Generating Personal Profiles

Jim Cowie, Sergei Nirenburg, Hugo Molina-Salgado
Computing Research Laboratory, New Mexico State University

Abstract

This paper presents an integrated NLP system which combines information retrieval of web accessible documents, multilingual generic (domain independent) summarization and machine translation to generate a single monolingual (English) document from multiple documents in several languages. In particular, this application generates a time-stamped list of events connected to a particular person. These are the personal profiles of the title. The system user can intervene in the document retrieval process, before the final resume is produced, to ensure that the documents selected are about the same person. As an alternative, automatic filters are also incorporated in the system, which attempt to ensure all the documents are in the some domain. One of the key factors in the document assembly process is the assignment of absolute dates to each sentence produced by the system. These are computed based on actual dates found in the document.

1 Introduction

Today's Internet technology is allowing direct on-line access to locations all around the globe. As a result, enormous quantities of multilingual information in the form of text are becoming available online every day. This growing source of information makes possible the development of new kinds of sources for users looking for very specific information. For example, users searching for information about the activities of a famous person or a world leader, can retrieve hundreds if not thousands of documents very rapidly. These documents may be written in languages the user does not know, which poses additional difficulties. Hence, to distill the desired information into a single text a user will need to translate and then select what is relevant to his or her needs. Certainly, having so many documents will require too much time reading and translating thus canceling the benefits of fast information retrieval technology to the point of making impractical the entire process of obtaining the relevant information. The aim of the work described here is to demonstrate a fully automatic approach which generates personal profiles from multilingual documents retrieved from the Internet. This has involved the integration of several multilingual tools - automatic language recognition, generic multilingual summarization, machine translation, date recognition, to produce a system that generates personal profiles. These profiles are lists of brief entries in English presented as HTML pages with links to the summaries and documents, in the original languages. We have tested the system on 18 people. An example of the current output of the system can be seen in Figure 1.

2 Cross Document Summarization

Cross Document Summarization consists of producing a summary out of a collection of documents that refer to the same topic. Several efforts to produce "Cross Document Summaries" can be found in the literature. McKeown, Jordan, and Hatzivassiloglou, (1998) suggested a methodology to produce a tailored summary out of several medical articles. The authors dealt with the problem of finding relevant information concerning the state of a patient across several online medical documents.

In a later work, McKeown, et. al., (1999) presented a more general approach which integrated other disciplines such as machine learning and statistical techniques combined with some linguistic features and also information fusion techniques, to select relevant phrases from the documents so they can be included in the final summary. Another approach to cross-document summarization uses cross document co-reference resolution to produce a summary out of a collection of related documents (Bagga and Baldwin, 1998).

3 The Personal Profile Cross Document Summarization System

The current implementation of the system takes as input a person's name in English, Spanish, and/or Russian. Additional search terms can be added to further constrain the search. A search is then carried out on a selected web search engine and the user can see the type of documents being found. If the search is successful then the user initiates generation of the personal profile. The main experiment described below used pre-retrieved documents.

The problem of generating the activity profile of a well known person, as carried out by our system, can be broken into three main steps:

I) **Collecting and preparing the data**

- Gathering documents from the web in English, Russian and Spanish.
- Filtering the documents to reduce the data to a collection of related documents.

II) **Individual Document Summarization**

(This is done for each document in the collection)

- Determining a date for the document
- Selecting concise relevant pieces of information from the filtered collection of documents.
- Determining a date for each of the selected extracts.
- Translating these pieces of text into English (our target language).

III) **Profile Generation**

- Merging the translated text extracts in chronological order to produce the cross document summary.
- Generating the output form for the end user.

The final result of processing the collection of related documents retrieved and filtered in the first step of our approach is a cross document summary about a specific individual.

3.1 Collecting and preparing the data

The data for the initial experiments was collected from the Internet by hand. However, we now use an automatic system to collect documents from the web by harvesting data from specific news sites. An automatic language recognition tool is used to ensure the corpus thus generated contained only documents in the three languages of interest.

Language recognition allows the system to convert the documents from any encoding to Unicode characters which is the expected encoding for language-specific tools (tokenizers and machine translation engines) Our language/encoding recognition program implements a statistical mixed-order n-gram algorithm, that during the training phase extracts the most important n-grams from the training data for each language/encoding pair, and then compares the document whose language/encoding is to be determined to the language models so created.

2.1 The next step is to filter the corpus to obtain only those documents where the person in question is mentioned. This is done by searching documents in each language for all the possible inflectional forms of the person's name. This process can be simultaneously used on a list of names in our experiments. This data is automatically grouped into 18 document sets, where each set contained documents about a specific person in up to three different languages (English, Russian and Spanish).

3.2 The Individual Document Summarization Process

Once the data collection phase is complete, a set of documents concerning the person in question is ready for processing. At this point, all the documents in the set are summarized by extracting from them sentences with information about our target.

Each document is summarized in its original language using a generic multilingual summarizer whose parameters are tuned to favor text extracts that mentioned the targeted person. A different set of parameter settings can easily shift the focus of processing to a place, and event or any other known entity.

