Improving Robust Domain Independent Summarization

Jim Cowie, Eugene Ludovik, Hugo Molina-Salgado
Dept. 3CRL, Box 30001, NMSU, Las Cruces, NM 88003, USA
(jcowie, eugene, hsalgado)@crl.nmsu.edu

Abstract

We discuss those techniques which, in the opinion of the authors, are needed to support robust automatic summarization. Many of these methods are already incorporated in a multi-lingual summarization system, MINDS, developed at CRL. The approach is sentence selection, but includes techniques to improve coherence and also to perform sentence reduction. Our methods are in distinct contrast to those approaches to summarization by deep analysis of a document followed by text generation.

KEYWORDS

Summarization, Multi-lingual Language Engineering, Robust Methods

1 Introduction

Summarization is the problem of presenting the most important information contained in one or more documents. The research described here focuses on multi-lingual summarization (MLS). Summaries of documents are produced in Spanish, Japanese, English and Russian using the same basic 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. We have developed a system design which allows the parameterization both of the summarization process and of necessary information about the languages being processed.

Document structure analysis (Salton & Singal 94, Salton et al. 95) is important for extracting the topic of a text. In a statistical analysis for example (Paice 90, Paice & Jones 93), titles and sub-titles would be given a more important weight than the body of the text. Similarly, the 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 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
- Document Structure Parsing
- Multilingual Sentence Segmentation
- Text Structure Heuristics

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.

