UCLIR: a Multilingual Information Retrieval Tool

Ahmed Abdelali, James Cowie, David Farwell and William Ogden

Computing Research Laboratory
Box 30001/3CRL
New Mexico State University
Las Cruces, New Mexico 88003
USA
{ahmed,jcowie,david,ogden}@crl.nmsu.edu

Abstract

In this paper the Unicode Cross-Language Information Retrieval system (UCLIR) is described. UCLIR accepts a query in one language and then retrieves relevant documents in any of several languages. The core process involves a suite of technologies including machine translation and standard monolingual information retrieval.

Key words: Cross-language Information Retrieval, Human-Computer Interaction, Machine Translation, Unicode.

1. Introduction

The problem of multilingual information retrieval is an extension of the general problem of monolingual information retrieval (IR). The goal of IR is to retrieve documents that most directly respond to a user's requests. Given the speed of access and the large scale of the information sources available today, users often wish to reach beyond a single information source in looking for relevant answers to their queries. The tool presented here, the Unicode Cross-language Information Retrieval system (UCLIR) focuses on providing users with the capability to exploring multiple multilingual resources using a range of different approaches. It does this by scaling up facilities developed for monolingual IR to include bilingual dictionary look up, multilingual search and the standardized encoding of different languages and their character sets through the use of Unicode.

The UCLIR retrieval system is based on URSA [Davis & Dunning 1995, Ogden & Davis 1998, Ogden, *et al*.1999, Ogden & Davis 2000], a multilingual IR system developed previously at the Computing Research Laboratory at New Mexico State University.

2. Unicode

Computers fundamentally deal only with numbers. They store letters and other characters by assigning a number to each distinct item. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. However, no single encoding system could represent sufficiently many characters as the number and diversity of languages grew. For example, the European Union alone requires several different encoding systems to cover all its languages. Even for a single language like English no single encoding was adequate for all the letters, punctuation, and technical symbols in common use.

In addition, these encoding systems often conflict with one another. That is, two different encoding systems might use the same number for two different characters or use different numbers for the same character. As a result, any given computer (in particular any server) needs to support many different encoding systems. Even so, whenever data is passed between different encodings or platforms, there is always the risk that that data may be corrupted.

Unicode provides a unique number for every character, no matter what the language, no matter what the platform, no matter what the program. Such industry leaders as Apple, HP, IBM, Microsoft, Oracle, SAP, Sun, Sybase, Unisys and many others have adopted the Unicode Standard. Unicode is required by modern standards such as XML, Java, ECMAScript (JavaScript), LDAP, CORBA 3.0, WML, etc., and is the official encoding for implementing ISO/IEC 10646. It is supported in many operating systems, all modern browsers and many other products. The emergence of the Unicode Standard and the availability of tools supporting it are among the most significant recent global trends in software technology [Unicode 2002].

3. URSA

URSA, the Unicode Retrieval System Architecture, is a high-performance text retrieval system that can index and retrieve Unicode texts. As result, URSA has the capacity to index and retrieve documents in every language that can be represented in UNICODE (almost 1000 languages are already included in the UNICODE standard) and provides tools for converting texts into UNICODE and back again into the original encoding. URSA also has a comprehensive set of query and document weighting functions commonly used for information retrieval. The complete suite of weighting and ranking functions implemented in URSA represents the bulk of the weighting schemes developed in the past 40 years of text retrieval research and includes many of the recent successful document weighting schemes from Cornell [Buckley, *et a*l. 1998] and City University of New York[Kwok *et al*., 1998]. Further, by using a posting compression scheme that is both simple enough to allow for the efficient merging of posting data as well as for its rapid decompression and yet is specifically tuned to the kinds of data in the postings, URSA indexes are only about 12%-25% of the size of the original texts. Finally, the URSA tools are robust enough to be used in industrial-grade applications and are based on a very simple object-oriented API [Ogden, *et al*. 1999].

4. System Description

4.1 Retrieval

The system operates in any of three different modes: using a multilingual query, using an English query without user involvement in the formulation of the multilingual queries or using an English query with user involvement in the formulation of the multilingual queries.

