Columbia Newsblaster: Multilingual News Summarization on the Web

David Kirk Evans **Judith L. Klavans** **Kathleen R. McKeown**
Department of Computer Science
Columbia University, NY, NY 10027
{devans, klavans, kathy}@cs.columbia.edu

Abstract

We present the new multilingual version of the Columbia Newsblaster news summarization system. The system addresses the problem of user access to browsing news from 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 Introduction

The Columbia Newsblaster[1] system has been online and providing summaries of topically clustered news daily since late 2001 (McKeown et al., 2002). The goal of the system is to aid daily news browsing by providing an automatic, user-friendly access to important news topics, along with summaries and links to the original articles for further information. The system has six major phases: **crawling**, **article extraction**, **clustering**, **summarization**, **classification**, and **web page generation**.

The focus of this paper is to present the entire multilingual Columbia Newsblaster system as a platform for multilingual multi-document summarization experiments. The phases in the multilingual version of Columbia Newsblaster have been modified to take language and character encoding into account, and a new phase, **translation**, has been added. Figure 1 depicts the multilingual Columbia Newsblaster architecture. We will describe the system, in particular a method using machine learning to extract article text from web pages that is applicable to different languages, and a baseline approach to multilingual multi-document summarization.

1.1 Related Research

Previous work in multilingual document summarization, such as the SUMMARIST system (Hovy and Lin, 1999) extracts sentences from documents in a variety of languages, and translates the resulting summary. This system has been applied to Information Retrieval in the MuST System (Lin, 1999) which uses query translation to allow a user to search for documents in a variety of languages, summarize the documents using SUMMARIST, and translate the summary. The Keizei system (Ogden et al., 1999) uses query translation to allow users to search Japanese and Korean documents in English, and displays query-specific summaries focusing on passages containing query terms. Our work differs in the document clustering component – we cluster news to provide emergent topic structure from the data, instead of using an information retrieval model. This is useful in analysis, monitoring, and browsing settings, where a user does not have an a priori topic in mind. Our summarization strategy also differs from the approach taken by MuST in that we focus our effort on the summarization system, but only target a single language, shifting the majority of the multilingual knowledge burden to specialized machine translation systems. The Keizei system has the advantage of being able

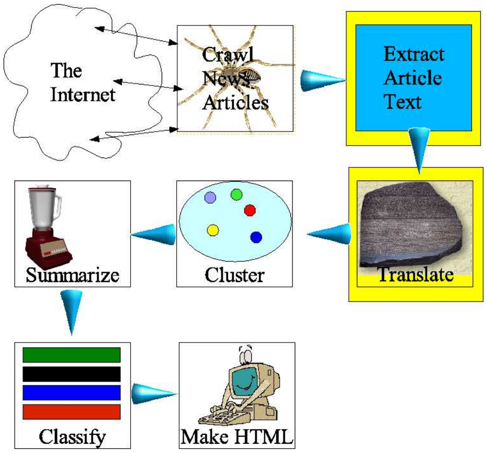


Figure 1: Architecture of the multilingual Columbia Newsblaster system.

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

to generate query-specific summaries.

Chen and Lin (Chen and Lin, 2000) describe a system that combines multiple monolingual news clustering components, a multilingual news clustering component, and a news summarization component. Their system clusters news in each language into topics, then the multilingual clustering component relates the clusters that are similar across languages. A summary is generated by linking sentences that are similar from the two languages. The system has been implemented for Chinese and English, and an evaluation over six topics is presented. Our clustering strategy differs here, as we translate documents before clustering, and cluster documents from all languages at the same time. This makes it easy to add support for additional languages by incorporating a new translation system for the language; no other changes need to be made. Our summarization model also provides summaries for documents from each language, allowing comparisons between them.

2 Extracting article data

2.1 Extracting article text

To move Columbia Newsblaster into a multilingual capable environment, we must be able to extract the "article text" from web pages in multiple languages. The article text is the portion of a web page that contains the actual news content of the page, as opposed to site navigation links, ads, layout information, etc. Our previous approach to extracting article text in Columbia Newsblaster used regular expressions that were hand-tailored to specific web sites. Adapting this approach to new web sites is difficult, and it is also difficult to adapt to foreign languages sites. We solved this problem by incorporating a new article extraction module using machine learning techniques. The new article extraction module parses HTML into blocks of text based on HTML markup and computes a set of 34 features based on simple surface characteristics of the text. We use features such as the percentage of text that is punctuation, the number of HTML links in the block, the percentage of question marks, the number of characters in the text block, and so on. Since the features are relatively language independent they can be computed for and applied to any language.

Training data for the system is generated using a GUI that allows a human to annotate text candidates with one of fives labels: "ArticleText", "Title", "Caption", "Image", or "Other". The "ArticleText" label is associated with the actual text of the article which we wish to extract. At the same time, we try to determine document titles, image caption text, and image blocks in the same framework. "Other" is a catch-all category for all other text blocks, such as links to related articles, navigation links, ads, and so on. The training data is used with the machine learning program Ripper (Cohen, 1996) to induce a hypothesis for categorizing text candidates according to the features. This approach has been trained on web pages from sites in English, Russian, and Japanese as shown in Table 1, but has been used with sites in English, Russian, Japanese, Chinese, French, Spanish, German, Italian, Portuguese, and Korean.

