Two Experiments in Situated MT

Jim Cowie and Sergei Nirenburg
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
{jcowie, sergei}@crl.nmsu.edu

Abstract

More often than not, MT these days is delivered as a component of a comprehensive end-to-end NLP application. This paper presents two applications that integrate MT with other NLP processes. The first of the two combines MT with crosslingual information retrieval. The second environment uses MT, together with summarization and information extraction techniques, to generate monolingual (English) documents based on information extracted from documents in various languages. In particular, this application generates a time-stamped list of events connected to a particular person. One of the key factors in the document assembly process is the assignment of absolute dates to each sentence produced by the system. Both applications use a general purpose computational architecture that centers on an annotated document collection.

1. Introduction

The past decade saw a gradual realization that even imperfect MT results can be useful (this trend was detected early by Church and Hovy 1993). So, it is not surprising that much of the system building and deployment in the field has recently involved situating MT in end-to-end applications. In this paper, we concentrate on two specific applications. The first one is devoted to crosslingual information retrieval and the second, to information extraction from documents in a multilingual collection.

Both applications rely on a generic system architecture, see Figure 1. The crucial knowledge source in the architecture is an indexed and searchable document collection annotated in various useful ways. The high-level annotations may, in different applications, include:

- indices for dates, abbreviations, **names** of people, places and organizations mentioned in the documents,
- results of tokenization as well as morphological, syntactic and semantic **analysis** of texts;
- **summaries** of the original documents in the original and other languages,
- **translations** of these documents into other languages,
- filled **extraction templates** whose fillers are words and phrases in a natural language, formal representations ("text meaning representations") of the meaning of texts or excerpts from the texts and
- "**facts**" from an ontologically motivated fact database whose entries are instances of ontological concepts (e.g., *John F. Kennedy*, an object instance, or *2000 Sydney Olympics*, an event instance), with fillers in a fully machine-tractable metalanguage capable of supporting question answering and a variety of automatic decision-making or other reasoning processes.

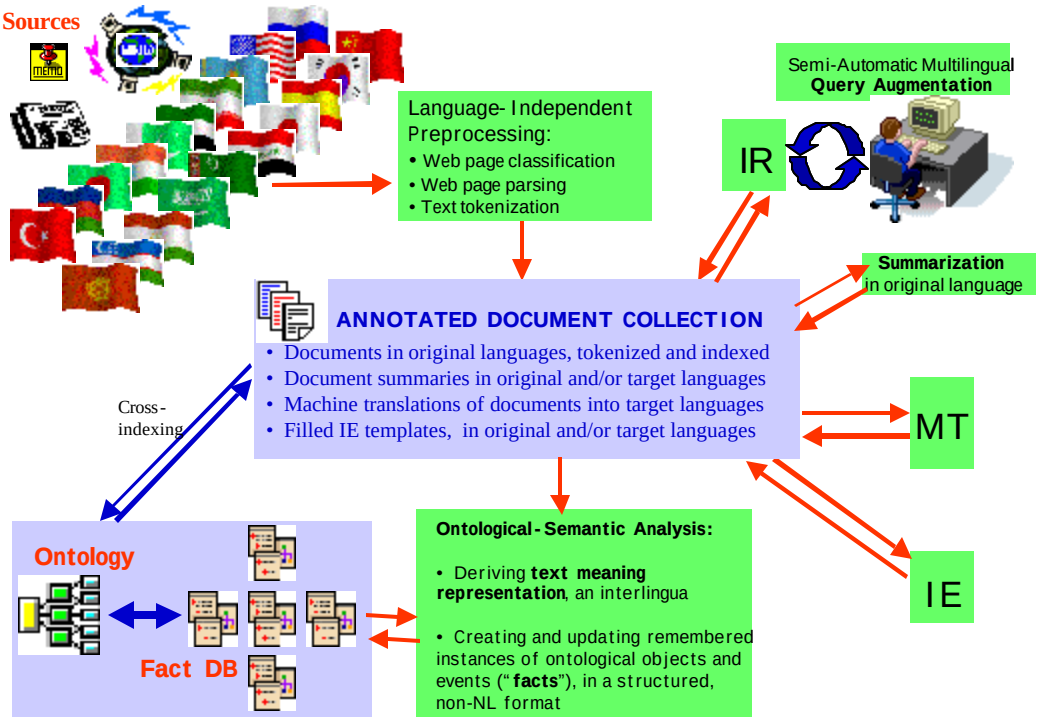

Figure 1. The architecture of an integrated text processing system. Specific end-to-end applications can be configured using a subset of the processing engines and knowledge repositories. The IR, IE, summarization and MT engines use the annotated document collection both as a source of input data and as a repository of the output from their processing. The fact database is cross-indexed with the documents in the collection. The architecture itself can also be extended: for example, additional engines, such as speech and video processing capabilities, can be incorporated. The diagram does not show the knowledge acquisition components of the architecture, environments and engines for the acquisition of the fact database or the ontology or the basic static knowledge sources for the application engines – lexicons, grammars and other rule sets.