4.1.1. Multilingual query

The multilingual query mode is an extension of the CRL's monolingual approach to IR query that essentially allow users to access resources in other languages by formulating queries in other languages. The interface (see Figure 1 below) will accept a Unicode text containing terms in various languages regardless of the format or form of the query. The documents retrieved against the query will in general be in the same languages as the query terms. The most relevant documents retrieved may end up being in one language, reflecting the likelihood of co-occurrence of the terms of the original query. The relevance of the documents retrieved is computed globally over the entire multilingual resource. Even in the case of cross-language homographs (for instance, up to 40% of the words in Englsh are of Latin or French origin), retrieval would be only slightly affected because the co-occurring words of the query will cause the system to give any documents containing the noisy word a lower score. In another words, even though some of the documents may be mistakenly retrieved on the basis of the accidental homograph, they will be ranked lower than those retrieved on the basis of the intended query for each language.

4.1.2. English query: non-interactive approach

The second mode relies on a set of bilingual dictionaries for translating an English query into the different target languages. Once the set of translations has been compiled, they are validated against the index word list of the resources and all the terms that are not in the resources are dropped from the query. The resulting query is then used to retrieve the relevant documents from the resources and returned to the user (see Figure 2 below). With this approach the user has no control over the multilingual query formulation process. Rather success depends on the quality of the bilingual

dictionaries and the size of the index word list of the resources.

4.1.3. English query: interactive approach

Using this third mode, the user becomes involved in making decisions about the terms selected for the