Language	Training set	Precision	Recall
English	353	89.10%	90.70%
Russian	112	90.59%	95.06%
Russian	English Rules	37.66%	73.05%
Japanese	67	89.66%	100.00%
Japanese	English Rules	100.00%	20.00%

Table 1: Article extractor performance for detecting article text in three languages.

The English training set was composed of 353 articles, collected from 19 web sites. Using 10-fold cross-validation, the induced hypothesis classify into the article text category with a precision of 89.1% and a recall of 90.7%. Performance over Russian data was similar, with a precision of 90.59% and recall of 95.06%. We evaluated the English hypothesis against the Russian data to observe whether the languages behave differently. As expected, the English hypothesis resulted in poor performance over the Russian data, and we saw comparable results for Japanese. The same English hypothesis performs adequately on other English sites not in the training set, so the differences between languages seem to be significant.

2.2 Title and date extraction

The article extraction component also determines a title for each document, and attempts to locate a publishing date for the articles. Title identification is important since in a cluster, sometimes with as many as 60 articles, the only information the user sees are the titles for the articles; if our system chooses poor titles, they will have a difficult time discriminating between the articles. If the article extraction component finds a title it is used. Unfortunately, this process is not always successful, so we have a variety of fall-back methods, including taking the title from the HTML TITLE tag, using heuristics to detect the title from the first text block, and using a portion of the first sentence. These approaches led to many uninformative titles extracted from the non-English sites, since they were developed for English news. We implemented a system to identify titles that are clearly non-descriptive, such as "Stock Market News", that would apply to non-English text as well. We record the titles seen and rejected over time and use the list to reject titles with high frequency. A title with high frequency is assumed to be not descriptive enough to give a clear idea of the content of an article in a cluster of similar articles. To correctly extract dates for articles, we use heuristics to identify sequences of possible dates, weigh them, and choose the most likely date as the publication date. Regular expressions for Japanese date extraction were added to the system.

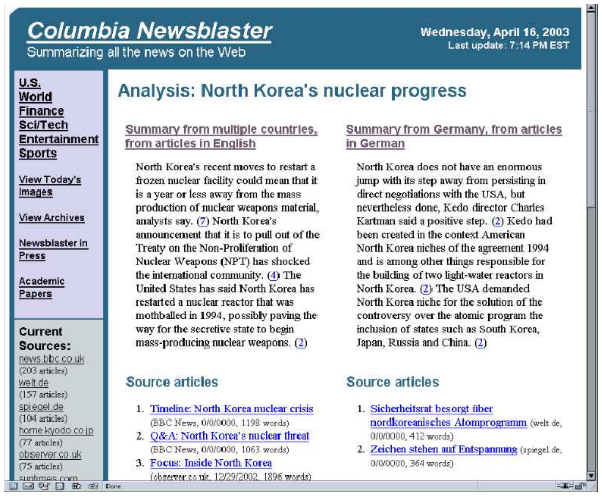
3 Multilingual Clustering

The document clustering system that we use (Hatzivassiloglou et al., 2000) has been trained on, and extensively tested with English. While it can cluster documents in other languages, our goal is to generate clusters with documents from multiple languages, so a baseline approach is to translate all non-English documents into English, and then cluster the translated documents. We take this approach, and further experimented with using simple and fast techniques for glossing the input articles for clustering. We developed simple dictionary lookup glossing systems for Japanese and Russian. Our experimentation showed that full translation using Systran outperformed our glossing-based techniques, so the glossing techniques are not used in the current system.

4 Multilingual Summarization Baseline

Our baseline approach to multilingual multi-document summarization is to apply our English-based summarization system, the Columbia Summarizer (McKeown et al., 2001), to document clusters containing machine-translated versions of non-English documents. The Columbia Summarizer routes to one of two multi-document summarization systems based on the similarity of the documents in the cluster. If the documents are highly similar, the Multigen summarization system (McKeown et al., 1999) is used. Multigen clusters sentences based on similarity, and then parses and fuses information from similar sentences to form a summary.

The second summarization system used is DEMS, the Dissimilarity Engine for Multi-document Summarization (Schiffman et al., 2002), which uses a sentence extraction approach to summarization. The resulting summary is then run through a named entity recovery tool (Nenkova and McKeown, 2003), which repairs named entity references in the summary by making the first reference descriptive, and shortening subsequent reference mentions in the summary. Using an unmodified version of DEMS, summaries might contain sentences from translated documents which are not grammatically correct. The DEMS summarization system was modified to prefer choosing a sentence from an English article if there are sentences that express similar content in multiple languages. By setting different weight penalties we can take the quality of the translation system for a given language pair into account.