During the summarization step several complex tasks are performed:

a) Automatic language and encoding recognition. See 3.1 above.
b) Text extraction from HTML files, so the information needed is free of markup tags. This task presents a great challenge, due to the complex layout used by different web sites that include the use of frames, tables and also dynamic html. At the moment, our methods are only a first approximation to an "ideal" HTML analysis module. A new parser for HTML text is under construction.
c) Multilingual paragraph, sentence and word tokenization to get the structure of the document. This stage is important for processing documents written in different languages.
d) Date stamp determination for the documents being summarized, using a multilingual date recognition package.
e) Sentence scoring and sentence ranking carried out to produce a final set of relevant text extracts in the original language that can be considered a person-biased summary of the document. For each sentence extracted from the document, a date stamp is determined using our multilingual date recognition package. If no date is found in the sentence itself, or if a partial date is found, the date for the sentence is completed by inheriting part or all the date previously determined for the document.
f) Translation to the target language (English). After the individual document summary is completed for one document, if the original language of the document was not English, all extracted sentences are translated.

3.3 Generating the profile

After all documents are summarized in this fashion, the translated text extracts are sorted according to their date stamps. The sorted sentences are arranged for viewing using HTML markup. Links are provided to the document summaries in the original languages and also to the complete documents. This is to allow verification of details and also to support system debugging.

3.4 Date Recognition and Utilization

Accurate date recognition is critical for the operation of this system. At the moment we rely on explicitly stated dates and not on referents like "tomorrow", "last year" etc. These have been used in previous systems developed by the authors for English (Cowie et al., 1993), but insufficient time has been available to extend the capability of attaching explicit dates to time referents for our other two languages. This type of *language ecology* capability is required for many tasks and should be developed as a shareable resource

The current date recognizer contains language specific date formats stored in patterns containing up to three letters, each representing year, day or month. Years expressed in a 4-digit format are represented by letter 'Y', 2-digit years – by 'y'. For months in a 2-digit format, letter 'm' is used, whereas month names, like January, are represented by letter 'M'. For instance, the full date in American format 11/21/1995 is represented by the pattern "mDY". The full date pattern with a year expressed by two digits is "mDy". An incomplete date like "January 22" can be presented by a pattern "MD". Once a date is detected in say, a Russian sentence, its elements are extracted, evaluated and converted to a standard language-independent format.

At the moment there are only two heuristics for date establishment. An explicit date in a sentence overrides the document date, otherwise the document date should be used. A more sophisticated treatment of date elicitation is obviously desirable. However, the primitive method described here already produces usable results.

4 Results

The experiments were performed in a collection of 923 documents, which were retrieved from the Internet. These documents could be written in any of the three languages considered in this work, that is English, Russian or Spanish. Profiles were generated for the following people: Adams, Annan, Berezovsky, Cardozo, Castro, Cook, Frei, Gates, Hashimoto, Hussein, Lebed, Lukashenka, Menem, Mubarak, Nemtsov, Pope John Paul II, Primakov, and Pujol. The resulting documents, we feel, show that the method has promise. Significant sections of each document can be read as a sequence of biographical notes on the person selected. Thus, for our first example we can see at a glance when and where Robin Cook was born, educated, and first elected. Later Spanish sources provide more regional information on the discussions over the Falkland Islands. Interspersed are some more useless pieces of information, such as the date of a fax, which was in fact the source of other information in the summary document.

5 Further Work

There are many problems concerning Cross Document Summarization that need to be treated to improve a system like this, such as the possibility of finding relevant sentences from different documents that contradict each other or sentences that talk about the same events, therefore causing repetition in the final cross document summary. Also adding techniques for anaphora resolution, a problem for any summarization technique that uses text extraction to produce summaries, will improve the quality of our system.

Another key problem that needs to be mentioned here is Cross-Document Co-reference. When we collect documents relevant to a specific person, we use the person name as a query term for the retrieval process. Now, if it happens that there are more than one well (or not so well) known person with the same name, we can (and do) end up with a ambiguous set of documents that are going to contain articles

about Berezovsky, the musician, and articles about Berezovsky, the politician, and text extracts talking about both will be present in the cross document summary, although we are interested only in one of them. An approach we are testing to solve this problem is to filter the documents using domain information, so for example only documents about politics are selected if we are interested in the politician or only documents about music are selected if we are interested in the musician. This disambiguation is now done during the retrieval process by modifying the query to include selected terms about the relevant domain. This, however, will have some undesirable effects in reducing the scope of the information available to the system.

One fundamental problem encountered in our present system is the attachment of dates to each sentence (or sub-sentence) in a document. A base date for the document needs to be established as a reference point. Currently this is determined by the dates at the front of the document. Datelines, however, regularly appear at the end of web and newswire documents. In addition some recognition of the temporal discourse structure of the document needs to be carried out. Does a sentence, or paragraph contain a change to the base date, which should apply to the rest of the document, or is it a side reference, with its own date, but having no bearing on the material that follows? A study of this temporal co-reference, which may be affected by genre, topic and possibly other factors would be very useful for developing our work further.