2. Crosslingual Information Retrieval

The purpose of this application is to allow the user to manipulate documents in a language that he or she does not know or does not know well, for example, being capable of reading and understanding but not of spelling or generating reliably. A typical objective is to select a subset of documents to be processed further, for example, to be translated by humans. Of course, it would have been even better if good quality machine translation were possible. But a good quality machine translation system might not be available. So, the "second best" option is the human-computer interactive process described here that can make use of less-than-perfect translation systems.

Users can specify a query by either typing or pasting in text in English or in a source language (SL). The system then performs morphological analysis on the queries and looks up the lemmata of the query words or phrases in an SL – English dictionary, displaying the results for the user to select those word senses that are appropriate for the given query.

Indeed, the English words *bank* or *interest* can have many meanings, and a user can select an appropriate SL term. Alternatively, the system can automatically translate an English query into SL and then display all the meanings of the SL words and phrases in the query, with their translations, for user filtering (see Figure 2).

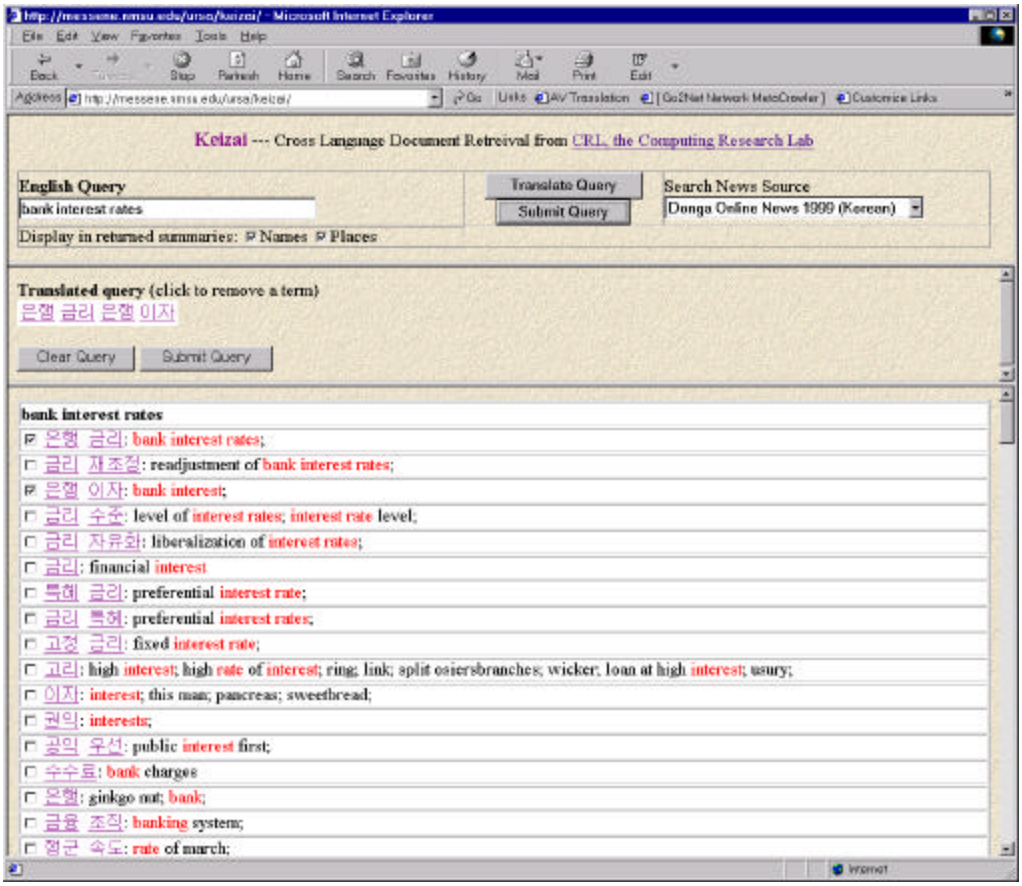

Figure 2. The user selects the appropriate senses of the words in a Korean query. The selected terms make up the final query.

