

# Keizai: An Interactive Cross-Language Text Retrieval System

William Ogden, James Cowie, Mark Davis, Eugene Ludovik, Sergei Nirenburg, Hugo Molina-Salgado, and Nigel Sharples

Computing Research Lab  
New Mexico State University  
Las Cruces, NM 88001  
USA

## Abstract

Can we expect people to be able to get information from texts in languages they cannot read? In this paper we review two relevant lines of research bearing on this question and will show how our results are being used in the design of a new Web interface for cross-language text retrieval. One line of research, “Interactive IR”, is concerned with the user interface issues for information retrieval systems such as how best to display the results of a text search. We review our current research, on “document thumbnail” visualizations, and discuss current Web conventions, practices and folklore. The other area of research, “Cross-Language Text Retrieval”, is concerned with the design of automatic techniques, including Machine Translation, to retrieve texts in languages other than the language of the query. We review work we have done concerning query translation and multilingual text summarization. We then describe how these results are being applied and extended in the design a new demonstration interface, *Keizai*, an end-to-end, Web-based, cross-language text retrieval system.

## 1 Introduction

A Cross-Language Text Retrieval (CLTR) system is designed to retrieve text documents in a language different from the language used to specify the information needed.

Several compelling research questions have emerged from our work and that of others on CLTR systems. Among these has been the basic question of whether machine translation technology in the broadest sense is required to translate queries and documents, or whether alternative techniques can substitute for machine translation in many circumstances. The latter question becomes particularly important for

languages for which there is little support from commercial machine translation vendors, especially languages that are spoken by small numbers of people. We provide an overview of our work in cross-language technology and show how research has begun to answer some of these questions in recent years, while bringing about new, additional questions at the same time.

Another important aspect of our research has been focused on trying to understand how to create interfaces and systems that are useful to people. This has included iterative development of systems like our current prototype, *Keizai*, as well as empirical user studies evaluating retrieval results visualization interfaces for monolingual and cross-language text retrieval systems.

When conducting empirical user studies, we have found that careful consideration of how users process the task sheds enormous light on developing effective user interfaces. There are at least a couple of ways a CLTR system may be used. A bilingual user who has good reading skills in their second language may have poorer language productive skills and thus cannot express their information need in their second language as well as they can in their first language. Thus, the CLTR system allows them to find documents in their second language using their first language. A second type of user is a person who is monolingual but has an interest in finding information in documents that are written in foreign languages. They may have access to translation resources but want to limit their use to control costs. Thus, they want to be able to evaluate the relevance of a document to their query before committing resources for a full translation.

Two elements of the CLTR system interface could benefit the two types of users identified above. The first element, the cross-language query formulation interface, should help the user construct good retrieval requests in the target foreign language. The

second element, the retrieval results display, should help users judge the relevance of the retrieved documents set.

## 2 Research in Cross-language Text Retrieval

Although it is possible to translate all of the documents into the query language, for large collections the most economical approach to CLTR is to simply translate the query at retrieval time into the document languages. This presupposes that the query can be translated in a reasonably accurate fashion and that monolingual retrieval systems are available for all of the document languages.<sup>1</sup>

As with Machine Translation (MT) in general, query translation in a CLTR system can be done many different ways. An advanced MT system might, for example, perform sophisticated parsing and analysis of the query, derive an inter-lingual semantic representation and generate a new query from it. At the opposite end of the spectrum, a system could use shallow translation techniques to simply substitute terms from a transfer dictionary, ignoring the ambiguities of polysemous candidates. In shallow translation approaches, the monolingual text retrieval engine operating on the translated query bears the burden of weighting the query terms by virtue of their co-occurrences, hopefully reducing the effect of poorly translated terms. Between these two extremes are a range of approaches.

To overcome the limitations of general-purpose transfer dictionaries, Salton (1971) used tuned lexicons and thesauri built from controlled vocabulary to good success in specific text retrieval problems. Despite the growing availability of machine-readable dictionaries, however, preparing special-purpose lexical resources remains a daunting task.