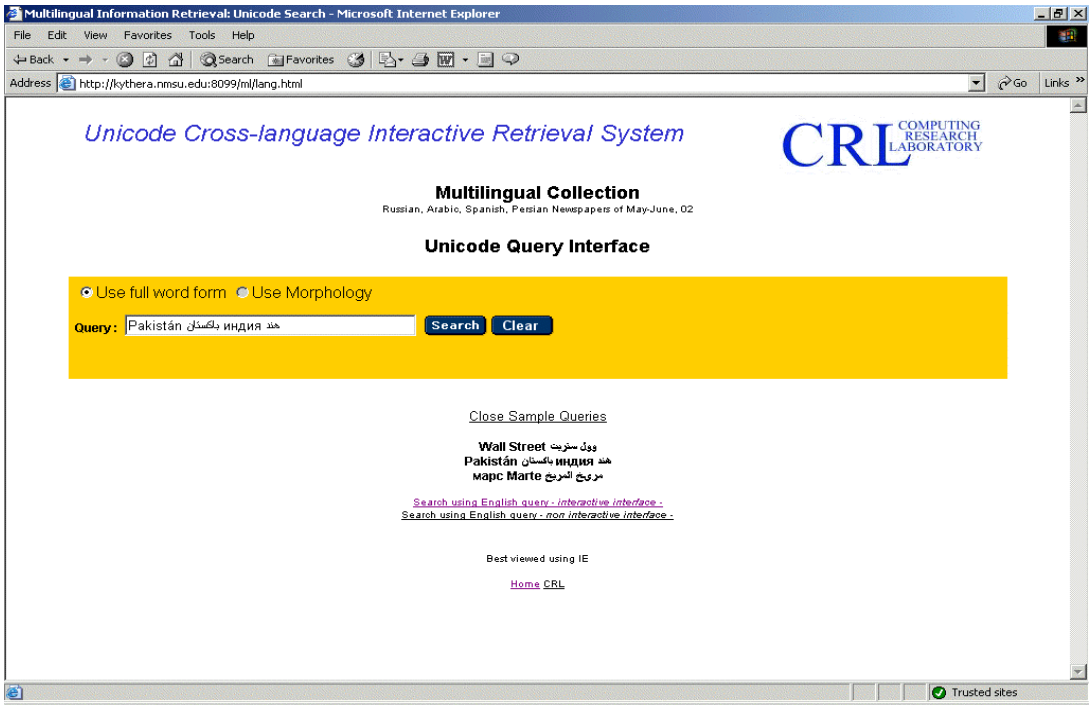

Figure 1. Main interface for Unicode query

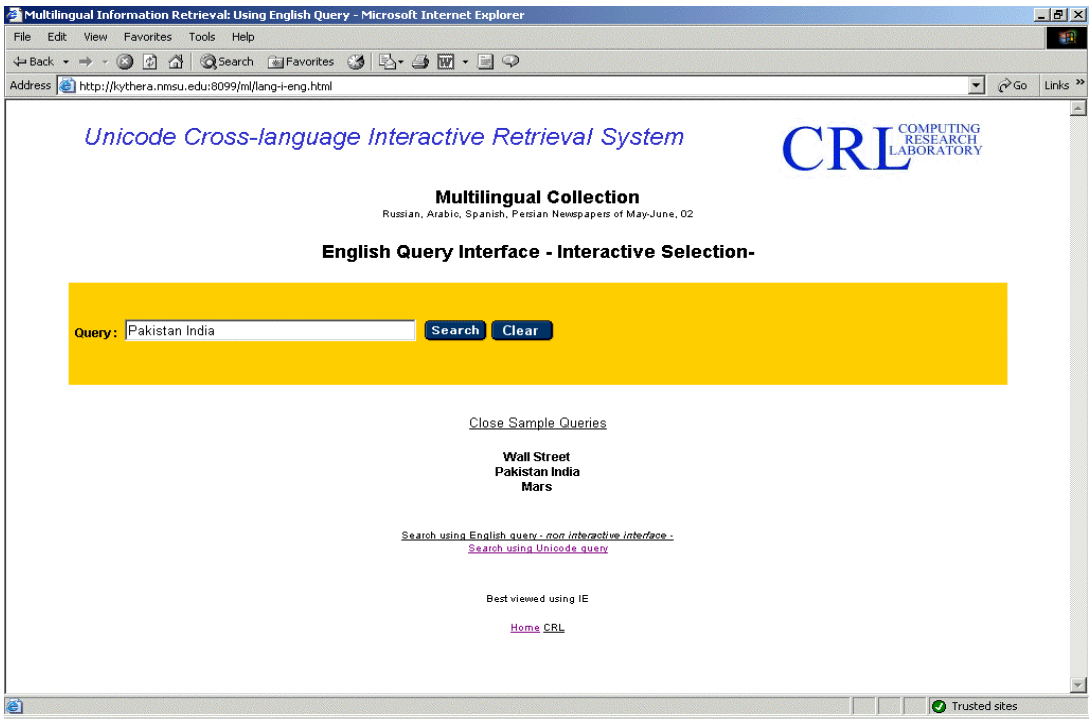

Figure 2. English query interface

various queries in the different languages. Knowledge of the other language is not required. First, the English query is passed to the bilingual dictionaries which, in turn, produce translations for each query term. The translations are checked against the index word list and only those terms that have a match are kept for further processing. The interface presents these terms to the user along with details about their English correspondences and other information (i.e., part of speech, domain, and so on). It also presents visual distinctions among the terms of the different languages (see Figure 3 below). The user is then responsible for choosing the most appropriate terms, based on their meaning, to use for the multilingual queries. Once a query ready, it is used to retrieve documents from the indexed resources and presented to the user.

4.2 Entities Selection

The options available for viewing the retrieved documents include the highlighting of the proper names (e.g., of people, organizations and locations) in the text, as well as of the query terms (see Figure 4 on next page for a general view and Figure 5 on next page for the pop-up view of a particular text). These functionalities allows users to get more information about the text in terms of co-occurrence and other relations between the entities mentioned in the text.

4.3. Translation

The documents retrieved in the different languages may be translated into English. This is a useful tool for checking on and evaluating the success of the cross-language retrieval process. To carry out the translation there are two alternative techniques: word-level translation and document-level translation.

4.3.1. Word-level translation

This option allows the user to see the translation of any single word in the document by clicking on it (see Figure 6 on page N). The translation provided includes lexical information about the word.