Once the query is submitted, the system performs retrieval on the SL sources as well as detection of proper names, for which it uses both an onomasticon, a lexicon of proper names, and a set of language-dependent heuristics. The results of this process are displayed as a set of thumbnails in which the appearance of the key words and phrases from the query as well as of names of places and people can be highlighted (see Figure 3). Documents with the highest concentration of highlighted elements can then be selected for translation. Any translation system can be used.

In the current version of the system, we have used several simple translation systems we have developed with Japanese, Korean, Russian, Spanish, Persian and Arabic as source languages. While the quality of the output is certainly not publication quality, it can be used for judgments of whether the information in the document is of interest. Our current interface displays alternative translations of words and phrases, as found in the system lexicons (see Figure 4).

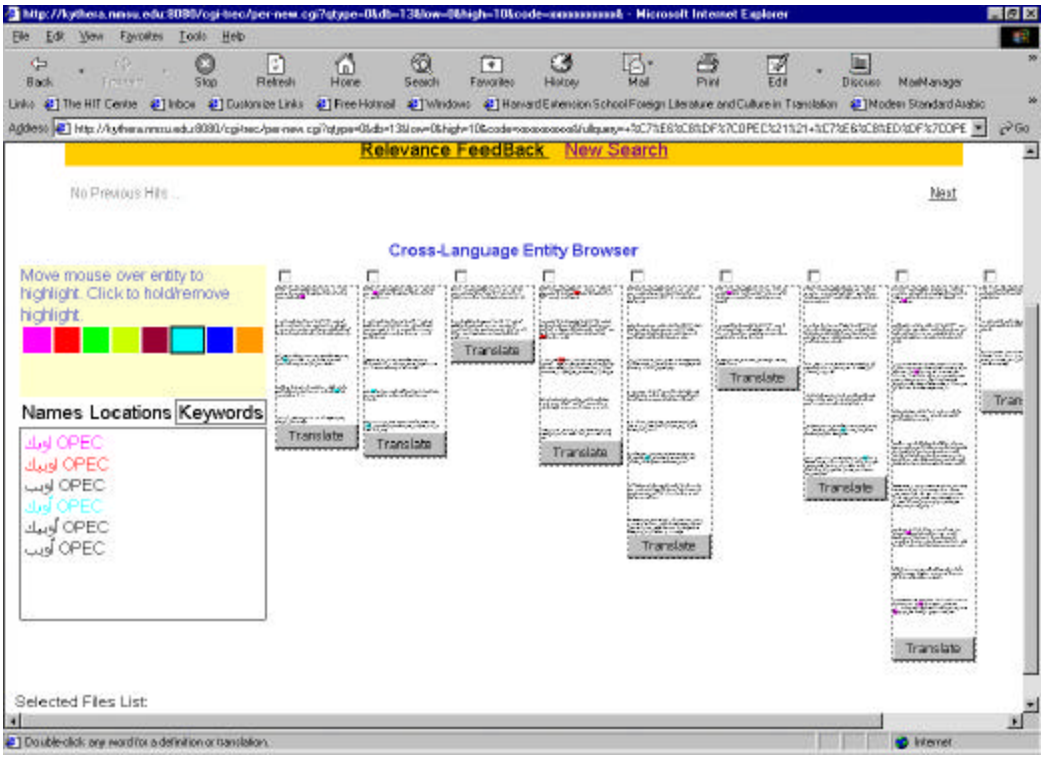

Figure 3. Occurrences of keywords in documents highlighted. This capability helps to select documents for translation.

3. Cross-Document Summarization

Cross-document summarization, producing a single summary out of a collection of documents in many languages that refer to the same topic, is another useful aid in information filtering. Our current system automatically generates information about personal profiles from multilingual documents retrieved from the Internet. This has involved the integration of multilingual tools such as automatic language recognition, generic multilingual summarization, machine translation, date recognition, to produce a system that generates "personal profiles" – information about specific personalities, organized by date. 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 5.

Several efforts to produce cross-document summaries can be found in 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).