Using massive bilingual and multi-lingual corpora as translation resources is another approach that has the

---

<sup>1</sup> To conduct our experiments in multilingual and Cross-language text retrieval we developed the Unicode Retrieval System Architecture (URSA) as a core retrieval technology. URSA is a high-performance text retrieval system that can index and retrieve Unicode text. URSA provides a C library for application development, and a group of stand alone tools that can process documents, queries and information requests in any of the scores of languages that can be represented as Unicode text. For more information see <http://crl.nmsu.edu/Research/Projects/tipster/ursa>.

potential to overcome the limitations of the shallow methods, while still requiring less resources than the tuned lexical methods or the deep semantic MT approaches. Text corpora contain examples of usage patterns in the query language that can be matched to examples in the target language if the sentences or paragraphs of the texts are aligned to one another. Although text corpora offer an intriguing possibility for CLTR query translation, the lack of domain-specific texts or a suitably large range of texts means that general purpose query translation systems remain elusive.

For example, one method of using bilingual corpora, automatically constructs a multilingual semantic space using Lexical Semantic Indexing (LSI). This approach does very well within domains similar to the domain of the bilingual corpora used to train the system but do considerable worse when searching outside of these domains (Dumais, 1997).

Several other researches have explored methods to improve the CLTR performance when using general bilingual lexicons with automatic disambiguation techniques. For example, Ballesteros and Croft (1997) improved CLTR performance using query expansion through automatic relevance feedback. Their best performance, however, was still 32% less than a monolingual baseline.

In our own work, (Davis & Dunning 1996, Davis 1996, Davis & Ogden, 1997) we have used combinations of shallow dictionary-based and corpus-based method. We use the corpora to derive query translations directly by electing terminology from the target language portion of the corpus. We have also used bilingual corpora to try to optimize translated queries by eliminating ambiguous terms incrementally from the target query according to the similarity of the query and target query's retrieval characteristics. In the best cases, we can approach 80-90% of the performance of a monolingual baseline system.

To briefly summarize the work on CLTR to date, we feel that minimal bilingual dictionaries are useful and plentiful but the quality of these resources make a big difference and vary considerably. The problem of selecting the right translation term can be partially overcome by automatic techniques including some corpora based methods. Parallel corpora are hard to come by, however, and produce noisy results outside of the subject domain of the parallel texts. We believe that the CLTR user can play an important role in aiding the disambiguation process through an interactive user interface. We have begun to explore interactive methods for query formation that could help solve some of these problems and describe this work below.

### 3 Interactive CLTR

In an important study, Resnick (1997) showed how users might be able to make judgements about information contained in documents in languages they cannot read. People who knew no Japanese, were asked to sort Yellow Page entries that had been translated from Japanese to English using a very simple word-for-word “gist” method. The sorting was compared with sorts done with same entries in original English and with a random sort. The “gist” sorts were more consistent with the original English than with the random sorts indicating that people were able to glean useful information from simple translations. This result gives hope for designing CLTR systems that can be used by the monolingual person. In this section, we evaluate a system to meet these needs.

In an effort to understand how CLTR systems might be improved, we have recently conducted some preliminary experiments on an interactive, cross-language text retrieval system. The system, *Arctos*, provides a user with a browser-based interface with which to enter English queries. After an initial query is entered, the query is translated using a simple word-for-word or phrasal translator. The user can then interactively improve the query translation using links to on-line bilingual translation resources and then submit the query for retrieval against document collections in the target language. The retrieved documents are presented using document thumbnails and query term highlighting. Further, the user can have the returned document translated from the target language to English by the Babelfish translation engine from Systran and made available by Alta Vista.<sup>2</sup>

The interactive task in this preliminary study was for one user to use English TREC CLTR track topics to retrieve and judge the relevance of German documents. The user, who judged himself to have no German language knowledge, formed his own English query based on the TREC topic statement in English. He then modified and improved the German query while examining documents and using the on-line dictionary resources, submitted the modified query to the URSA<sup>1</sup> engine for retrieval, examined the retrieved documents using the German equivalents and document thumbnail interface, submitted the documents to the Babelfish translation engine to translate to English, and judged the top 10 documents retrieved as either relevant or non-relevant. No time limit was set and the user spent approximately 8 hours working on 22 TREC topics.

		User Judgements	
		relevant	not relevant
NIST Judgements	relevant	69	1
	not relevant	11	43

Figure 1. Number of judged documents in the *Arctos* study

The results are shown in Figure 1 and consist of a comparison of the relevance judgements made by the user in the study to the “correct” judgements provided for the TREC-6 Cross-Language evaluation track. Documents were excluded if this system retrieved them but they were not in the TREC judgments for these queries. From these numbers, we summed the counts and calculated a False Hit Ratio, that is, the ratio of the number of documents that were irrelevant but judged relevant to the number relevant and judged so. The false hit ratio for this experiment is only 15.9%. We further calculated the False Drop Ratio at 2.32%, which indicates the relative number of times a relevant document was incorrectly judged irrelevant to the number of times an irrelevant document was also judged irrelevant. The combined performance figures for this system indicate a very low percentage chance of error in using this cross-language retrieval system.