The core engine is 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. The most promising components are part of speech tagging, anaphora resolution, and semantic methods to allow concept identification to supplement word

frequency analysis. Part of speech tagging has already been used to perform sentence length reduction by stripping out "superfluous" words and phrases. The other methods will be used to maintain document coherence, and to improve sentence selection and reduction.

In this paper we describe the architecture and performance of the current system and our plans for incorporating new NLP methods.

2 MINDS - Multi-Lingual Interactive Document Summarization

2.1 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 (MT) of entire documents cannot yet meet the requirements of such a task. Multilingual summarization (MLS) introduces the possibility of translating a summary rather than the entire document to the language of the summary (i.e., English). We hope that MLS and MT can mutually benefit from one another since summarization offers MT the benefit of not having to translate entire texts and also spares a user from having to read through an entire document produced by an MT system.

2.2 Overview

The MINDS system is a multilingual domain independent summarization system, which is able to summarize documents written in English, Japanese, Russian and Spanish. The system is intended to be rapidly adaptable to new language and genres by adjusting a set of parameters. A summarization system for Turkish has just been added to the system. This required about one programmer day of effort, mostly spent in preprocessing the language

Figure 1. Overview of the MINDS Architecture

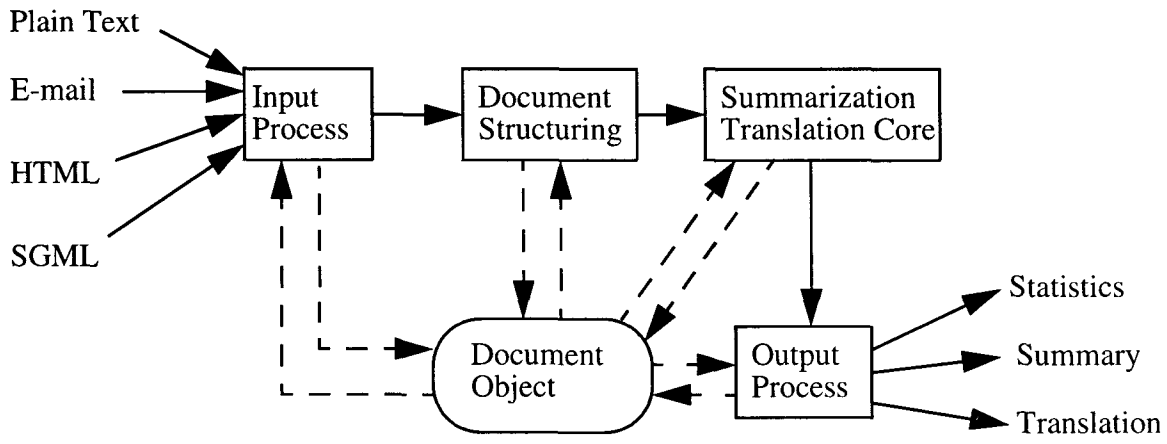
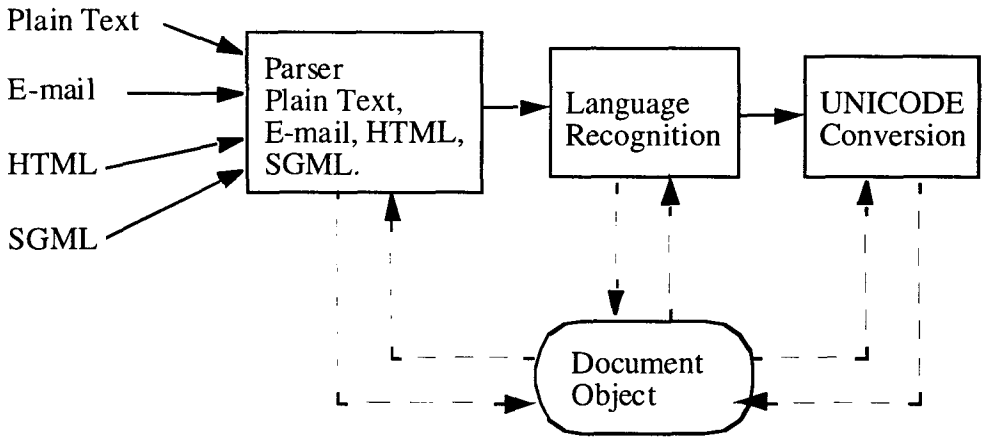