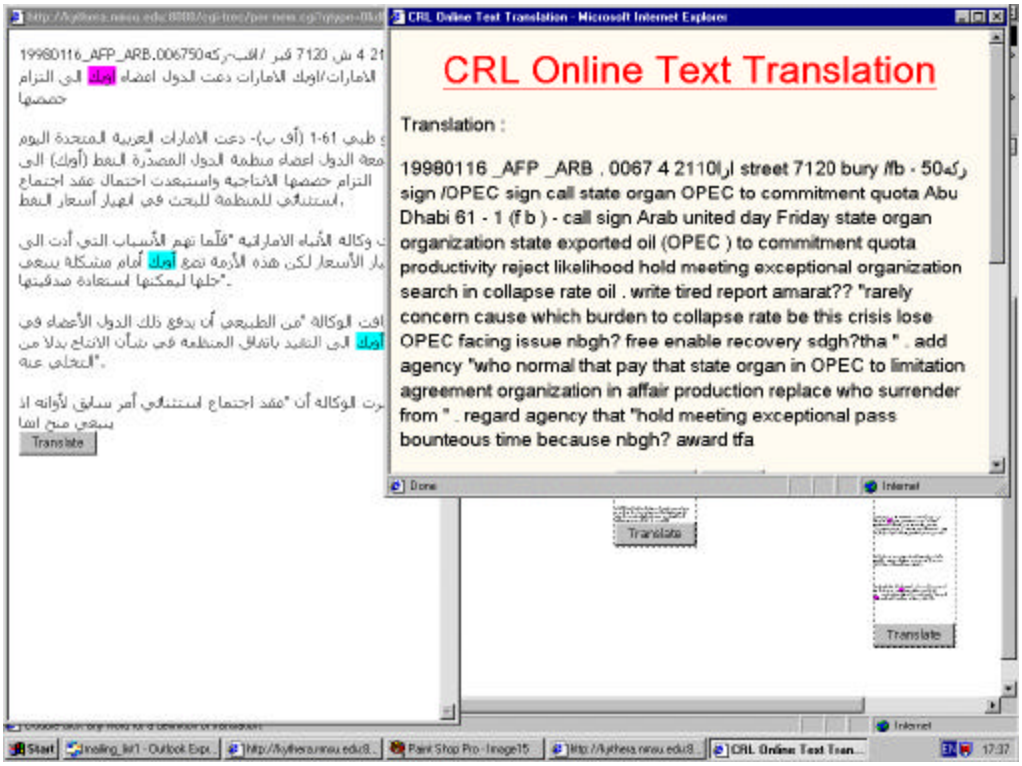

Figure 4. An Arabic text and its translation into English using a simple translation system.

The current implementation of the system takes as input a person's name in English, Spanish, and/or Russian. At present the user must supply the names (and morphological variants) in each language. 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:

1) **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.

2) **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).

3) **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.

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 (Cowie et al. 1998). An automatic language recognition tool (Ludovik et al. 1999) is used to ensure the corpus thus generated contains only documents in the three languages of interest.

The **language recognition** (LR) module also recognizes codesets which then allows a code conversion step to convert documents from any encoding to Unicode characters which is the expected encoding for language-specific tools (tokenizers and machine translation engines). The LR module 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.

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.

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) **Text extraction from HTML files**, to eliminate markup tags. This task presents a challenge, due to the complex layout used by different web sites that include the use of frames, tables and dynamic html. A special parser for HTML text has been adapted for this purpose (Ragget 2000).
   Algorithms for summarization and translation typically act on 'flat' text. The module for zoning and parsing web documents:- finds textual content, recognizes and analyze frames with reference to their relevance to the meaning of the document and includes the content of hyperlinked "source" pages

b) Multilingual paragraph, sentence and word **tokenization** to get the structure of the document. This stage is important for processing documents written in different languages.

a) **Date stamp determination** for the documents being summarized, using a multilingual date recognition package.

b) **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-oriented 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 reconstructed from the date previously determined for the document.
c) **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. Any available MT engine could be used.

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.

3.1 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 have not yet been extended 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 relies on a simple grammar that uses 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 the 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*, will be represented 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, we are using very simple 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. 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 temporal co-reference, which may be affected by genre, topic and possibly other factors would be very useful for developing our work further.

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. 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, the former UK Foreign Secretary, was born, educated, and first elected. Later Spanish sources provide more regional information related to 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.

There are many problems concerning cross-document summarization that need to be treated to improve a system, such as the possibility of finding relevant sentences from different documents that contradict each other or sentences that refer to the same events, therefore causing repetition in the final cross document summary. 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.