What led to the relatively good performance by this user who could search for and identify relevant German documents with no German language knowledge? The user reported using a very wide range of techniques to modify queries, and reported particular difficulties with generating good phrasal equivalents in German, where compound nouns are often extremely important as query elements. The primary resource used to improve and evaluate queries was an on-line bilingual dictionary which was used to “back-translate” query terms selected by the URSA query translation process. The dictionary interface listed all matching entries with their English definitions and the user could easily pick correct query terms to expand the search. This process of user-aided query expansion has been streamlined in the *Keizai* prototype described below. Whereas the *Arctos* interface required cutting and pasting from the dictionary pages to the query entry pages, the *Keizai* interface integrates the two processes.

<sup>2</sup> [URL] <http://babelfish.altavista.com>

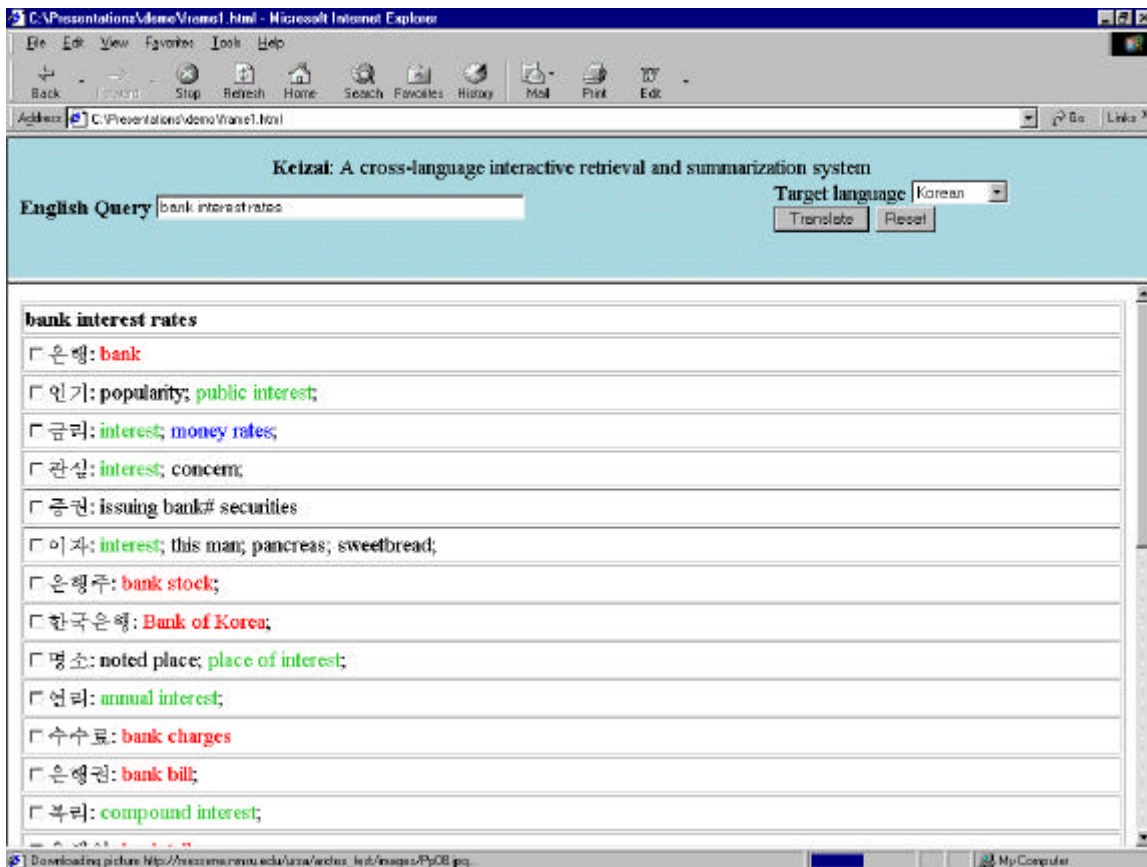


Figure 2 Keizai query term selection

#### 4 The Keizai CLTR system prototype

The Keizai project is focused on providing a demonstration of an end-to-end Web-based cross-language text retrieval system. Beginning with an English query, the system will search Japanese and Korean Web data and display English summaries of the top ranking documents. User should be able to accurately judge which foreign language documents are relevant to their query. The system currently searches archived Web news in Japanese from the Asahi Shimbun news service and in Korean from the Dong-A Ilbo.