Figure 2. Input Processing Stage



resources used by the system. The types of summarization information used are also intended to be adjustable by a user "on the fly", to allow the tuning of the summarizers output based on length of summary needed, type of document structure, topic focus.

The MINDS summarization system is composed of four stages. First we have an Input Process stage, whose main function is to get the relevant text in the document in UNICODE encoding. The second stage is a Document Structuring Stage, where paragraph and sentence recognition, and word tokenization are performed. All the information about the document structure is stored in a "Document Object" that will be used in the Summarization-Translation stage. In the Summarization-Translation Stage, the text is summarized using sentence extraction techniques, where the sentence scoring and ranking is mainly based on text-structure based heuristics supplemented by word frequency analysis methods and in some cases by information from a Name Recognition module. Once the summary is ready in the original

Figure 3. Document Structuring Stage

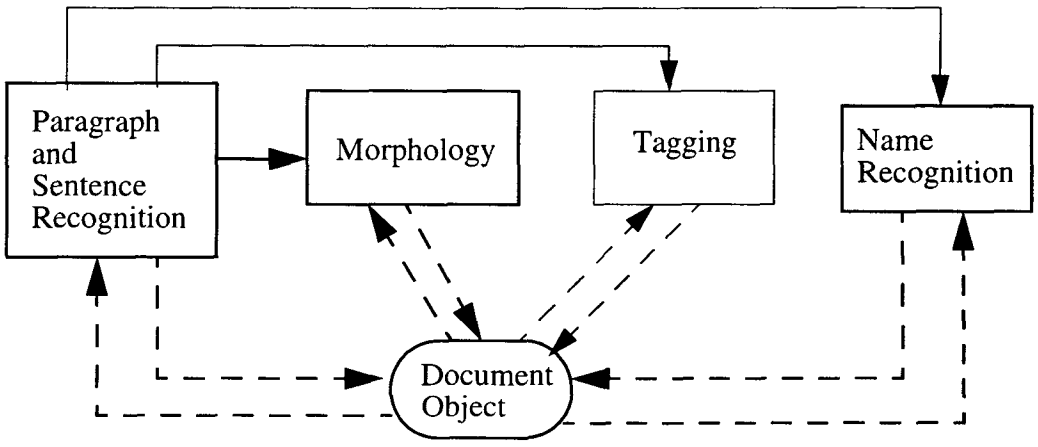
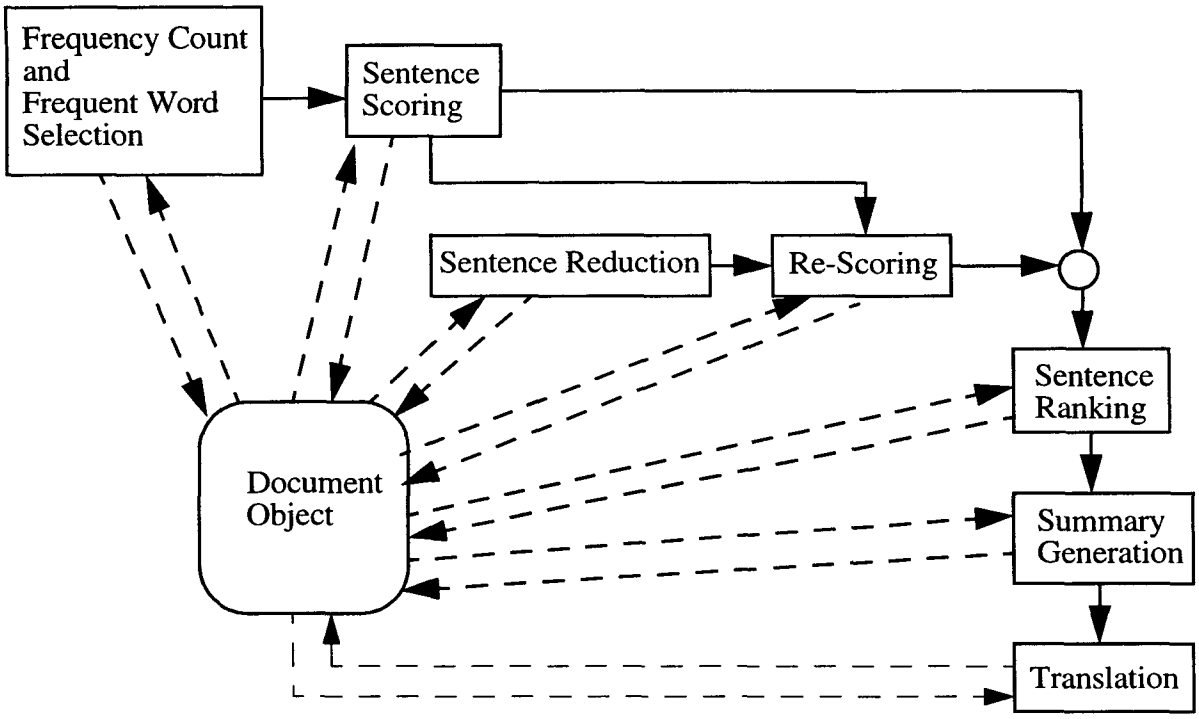


Figure 4. Summarization and Translation Stage



language, MINDS uses MT engines from other ongoing CRL projects to translate the summary to English. The final stage is the Output Process that generates the summary output form; SGML, HTML, or Plain text. This may also involve conversion from UNICODE to the original encoding of the document.