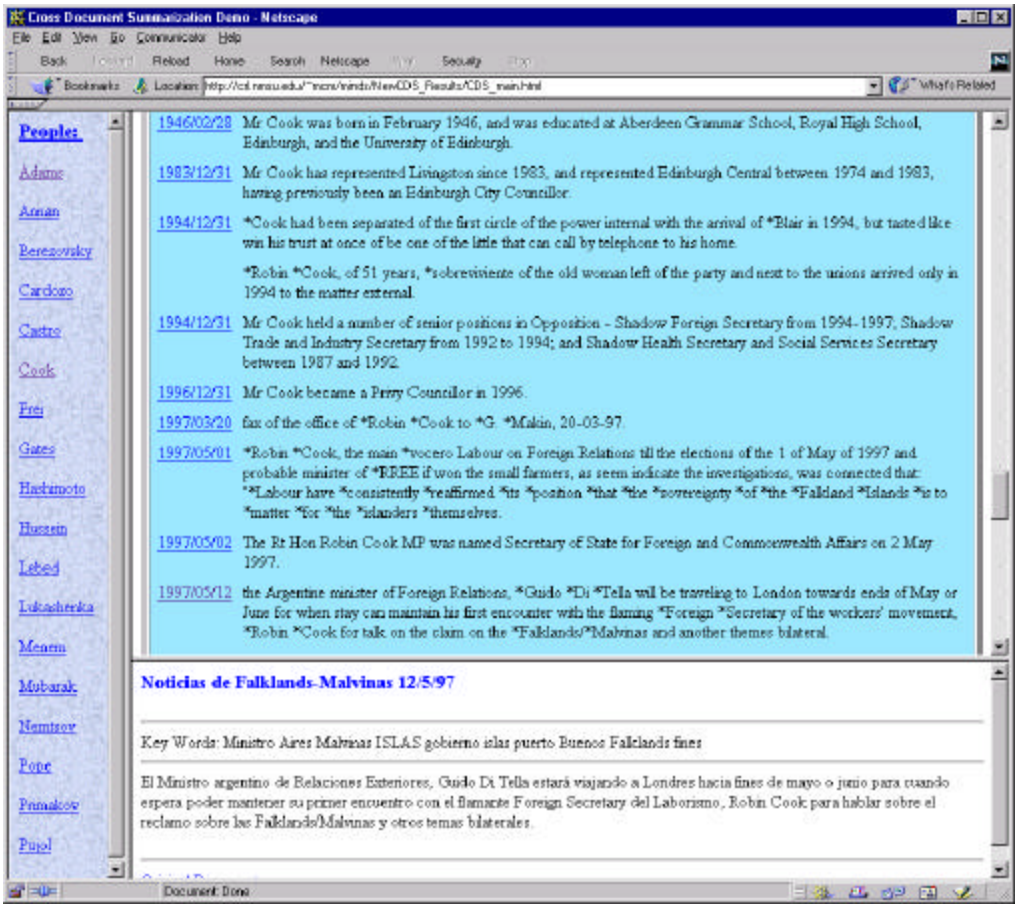

Figure 5. A personal profile of Robin Cook, former UK Foreign Secretary. The asterisks in the text are words transliterated by the translation systems. Our MT system resorts to transliteration if it fails to provide a real translation. The dates provide links to the summaries shown in the lower frame, and this in turn contains a link to the original document. The text shown here is from a Spanish source.

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 two or more people in the

news with the same names, we can (and do) end up with a ambiguous set of documents. In our sample we had to deal with a concert pianist and a politician both named Boris Berezovsky. 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. A second approach is to let the user see translated document summaries of the retrieved documents and select those he wishes to pass to the profile generator. An interface has been developed which allows this human filtering to be carried out rapidly.

Acknowledgements

Our thanks to our colleagues at CRL, who have developed many of the components described here: Ahmed Abdelali, Mark Davis, Mark Leisher, Yevgeniy Lyudovyk, Bill Ogden, Hugo Molina Salgado, Svetlana Sheremetyeva. and Ron Zacharski.

References

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

Church, K and E. Hovy. (1993). Good Applications for Crummy Machine Translation. *Machine Translation*, 8.

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.

Cowie, J., E. Ludovik and R. Zacharski. (1998) An Autonomous, Web-based, Multilingual Corpus Collection Tool. In *Proceedings of the Natural Language Processing and Industrial Applications*, Moncton, Canada.

Ludovik, E., R. Zacharski, and J. Cowie (1999) Language recognition for mono- and multi-lingual documents. *VEXTAL (Venezia per il Trattamento Automatico dell Lingue)*, Venice, 22 - 24 November

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*.

Ragget, D. (2000) Jtidy HTML Parser, http://www3.sympatico.ca/ac.quick/jtidy.html