The prototype system has undergone extensive revision and iterative design. The design of the query interface and query translation module has been improved through several design reviews and usability walkthroughs. As a result of these and our experience with Arctos and other CLTR research, the test system now implements a new approach to query translation that presents extended English definitions of query terms and phrases alongside their Japanese or Korean translations. The user can select the English definition that most accurately reflects their intention in the original query. Query term disambiguation is handled by user intervention without the need for morphological analysis and segmentation. The correct surface forms of the

Korean or Japanese query terms are selected by a user who knows no Korean or Japanese. At the same time, the query translation module has been improved to only select terms that actually occur in the target data.

Figure 2 shows the current design. In this example, the user has entered an English query, “bank interest rates”. The system has selected Korean query terms from a Korean-English lexicon that could translate to one or more of the query terms. A pre-search eliminates terms that will not return any documents, and then the terms are sorted by frequency. The user can simply select those terms whose definitions are consistent with their information need.

Similar progress has been made in the WWW-based presentation interface. (See Figure 3). Dynamic HTML capability is being used to quickly and compactly display document summaries and the distribution of keywords in documents are displayed in both the original and English forms. Query term occurrences for each document are being represented. Documents are being summarized in the original language, Japanese and Korean, by MINDS<sup>3</sup>, an internally developed summarization engine. Summarization in MINDS is based on a

<sup>3</sup> Multilingual Interactive Document Summarization [URL] <http://crl.nmsu.edu/Research/Projects/minds>