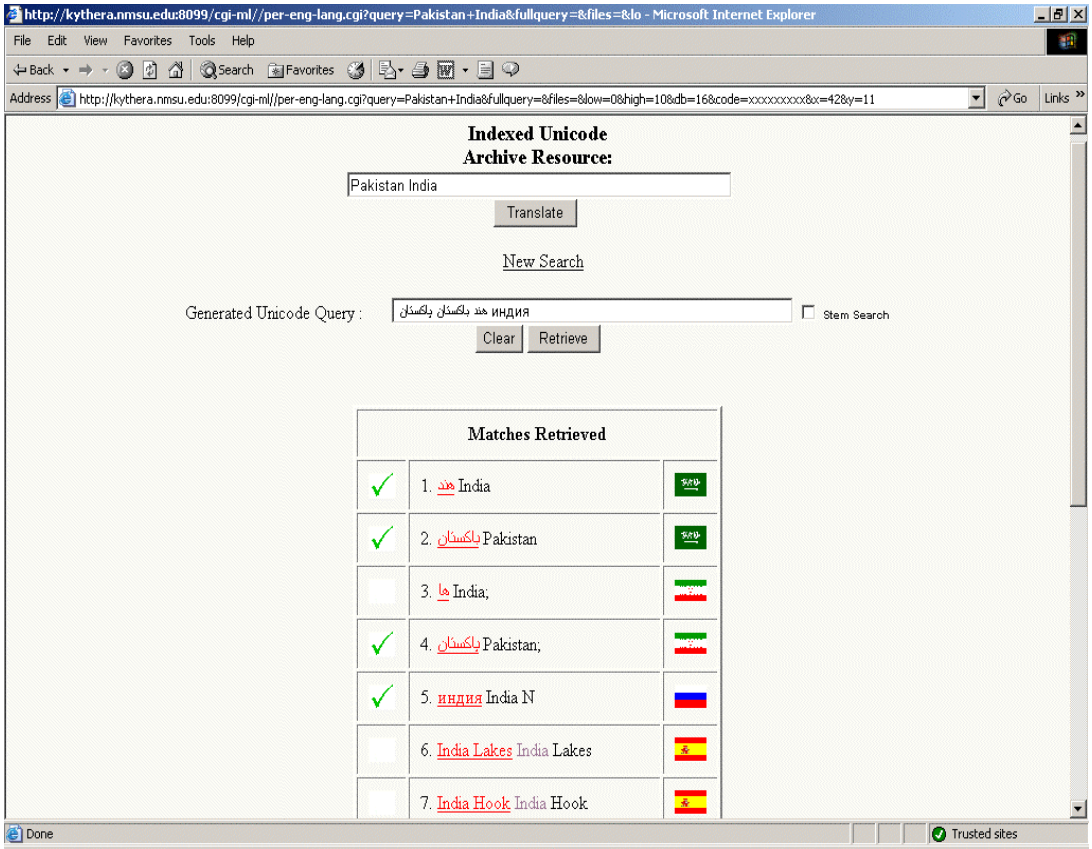

Figure 3. Interactive multilingual query building

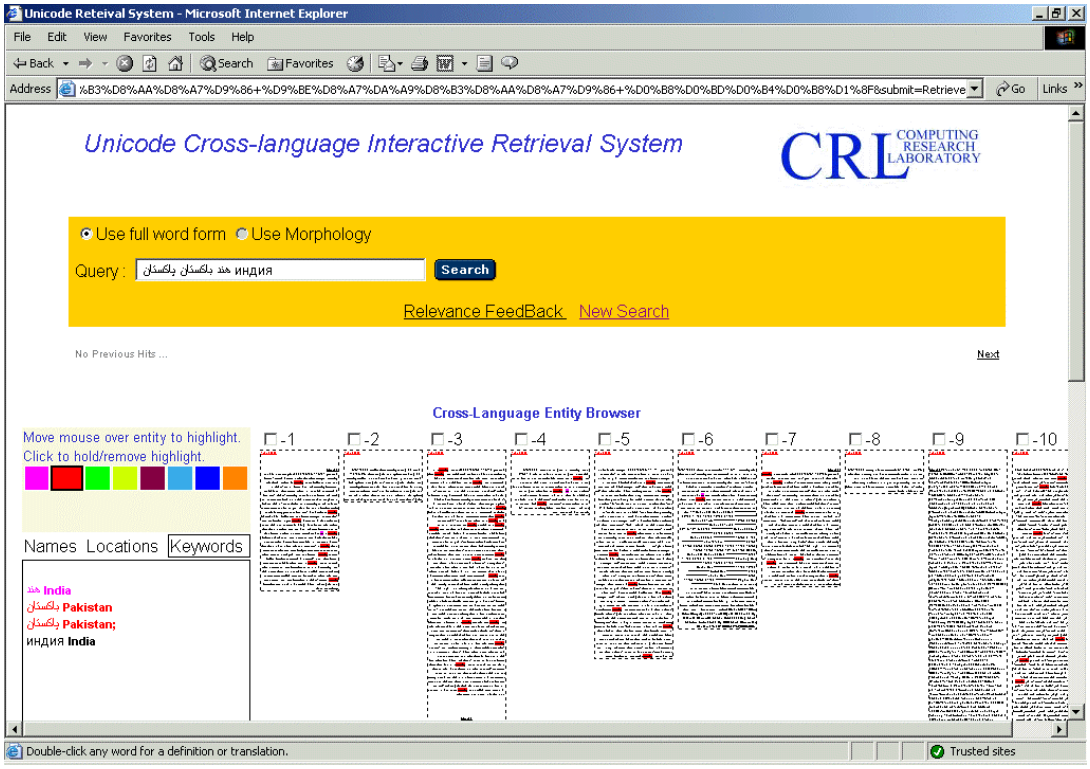

Figure 4. Retrieved document and entity selection

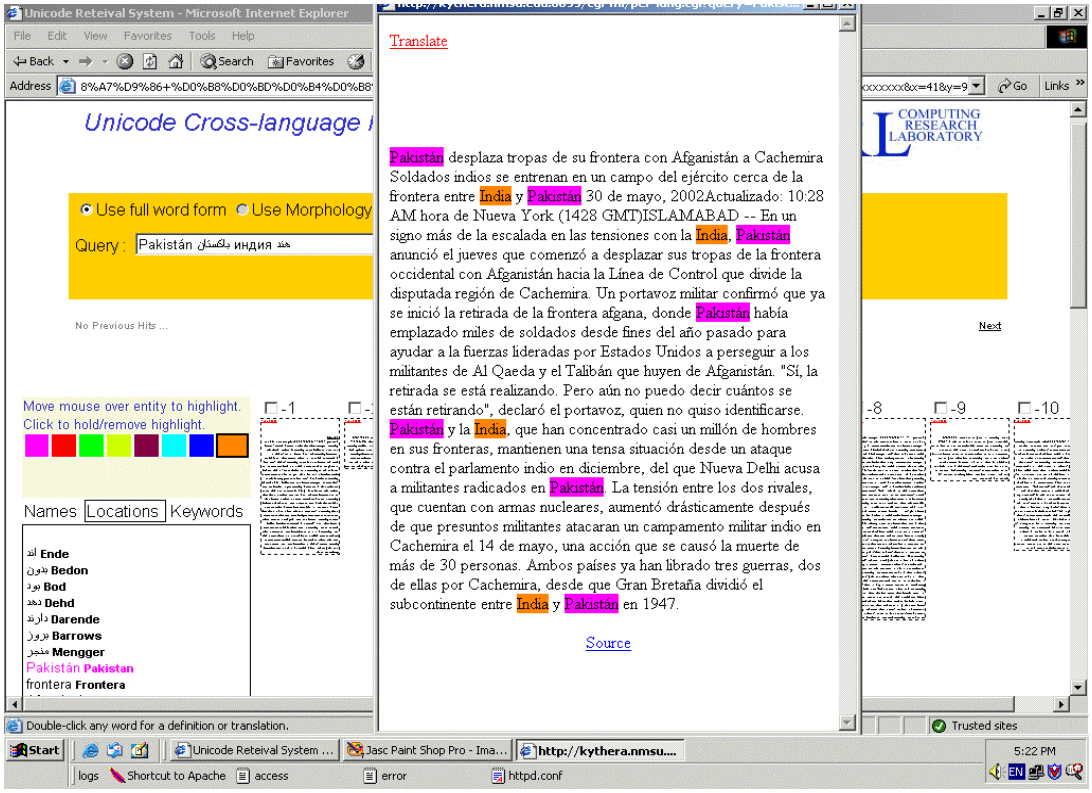

Figure 5. Retrieved document viewing

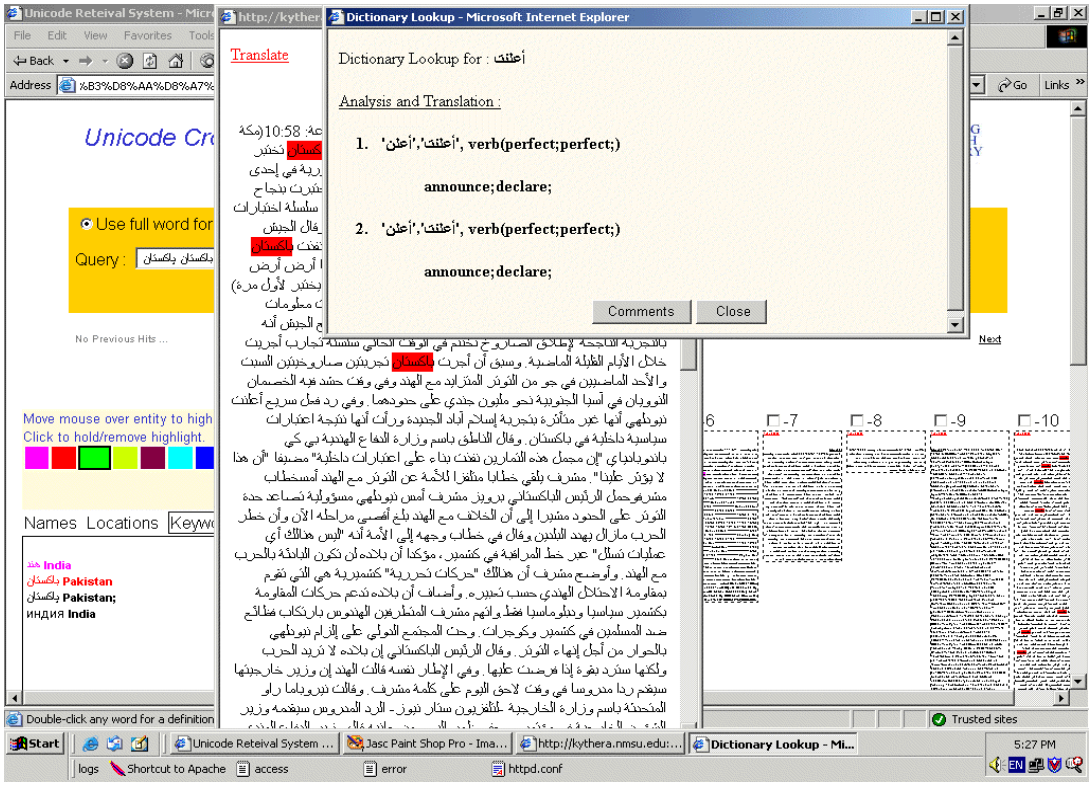

Figure 6. Arabic word translation

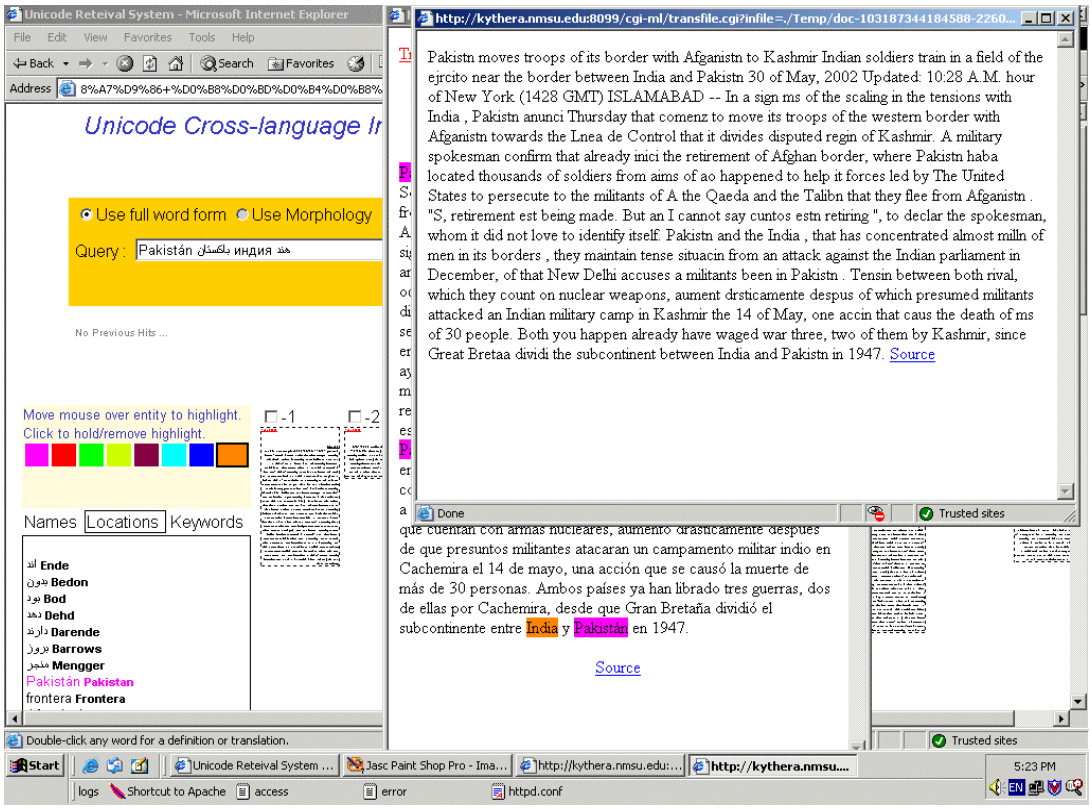

Figure 7. Retrieved document English translation

4.3.2. Document-level Translation

This option allows the user to translate the full document from its original language into English. The quality and type of translation depends entirely on the translation system used to handle the task (see Figure 7 on page N).

5. Evaluation

