Columbia Newsblaster: Multilingual News Summarization on the Web

David Kirk Evans
Columbia University
devans@cs.columbia.edu

Judith L. Klavans
Columbia University
klavans@cs.columbia.edu

Abstract

We propose to show the new multilingual version of the Columbia Newsblaster news summarization system. The system addresses the problem of user access to browsing news in multiple languages from multiple sites on the internet. The system automatically collects, organizes, and summarizes news in multiple source languages, allowing the user to browse news topics with English summaries, and compare perspectives from different countries on the topics.

1 Multilingual Columbia Newsblaster

We propose to present a demo of the multilingual version of Columbia Newsblaster which is currently in development. Columbia Newsblaster[1] (McKeown et al., 2002) is a system for news browsing that crawls news from the web, clusters related articles, summarizes the multi-document clusters, and presents the clusters in a browsing interface separated into pre-defined categories.

In this demo, participants will be able to

- Browse summaries of current news from multiple languages
- View news clusters with documents from a particular language
- Compare summaries from documents in different languages

The Multilingual version of Columbia Newsblaster is built upon the English version of Columbia Newsblaster, taking advantage of the existing system by translating documents into English. The system has six major phases: **crawling**, **article extraction**, **clustering**, **summarization**, **classification**, and **interface generation**. In the demo, we will show the new **translation** phase, and changes made to encode all text in UTF8 Unicode. Figure 1 depicts the multilingual Columbia Newsblaster architecture, highlighting the changes made to support multiple languages.

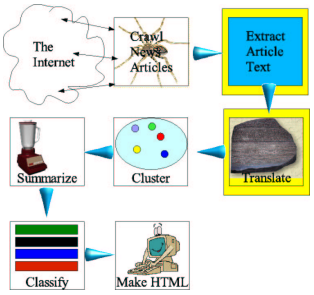

Figure 1: Multilingual CU Newsblaster Architecture.

The multilingual version of Columbia Newsblaster crawls web sites in foreign languages and extracts article text from the HTML pages. Non-English documents are translated and clustered with English documents using the existing document clustering system. For translation we use an interface to the babelfish translation system, and an interface to a statistical Arabic machine translation system from IBM for Arabic documents. The resulting document clusters are then summarized, with different summaries created for documents based on country and language.

Extracting article text

Our previous approach to extracting article text was hand crafted for English sites; to support non-English sites we incorporated a new article extraction module that uses machine learning techniques to identify the article text. The new module parses HTML into blocks of text and computes a set of 34 simple text features for each text block. Training data is generated using a GUI, and Ripper (Cohen, 1996) is used to induce a hypothesis for categorizing text. The is currently working with sites in English, Russian, Japanese, Chinese, French, Spanish, Ger-

[1] http://newsblaster.cs.columbia.edu/

Figure 2: A screen shot comparing a summary from English documents to a summary from German documents.

man, Italian, Portuguese, Korean, and Arabic. Over an English training set composed of 353 articles, the extractor had 89% recall and 90% precision. It had similar performance for Russian and Japanese training sets. Using rules learned for one language on a different data set significantly degraded performance, showing that the system is able to adapt to different sites and languages. Adding new sites is easy; a human annotates web pages using the GUI, and a new categorization hypothesis is learned from the new training data.

Summarization

We are currently working on approaches to multilingual document summarization that attempt to replace ill-formed clauses in the summary from machine translated output with clauses from the input English articles.

The system generates summaries based on the language and country of the articles, and allows the user is able to compare them side-by-side. Figure 1 shows two summaries of articles about talks between America, Japan, and Korea over nuclear arms. One summary covers articles from America, the United Kingdom, Japan, and Germany compared to a summary from articles just from Germany.

Related Research

The SUMMARIST summarization system (Hovy and Lin, 1999), integrated into the MuST System (Lin, 1999) uses query translation to allow a user to search for documents in a variety of languages, extracts sentences from documents in a variety of languages, and translates the resulting summary. The Keizei system (Ogden et al., 1999) uses query translation to search Japanese and Korean documents in English, and displays query-specific summaries focusing on passages containing query terms. (Chen and Lin, 2000) describe a system that separately clusters English and Chinese news into topics, relates clusters that are similar across languages, and generates a summary by linking sentences that are similar from the two languages. Our system uses clustering instead of search to organize documents, and produces summaries from the translations. We support many languages (8 currently) and also provide summaries for documents from each language, allowing comparisons between them.

Evaluation

The ideal summary for a set of documents differs greatly depending on the intended application for the user. We are participating the Document Understanding Conference (DUC)[2] conference this year, and will use the foreign language summarization track, with English and machine-translated Arabic input documents to evaluate the approaches we are using.

Conclusions

In this demo we present a system that generates English summaries of news from around the world that allows access and browsing capabilities to foreign language news. We have solved the problem of extracting article text in a portable manner applicable to many languages, and address the problem of generating English summaries with text from heterogeneous machine-translation systems.

References

Hsin-Hsi Chen and Chuan-Jie Lin. 2000. A multilingual news summarizer. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 159–165.

William W. Cohen. 1996. Learning trees and rules with set-valued features. In *AAAI/IAAI, Vol. 1*, pages 709–716.

E.H. Hovy and Chin-Yew Lin. 1999. Automated text summarization in summarist. In I. Mani and M. Maybury, editors, *Advances in Automated Text Summarization*, chapter 8. MIT Press.

Chin-Yew Lin. 1999. Machine translation for information access across the language barrier: the must system. In *Machine Translation Summit VII*, September.

Kathleen R. McKeown, Regina Barzilay, David Kirk Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Human Language Technology Conference*.

William Ogden, James Cowie, Mark Davis, Eugene Ludovik, Hugo Molina-Salgado, and Hyopil Shin. 1999. Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system. In *SIGIR/DL Workshop on Multilingual Information Discovery and Access (MIDAS)*, August.

[2]http://duc.nist.gov/