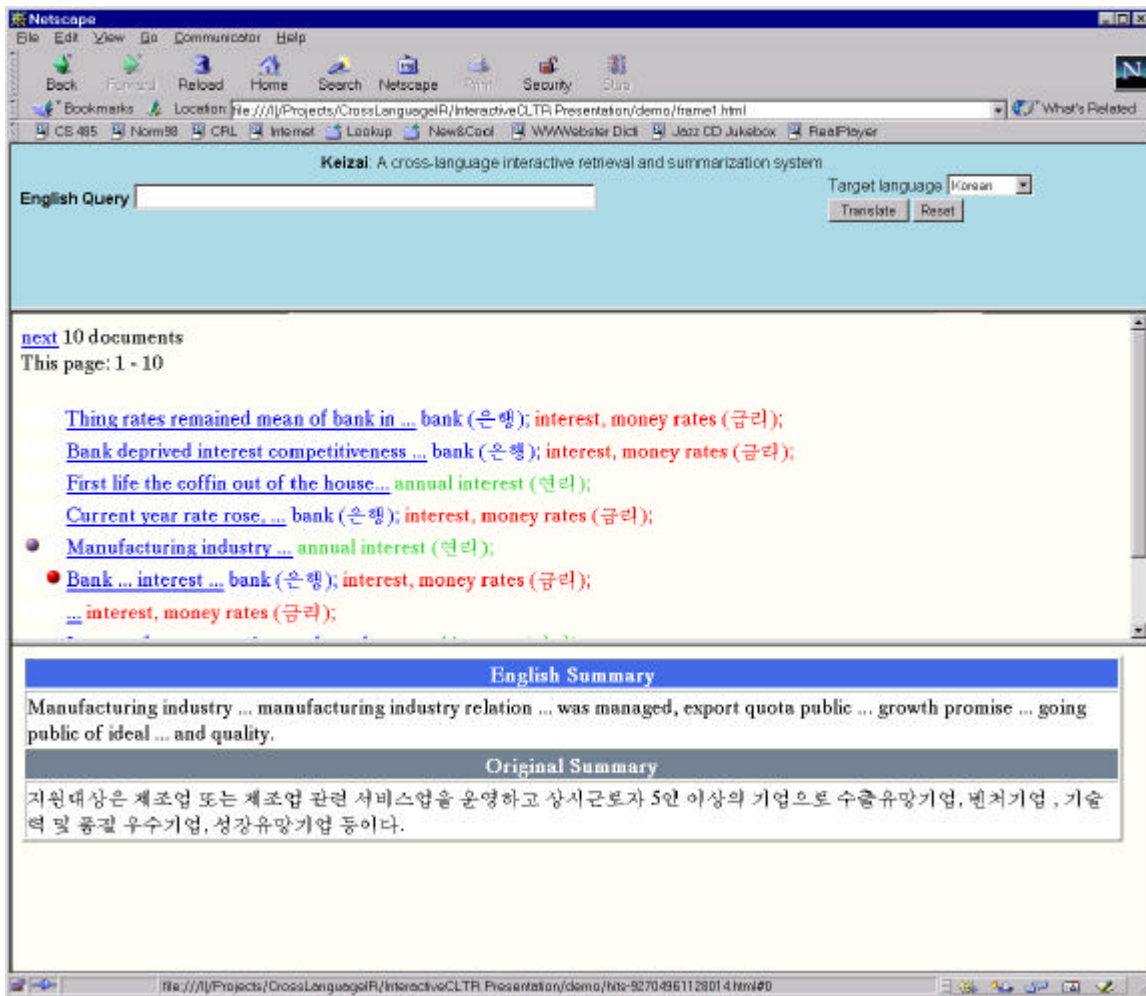


Figure 3. Keizai results display

variety of statistical and symbolic techniques and is parameterized to produce a variety of summary types for different goals. In this case, query specific summaries are produced which focus on passages containing query terms. The original summaries are then translated to English using the machine translation facilities for Japanese and Korean provided by our Corelli<sup>4</sup> MT architecture. Documents are sorted by a relevance measure and are represented by the first line of the summary for each retrieved document which is displayed in the top portion of the results display along with the color coded query terms occurring in the document. Both the original and the English summaries are dynamically displayed in the bottom portion of the display for each document when the mouse pointer is positioned over the document line giving users a very quick way of evaluating document relevance.

The *Keizai* demonstration interface brings together many of the findings of our work on CLTR and Interactive IR. However, there are a number of

research issues that we have yet to test with the system. Of primary importance is the quality of the translated summaries. We need to do more work here and we plan to evaluate the effectiveness of the *Keizai* with test users in the near future.

## 5 Related Work

There are a number of similar CLTR efforts. Perhaps the most similar is the MULINEX project<sup>5</sup> (Erbach Neumann, and Uszkoreit, 1997). The current demonstration (July, 99) is a cross-language retrieval system in English, German, and French that contains a “query assistance” feature, which allows users to pick alternative translations for query terms in much the same way as the *Keizai* system does. It also provides translations of document summaries but, unlike *Keizai*, does not include information about query term distribution in the returned documents.

<sup>4</sup> [URL] <http://crl/nmsu.edu/Research/Projects/corelli>

<sup>5</sup> [URL] <http://mulinex.dfki.de/index.html>

## 6 References

Ballesteros, L., and Croft, B., (1997). "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval", in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, July 1997, pp. 84-91. Earlier work by the same authors is available at <http://ciir.cs.umass.edu/info/psfiles/irpubs/ir.html>

Erbach, G., Neumann, G., and Uszkoreit, H., (1997). MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web", in *Cross-Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05. Available at <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

Davis, Mark. (1996). "New experiments in cross-language text retrieval at NMSU's Computing Research Lab". In Harman, D. K., ed., *The Fifth Text REtrieval Conference (TREC-5)*. NIST. 1996. Available at <http://trec.nist.gov/pubs.html>.

Davis, Mark, and Dunning, Ted, (1996). "A TREC evaluation of query translation methods for multilingual text retrieval", in *Proceedings of the 4th Text Retrieval Conference (TREC-4)*. 1996. Available at <http://trec.nist.gov/pubs.html>.

Davis, M. and Ogden, W. C., (1977). "QUILT: Implementing a large-scale Cross-Language Text Retrieval System." In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, July, pp. 92-98.

Dumais, S. T., Letsche, T. A., Littman, M. L., and Landauer, T. K., (1977). "Automatic Cross-Language Retrieval Using Latent Semantic Indexing," in *Cross-Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05. Available at <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

Resnik, P. (1997). "Evaluating Multilingual Gisting of Web Pages", in *Cross-Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05. Available at <http://www.clis.umd.edu/dlrg/filter/sss/papers/> .

Salton, G. (1971) *Automatic Processing of Foreign Language Documents*, Prentice-Hall, Englewood Cliffs, NJ.