Evaluating system performance has been the goal of numerous experiments carried out as a central component of the system's development. One of the evaluations was done as part of the Arctos project. Here, the user was provided with a browser-based interface with which to enter English queries. After an initial query is entered, the query was translated using a simple word-for-word or phrasal translator. The user was then asked to interactively improve the query translation using bilingual translation resources and finally submit the query for retrieval against a document collection in the target language. The retrieved documents were presented using document thumbnails and query term highlighting. The user was also asked have the documents translated from the target language into English by the Babelfish translation engine from Systran that has been made available by Alta Vista.

For the interactive task in this preliminary study the user was to use an English TREC CLTR track topics to retrieve and judge the relevance of German documents. Users, who claimed to have no knowledge of German, formed their English query based on the TREC topic statement in English. They then constructed and modified a German query using the on-line dictionary resources. The modified query was then submitted to the URSA engine for retrieval and the retrieved documents were examined using the German equivalents with English glosses and document thumbnail interface. The documents that appeared to be most promising were translated into English and the top 10 documents retrieved were finally judged to be either relevant or non-relevant.

The results are shown in Figure 8. They consist of a comparison of the relevance judgments made by the user in the study with the "correct" judgments provided by NIST for the TREC-6 Cross-Language evaluation track. Documents were excluded if this system retrieved them but they were not among those for which TREC judgments for were available. From these numbers, a false hit ratio can be calculated, that is, the ratio of the number of documents that were non-relevant but judged relevant (11) to the number relevant and judged so (69). The false hit ratio for this experiment is only 15.9% In addition, a false drop ratio can be calculated, that is, the relative number of times a relevant document was incorrectly judged irrelevant (1) to the number of times an irrelevant document was judged irrelevant (43). The false drop ratio in this case is only 2.32%. The combined performance figures for this system indicate a very low percentage chance of error in using this cross-language retrieval system.