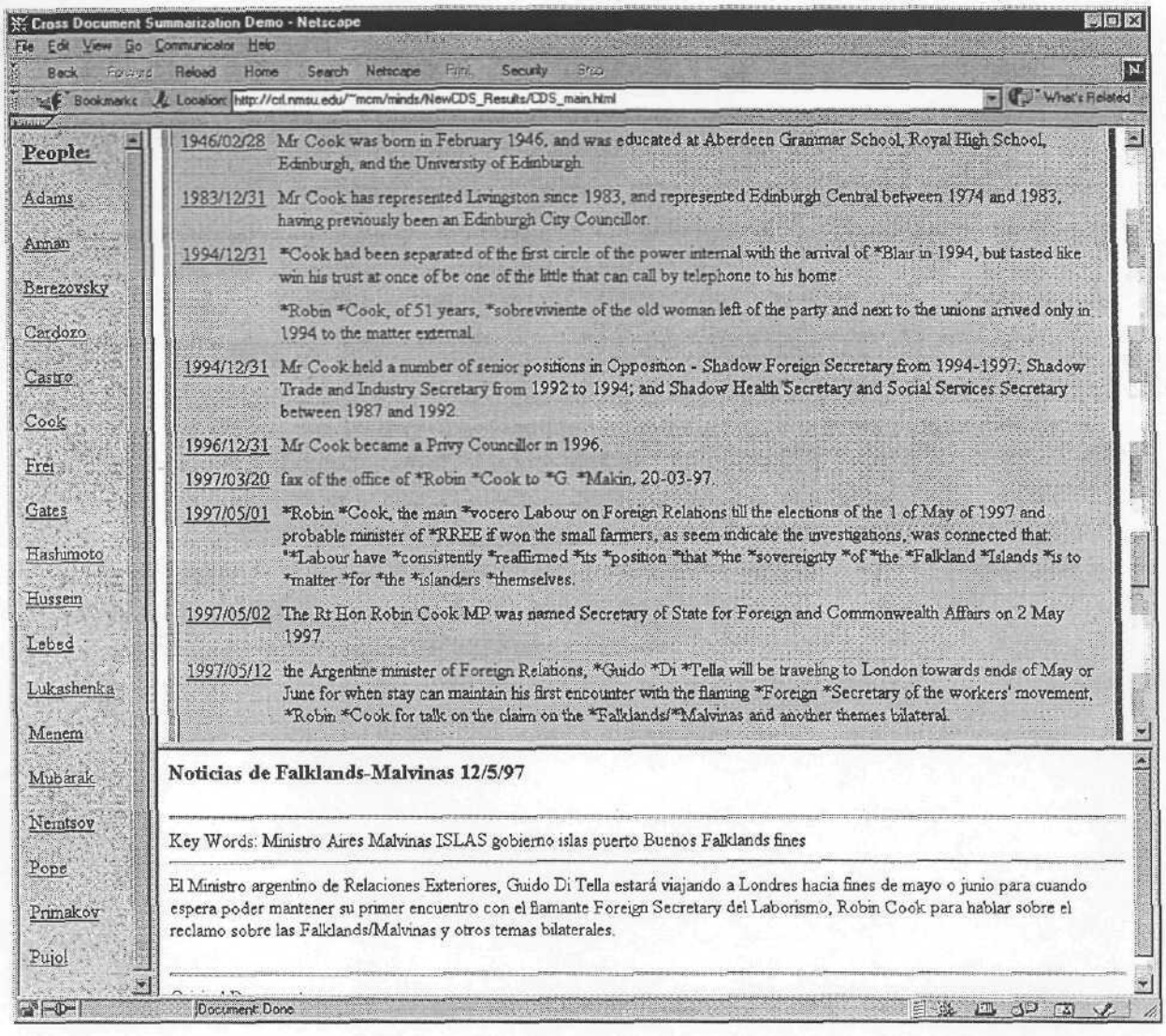

Figure 1. Personal Profile of Robin Cook, British Foreign Secretary
Notes: The asterisks in the text are words transliterated by the translation systems. Our MT system resorts to transliteration if it fails to provide a "real" translation.
The dates provide links to the summaries shown in the lower frame, and this in turn contains a link to the original document. The text shown here is from a Spanish source.

References

Bagga, A and Baldwin, B. (1998) Entity-Based Cross-Document Co-referencing Using the Vector Space Model, in Proceedings of COLING-ACL-98.

Cowie, J., L. Guthrie, T. Wakao, W. Jin, J. Pustejovsky and S. Waterman (1993) The Diderot Information Extraction System. In Proceedings of the First Conference of the Pacific Association for Computational Linguistics, (PACLING 93). Vancouver, Canada.

McKeown, K., Jordan, D., and Hatzivassiloglou, V. (1998) Generating Patient-Specific Summaries of Online Literature, In Proceedings of "Intelligent Text Summarization" (AAAI 1998 Spring Symposium Series), Stanford University, Stanford, California., pp 34-43.

McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., and Eskin, E. (1999) Towards Multidocument Summarization by Reformulation: Progress and Prospects, In Proceedings of "AAAI-99"