Analysis: North Korea's nuclear progress

Summary from multiple countries, from articles in English

North Korea's recent moves to restart a frozen nuclear facility could mean that it is a year or less away from the mass production of nuclear weapons material, analysts say. (7) North Korea's announcement that it is to pull out of the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) has shocked the international community. (4) The United States has said North Korea has restarted a nuclear reactor that was mothballed in 1994, possibly paving the way for the secretive state to begin mass-producing nuclear weapons. (2)

Source articles

1. Timeline: North Korea nuclear crisis (BBC News, 0/0/0000, 1198 words)
2. Q&A: North Korea's nuclear threat (BBC News, 0/0/0000, 1063 words)
3. Focus: Inside North Korea (observer.co.uk, 12/29/2002, 1896 words)

Summary from Germany, from articles in German

North Korea does not have an enormous jump with its step away from persisting in direct negotiations with the USA, but nevertheless done, Kedo director Charles Kartman said a positive step. (2) Kedo had been created in the context American North Korea niches of the agreement 1994 and is among other things responsible for the building of two light-water reactors in North Korea. (2) The USA demanded North Korea niche for the solution of the controversy over the atomic program the inclusion of states such as South Korea, Japan, Russia and China. (2)

Source articles

1. Sicherheitsrat besorgt über nordkoreanisches Atomprogramm (welt.de, 0/0/0000, 412 words)
2. Zeichen stehen auf Entspannung (spiegel.de, 0/0/0000, 364 words)

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

4.1 Similarity-based Summarization

As part of our multilingual summarization work, we are investigating approaches to summarization that use sentence-level similarity computation across languages to cluster sentences by similarity, and then generate a summary sentence using translated portions of the relevant sentences. The multilingual version of Columbia Newsblaster provides us with a platform to frame future experiments for this summarization technique. We are investigating translation at different levels - sentence level, clause level, and phrase level. Our initial similarity-based summarization system works at the sentence level. Starting with machine-translated sentences, we compute their similarity to English sentences that have been simplified(Siddharthan, 2002). Foreign-language sentences that have a high enough similarity to English text are replaced (or augmented with) the similar English sentence.

This first system using full machine translation over the sentences and English similarity detection will be extended using simple features for multilingual similarity detection in SimFinder MultiLingual (SimFinderML), a multilingual version of SimFinder (Hatzivassiloglou et al., 2001). We also plan an experiment evaluating the usefulness of noun phrase detection and noun phrase variant detection as a primitive for multilingual similarity detection, using tools such as Christian Jacquemin's FASTR (Jacquemin, 1994; Jacquemin, 1999).

4.2 Summary presentation

Multilingual Newsblaster presents multiple views of a cluster of documents to the user, broken down by language and by country. Summaries are generated for the entire cluster, as well as sub-sets of the articles based on the country of origin and language of the original articles. Users are first presented with a summary of the entire cluster using all documents, and then have the ability to focus on countries or languages of their choosing. We also allow the user to view two summaries side-by-side so they can easily compare differences between summaries from different countries. For example, figure 4.2 shows a summary of articles about talks between America, Japan, and Korea over nuclear arms, comparing the summaries from articles in English and German.

5 Evaluation

Evaluation of multi-document summarization is a difficult task; the Document Understanding Conference (DUC)[2] is designed as an evaluation for multi-document summarization systems. We participated in the DUC 2004 conference submitting the results of the summarization system used in Newsblaster, as well as an in-progress system described in Section 4.1 for multilingual cluster summarization. The results of the DUC evaluation will provide us with valuable feedback on the multi-document multi-lingual summarization components in Newsblaster.

6 Conclusions

In this paper we have described a multilingual version of Columbia Newsblaster, a system that runs daily offering users an accessible interface to online news browsing. The multilingual version of the system incorporates two varieties of machine translation, one for clustering, and one for translation of documents for summarization. Existing summarization methods have been applied to translated text, with plans for an evaluation of the current method, and incorporation of summarization techniques specific to translated documents. The system presents a platform for further multilingual summarization experiments and user-oriented studies.

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.

Vasileois Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*.

Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa Holcombe, Regina Barzilay, Min-Yen Kan, and Kathy McKeown. 2001. Simfinder: A flexible clustering tool for summarization. In *Proceedings of the North American Association for Computational Linguistics Automatic Summarization Workshop*.

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.

Christian Jacquemin. 1994. Fastr: a unification-based front-end to automatic indexing. In *In Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94)*, pages p. 34–47.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348.

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

Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI*, pages 453–460.

Kathleen R. McKeown, Regina Barzilay, David Kirk Evans, Vasileios Hatzivassiloglou, Min-Yen Kan, Barry Schiffman, and Simone Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the Document Understanding Conference*.

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*.

Ani Nenkova and Kathy McKeown. 2003. References to named entities: A corpus study. In *Short Paper Proceedings of NAACL-HLT*.

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.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*, March.

Advaith Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 60–65, Philadelphia, USA.

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