MINDS - Multi-lingual INteractive Document Summarization

Jim Cowie, Kavi Mahesh, Sergei Nirenburg, Remi Zajac

Computing Research Laboratory
New Mexico State University
New Mexico 88003
(jcowie, sergei, rzajac @crl.nmsu.edu)

Abstract

The research described here focuses on multi-lingual summarization (MLS). Summaries of documents are produced in their original language; corresponding summaries in English will eventually be generated. The source languages supported are Spanish, Japanese, English and Russian.

Background

The need for summarization tools is especially strong if the source text is in a language different from the one(s) in which the reader is most fluent. Interactive summarization of multilingual documents is a very promising approach to improving productivity and reducing costs in large-scale document processing. This addresses the scenario where an analyst is trying to filter through a large set of documents to decide quickly which documents deserve further processing. This task is more difficult and expensive when the documents are in a foreign language in which the analyst may not be as fluent as he or she is in English. The task is even more difficult when the documents are in several different languages. For example, the analyst's task may be to filter through newspaper articles in many different languages published on a particular day to generate a report on different nations' reactions to a current international event, such as a nuclear test on the previous day. This last task is currently infeasible for a single analyst, unless he or she understands each one of those languages, since machine translation of entire documents cannot yet meet the requirements of such a task.

Report Generation versus Summarization

Summarization is the problem of presenting the most important information contained in one or more documents. A fundamental distinction between summarization and other, well-developed areas such as information extraction (IE) and information retrieval (IR) is that unlike IE and IR, pure summarization is inherently untargeted. In the document-filtering scenario, the analyst is interested in knowing what are the main points in one or more documents, no matter what those points may be. In more targeted scenarios, however, the task is to find out what the documents have to say about a particular topic (or in a particular style). Target-specific summarization involves extracting information from multiple documents based on pre-specified templates, queries, or from different points of view (e.g., extract all facts, all opinions, all sarcasm, etc.).

Machine Translation and Summarization

Multilingual summarization (MLS) immediately introduces the problem of translating the documents to the language of the summary (i.e., English). Unfortunately, machine translation (MT) is not yet in a state where good quality translations of documents can be provided. Moreover, machine translation has always worked under the assumption that a text must be translated in its entirety. CRL hypothesizes that MLS and MT can mutually benefit from one another since summarization offers MT the benefit of not having to translate entire texts.

MINDS: Multilingual, Interactive Document Summarization

MINDS addresses primarily the multilinguality and multiple document dimensions and will

- integrate multi-lingual summarization and multi-document summarization capabilities using a multi-engine, core summarization system;
- provide fast, interactive document access through hypertext summaries;
- produce document cross-links (including links across languages) as a byproduct; and
- generate targeted documents linking a variety of information about individuals.

The goals of MINDS are to -

- produce summaries both in English and in the original language of a document;
- operate in near real time;
- support interactive document access. The user can "resummarize" interactively and access relevant parts of documents through the summary.

We assume that the documents are at most a few pages long. The proposed technique may not be effective for producing a summary of a book, for example. It is also assumed for the most part that the genre of the documents is news articles.

Core Summarization Engine

The core summarization problem is taking a single text and producing a shorter text in the same language that contains all the main points in the input text. We are using a robust, graded approach to building the core engine by incorporating statistical, syntactic and document structure analyses among other techniques. This approach is less expensive and more robust than a summarization technique based entirely on a single method. The core engine is being designed in such a way that as additional resources, such as lexical and other knowledge bases or text processing and MT engines, become available from other ongoing research efforts they can be incorporated into the overall multi-engine MINDS system. Ideally the core engine itself will remain language independent.

A prototype core engine has been built for English, Spanish, Russian, and Japanese documents. A demonstration of the core engine for web pages can be found at

http://crl.nmsu.edu/Research/Projects/minds/demonstrations.hmtl

Core Summarization Techniques

Document Structure Analysis

Document structure analysis is important for extracting the topic of a text [Paice and Jones, 1993; Salton and Singhal, 1994]. In a statistical analysis for example, titles and sub-titles would be given a more important weight than the body of the text. Similarly, introduction and conclusion for the text itself and for each section are more important than other paragraphs, and the first and last sentences in each paragraph are more important than others. The applicability of these depends, of course, on the style adopted in a particular domain, and on the language: the stylistic structure and the presentation of arguments vary significantly across genres and languages. Structure analysis must be tailored to a particular type of text in a particular language. In the MINDS system document structure analysis involves the following subtasks:

- Language Identification: CRL's language recognizer automatically selects subsequent processors based on the language of the document. This recognizes both the language and the character encoding based on a short section of the document.
- Document Structure Parsing: If the documents have SGML or HTML encoding then a parsing process separates the title and subheadings, sections and subsections, and other data and graphics. If the documents do not contain a markup then various heuristics are used to identify the components of the document.
- Multilingual Sentence Segmentation: Sentence segmentation is also language dependent. In particular the overloading of the full stop to carry out other functions, such as indicating abbreviation, varies from language to language. Chinese and Japanese do not suffer from this problem the stop character is unambiguous.
- Text Structure Heuristics: the MINDS summarizer uses rules based on document structure to rank sentences. Specialized sets of rules are needed for different styles, or domains. At the moment, however, it has not been found necessary to develop different rules for different languages.

In order to allow a multitude of techniques to contribute to sentence selection, the core engine adopts a flexible method of scoring the sentences in a document by each of the techniques and then ranking them by combining the different scores. Text-structure based heuristics provide the main method for ranking and selecting sentences in a document. These are supplemented by word frequency analysis methods.

Word Frequency Analysis

The word frequency analysis used in MINDS at present is naively simple. It is our intention to use more sophisticated techniques later in the development of the system. Our primary goal was to ensure that all four languages were handled in the same way. The basic technique is to sort the words in the document by frequency and select a few of the most frequent content words (i.e., words other than articles, prepositions, conjunctions and other closed-class words). Sentences containing those words get a score increment.

Word segmentation is based on white space characters for English, Russian and Spanish; for Japanese we rely at the moment on generating every sequential pair of characters in a text as a word. No morphological analysis is carried out.

Future developments will involved the incorporation of a Japanese word segmentation program; inclusion of morphological analysis for the other languages; incorporation of proper name recognition software to recognize more significant units in a text; and the use of corpus statistics to provide normalized weightings for word usage.

References

Paice, C.D. and Jones, P.A. 1993. The identification of important concepts in highly structured technical papers. In Proceedings of the 16th ACM SIGIR conference, Pittsburgh PA, June 27-July 1, 1993; pp.69-78.

Salton, G. and Singhal, A. 1994. Automatic text theme generation and the analysis of text structure. Technical Report TR94-1438, Department of Computer Science, Cornell University, Ithaca, N.Y.