2.3 Input Process Stage

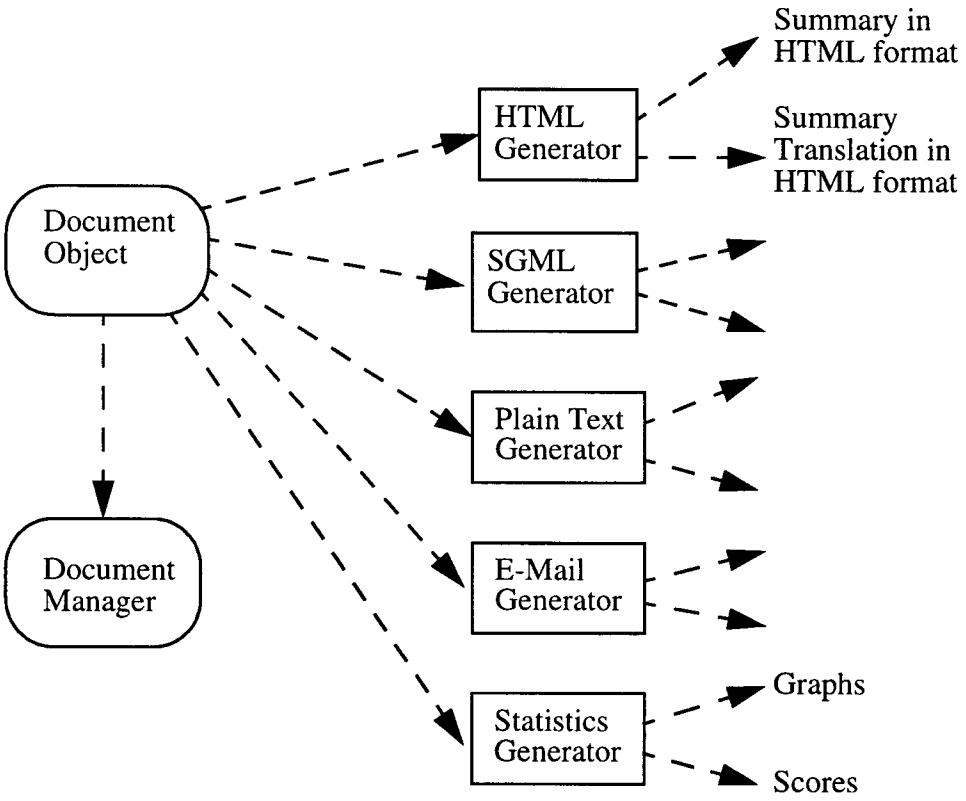
In the input stage, MINDS can accept documents written in different languages and codesets: currently English, Japanese, Russian, Turkish and Spanish. Also the documents can be in different formats such as SGML, HTML, E-mail or Plain text. A parsing stage identifies the document's format, selects and applies the appropriate parser and extracts the relevant text from the document. Once we have the text to be summarized a language recognition module determines the language in which the document is written and the text encoding used in the document. Given the encoding of the document the text is converted to UNICODE and all the rest of the processing is carried out on the UNICODE version of the text.

2.4 Document Structuring Stage

After the text to be summarized is available in UNICODE encoding, its structure needs to be determined. This is the job of the Document Structuring Stage. In this stage, three tokenization stages are performed. The first one pose of identifies the paragraphs in the document. The second tokenization stage identifies sentences within each paragraph. To identify sentence boundaries for many languages requires a list of abbreviations for the language. Languages such as Chinese and Japanese have an unambiguous "stop" character and thus do not present this problem. Finally, word tokenization is carried out to identify individual words in each sentence. Here Chinese and Japanese which do not use spaces between words require some segmentation method to be applied. The current system actually uses two character pairs, bi-grams, for all its calculations for Japanese. These bi-grams are produced starting at every character position in the document.

All the structuring information is stored in a "Document Object", which is the main data structure of the system, holding all the information gen-

Figure 5. Output Process Stage



erated during the processing. After the tokenization stage is complete and depending on the lexical resources available for each language, other stages are performed, such as Morphology, Proper Name Recognition and Tagging.

2.5 Summarization-Translation Stage

In the Summarization-Translation Stage, the importance of each sentence in the document is determined using a scoring procedure, which assigns scores to the sentences according to the position of the sentences in the document structure and according to the occurrences of key-words in the sentence which belong to the set of most frequent words in the document that are not in a "stop list" (the most frequent words in a language are considered irrelevant). We make the assumption that these key-word represent or identify the main concepts in the document, therefore if a sentence contains several of them, its score should be high so it could be selected as part of the summary. It is important to note here that we need a "stop list" for each language considered in the summarization system. Also, if a Proper Name Recognition module is available for a specific language, we use the information about person names, organization names, places and dates to contribute in the scores of sentences.

At this point if the lexical resources are available, an optional sentence length reduction can be carried out using information from a tagging stage. This sentence length reduction includes the elimination of adjectives from noun phrases, keeping only the head noun in a noun phrase, eliminating adverbs from verb phrases and eliminating most of the prepositional phrases. However, if a word selected for elimination is a key word, proper noun, the name of a place, a date or a number, the word is kept in the sentences. If this word happens to be in a prepositional phrase, then the prepositional phrase is kept in the sentence.

Once the scoring process is done, the sentences are ranked and a summary is generated using the sentences with the higher scores that together do

not exceed a predetermined percentage of the document's length. This summary is written in the document's original language, so a machine translation system is used to produce an English version of the summary.