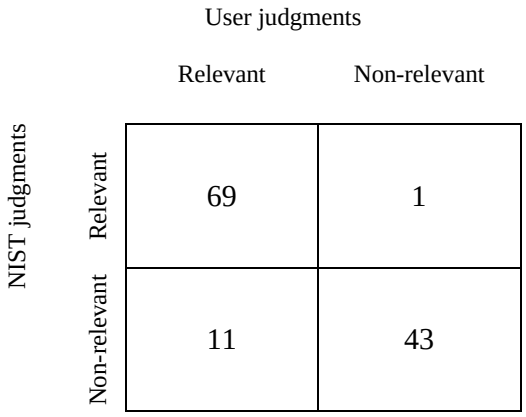

Figure 8. Document judged by the user compared to the NIST judgments

6. Conclusions

The system described here is dependent on the quality of a large set of language resources for its successful operation. These include:

- ?? morphological analyzers,
- ?? annotated bilingual dictionaries,
- ?? high-accuracy proper name recognizers,
- ?? high-quality transliteration procedures for proper names,
- ?? machine translation.

Each of these resources requires further research and development especially in regard to how their combined role in a multilingual IR system can be used to improve the accuracy of retrieval and understandability of the documents retrieved.

The effect of morphology on retrieval system performance, for example, is still an undecided issue. Most of the research is done on English documents and naïve stemmers, such as Porter [REF], seem to provide adequate retrieval. When morphologically rich languages are considered, however, many ambiguous stems may have to be considered. Since the weighting statistics of most IR engines assume a

single stem per token, further work is needed to determine appropriate weighting schemes when each token in a text may give rise to multiple indexing terms.

For the type of retrieval described here it is necessary to combine the richness of a monolingual dictionary, which attempts to explain the different senses of a word, with the preciseness of a pocket size bilingual dictionary, which attempts to give possible translation equivalents but little else. The detail of the monolingual dictionary would allow the user to pick the right terms and the precise translation equivalents would prevent query drift. It might be possible to semi-automatically combine resources of each type to provide a more appropriate CLIR dictionary.

Developing high accuracy name recognition for multiple languages seems to be a useful and fairly well understood task. These combined with onomasticons (proper name lexicons) to translate from one language to another and with intelligent transliteration should allow the enhancement of translation software. One common problem in most off-the-shelf translation software is its tendency to translate proper names as words in the language (e.g. *Castro* as *I castrate*). Name recognition can thus help users filter relevant documents and also improve the readability of translations.

The goal of the current system is to allow a user to find and filter documents in languages they do not speak. Plans are to expand it to handle more languages. In addition further evaluations on the use of multiple methods for skimming through document content and minimizing the need to read bad quality automatic translation must be carried out.

A version of the UCLIR system is publicly accessible at : *http://kythera.nmsu.edu:8099*

Referencias

Buckley, C., Mitra, M., Walz, J., and Cardie, C. (1998). Using clustering and superconcepts within SMART: TREC-6. In E. Voorhees and D. Harman (eds.), Proceedings of the sixth Text Retrieval Conference (TREC-6), pp.107-124. NIST Special Publication 500-240.

Davis, Mark, and Ted Dunning. (1995) Cross-Language Text Retrieval using Evolutionary Optimization. (EP95 in San Diego).

Kwok, K. L., Grunfeld, L., and Xu, J. H. (1998). TREC-6 English and Chinese retrieval experiments using PIRCS. In E. Voorhees and D. Harman, Proceedings of the Sixth Text REtrieval Conference (TREC-6), pp.207-214. NIST Special Publication 500-240.

Ogden, William, James Cowie, Mark Davis, Eugene Ludovik, Sergei Nirenburg, Hugo Molina-Salgado, and Nigel Sharples. (1999) Keizai: An Interactive Cross-Language Text Retrieval System. Paper presented at the Workshop on Machine Translation for Cross-language Information Retrieval, Machine Translation Summit VII, September 13-17, 1999, Singapore.

Ogden, William, and Mark Davis. (1998) Design, Implementation and User's Guide to URSA, the UNICODE Retrieval System Architecture.

Ogden, William, and Mark Davis. (2000) Improving Cross-Language Text Retrieval with Human Interactions. Hawaii International Conference on System Sciences, HICSS-33 January 4-7, 2000.

Unicode standards, The Unicode Consortium 2002 www.unicode.org