse of multimedia information.