2.6 Output Process

At this point in the summarization process we have a version of the document's summary in the original language and a version in English, both encoded using UNICODE and in plain text format. The Output Process stage takes these two versions of the summary and converts the one written in the original language to the original encoding of the document (identified by the Language Recognition module), then it converts the version in English from UNICODE to "8859_1" (ISO Latin-1). After the summaries are in the proper output encoding, the system generates the summary in one of the following formats: SGML, HTML, E-mail or Plain text according to the user's specification or to system parametrization, for example, if the summarization system is being used for web delivery, then the output format will be HTML by default.

3 Extending the Summarization Capability

Our goal is to improve the usability and flexibility of the summarization system, while still retaining robustness. This is one of the main reasons why we favor the sentence selection method rather than approaches based on deep analysis and generation (Beale 94, Carlson & Nirenburg 90). Though much disparaged for lack of readability, cohesion etc. systems based in the sentence selection method performed well in the recent Tipster summarization evaluation. In fact the readability as assessed by the evaluators was as high for summaries of about 30% of the document length as it was for the original documents. We are developing summarization techniques based on information extraction and text generation. These will not give very good coverage, because of their domain specificity, but do offer advantages, particularly in the area of cross document summarization.

Our experiments have shown for English that the inclusion of other language processing techniques can indeed increase the flexibility and performance of the summarizer. In particular proper name recognition, co-reference resolution, part of speech tagging and partial parsing can all contribute to the performance of the system.

The use of proper names allows the summaries to be weighted towards sections of the documents discussing specific individuals or organizations rather than more general topics. In terms of production of informative summaries, rather than indicative summaries, this may be an important capability. This technique was used to produce summaries evaluated using a "question and answer" methodology at the Tipster evaluation and produced a high performance here.

We have not incorporated co-reference resolution methods in our system yet, but it would seem that readability can be improved by the ability to replace pronouns with their referents would be useful. It remains to be seen, however, whether sufficient accuracy can be achieved to support this method. In cases like this where an error may be critical for a user of the system we would normally mark the fact that the text had been added by the system.

Part of speech tagging and phrase recognition allows us to carry out certain kinds of text compaction. This is particularly important when very short summaries (10%) of short documents are required. Our experiments with this kind of compaction have showed reductions of about 1/3 of the summary size with some loss of readability. A single sentence example shows the usefulness of this technique.

Original Sentence

> Browning-Ferris Industries Inc. was denied a new permit for a huge Los Angeles-area garbage dump, threatening more than $1 billion in future revenue and the value of a $100 million investment.

Shortened Sentence

Browning-Ferris Industries Inc. was denied a permit for a Los Angeles-area dump, threatening more than $1 billion in revenue and the value of a $100 million investment.

We hope eventually to have sentence reduction in place for all the languages we process, and that this will also improve the readability of MT output by allowing it to process significantly simplified input.

4 Conclusions

We feel that further research is warranted on improving summarization based on sentence selection and that its bad press is largely apocryphal and unjustified. In fact from a document analysts point of view material from the original document may be preferable, carrying as it does, the style and tone of the original document.

We see significant opportunities in carrying out further research to develop and integrate language processing and other intelligent techniques such as those described above. One particularly challenging type of document is the HTML pages found on the web. Here techniques to identify coherent sections of text are required as well as methods for summarizing tables and groups of frames.

References

(Beale 94) Beale, S. 1994 Dependency-directed text generation, Technical Report, MCCS-94-272, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

(Carlson & Nirenburg 90) Carlson, L. and Nirenburg, S. 1990. World Modeling for NLP. Center for Machine Translation, Carnegie Mellon University, Tech Report CMU-CMT-90-121.

(Cohen 95) Cohen, J.D. 1995. Highlights: Language- and domain-independent automatic indexing terms for abstracting. Journal of the American Society for Information Science, 463:162-174.

(Paice & Jones 93) 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.

(Paice 90) Paice, C.D. 1990. Constructing literature abstracts by computer: Techniques and prospects. Inf.Proc. & Management 261, 171-186.

(Salton et al. 95) Salton, G., Singhal, A., Buckley, C., and Mitra, M. 1995. Automatic text decomposition using text segments and text themes. Technical Report, Department of Computer Science, Cornell University, Ithaca, NY.

(Salton & Singal 94) 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.