Information Access across Languages on the Web: From Search Engines to Digital Libraries

Jiangping Chen, Yu Bao

Department of Library and Information Sciences, University of North Texas
1155 Union Circle #311068, Denton, Texas 76203-5017
Jiangping.Chen@unt.edu, Yu.Bao@my.unt.edu

Abstract:
Information access across languages challenges researchers and practitioners in many disciplines, especially machine translation (MT) and Cross-language Information Retrieval (CLIR). Google’s cross-language search is a model that integrates MT and CLIR technologies to help users find information on the Web that is not written in their familiar languages. This paper overviews the functions of Google’s cross-language search and its performance. It proposes strategies that digital libraries can apply for implementing multilingual information access through the discussion of the cases of five bilingual or multilingual digital libraries.

Keywords:
Digital Libraries, multilingual information access, Google, cross-language search

Introduction

Recent and continuing advances in online information systems are creating many opportunities and also new problems in information access and retrieval. Although English is the language of the Internet (Flammia & Saunders, 2007), online documents are available internationally in many different languages. This provides opportunity for users to directly access previously unavailable sources of information. Cross-language Information Access (CLIA) is the desideratum of many Web users: information can be understood and used from various Web pages in multiple languages.

Researchers in various disciplines have been diligently working on exploring computing algorithms and systems. CLIA research has been vigorously pursued through TREC (http://trec.nist.gov/), CLEF (http://www.clef-campaign.org/), an Asian Language Retrieval and Question-answering Workshop called NTCIR (http://research.nii.ac.jp/ntcir/) and other forums. Significant experimental results have been obtained in cross-language summarization workshops and cross-language named entity extraction challenges by the Association for Computational Linguistics (ACL) and the Geographic Information Retrieval track (GeoCLEF) of CLEF (Gey et al, 2006). Many related research projects have been funded by U.S. government agencies and the governments of other countries. Exploration and application of related technologies in areas such as digital libraries are ongoing and becoming increasingly popular (Larson, Gey, & Chen, 2002; Wang, Lu, & Chien, 2004; Monroy, Furuta, & Castro, 2007). However, there is little application of these technologies in existing information access systems such as commercial online services and digital libraries.

In 2004, search engines began to provide various language supports. Zhang ad Lin (2007) investigated the multiple language support features in 21 search engines. The selected search engines were categorized into regular search engines (such as Google, Yahoo, and MSN), meta-search engines (such as Exite, HotBot, and WebCrawler), and visualization search engines (Kartoo, Onlinelink, and Ujiko). Zhang and Lin summarized the characteristics and functions of these search engines in the following five aspects: the number of supported languages, visibility of language support, translation ability, result
presentation, and interface design. Google was identified as the regular search engine with the best multiple language support (Zhang and Lin, 2007, p530).

On May 23, 2007, Google launched its “Translated Search” in its Google Language Tools (http://www.google.com/language_tools) in addition to other language support services and tools. Here we use the term Cross-Language Search instead of “Translated Search” in order to reflect its relationship with the field of Cross-Language Information Retrieval (CLIR). Greg Notess (2008) found Google was the only search engine providing cross-language search. He briefly described the procedures of this new service and considered it useful for monolingual searchers to explore information content in other languages.

The launch of the cross-language search by Google was a breaking-through event because it signified the transition from CLIR research to its actual application. It was the first time that CLIR and machine translation (MT) were integrated to provide a real application on the Internet (Chen & Bao, 2009). In this paper, we would like to briefly summarize the cross-language search function provided by Google Language Tools (we use the acronym GLT to represent Google Language Tools in the remainder of this paper) and then discuss possible strategies that can put existing technologies to practical use.

This paper is organized as follows. The next section overviews the current research and progress in MT and CLIR, which constitute the major challenges for cross-language search, and then reports GLT’s cross-language search service and its performance through a small-scale evaluation. After that, we describe the characteristics of three digital libraries that provide multilingual information access. The next section proposes strategies that digital libraries can apply in order to serve global users based on previous analysis of Google cross-language search and the current status of multilingual information access in digital libraries. The paper concludes with suggestions for future research and development in multilingual information access on the Web.

Google’s Cross-Language Search

The General Process of Cross-Language Search

Cross-language search aims at facilitating information access across languages. It is built upon many years of research and development in Machine Translation (MT) and Cross-Language Information Retrieval (CLIR),

Machine Translation (MT) has been an important field in Artificial Intelligence. MT automates the process of language translation, which normally includes analyzing and understanding information in one language and expressing it in another language. MT is difficult and complex because it involves understanding and interpretation of the connotative meaning in the original language and its expression in the target language using correct terminology and syntax. Machine translation systems apply various translation strategies to automatically convert text or speech from one language into one or more other languages. Manning and Schutze (1999, p.464) summarized four different levels of translation strategy for machine translation. The lowest level is word-level translation, or word-for-word substitution in which the system attempts to find a word in the target language for each word in the original language. Other levels include syntax-based, semantics-based, and knowledge-based translation, which also consider the structure and semantics of the translated text. The desired translation is the one that expresses the exact meaning in the source text with correct syntax. Current MT research focuses on statistical modeling based approaches for translation. MT systems build statistical models automatically “learned” from parallel corpora (texts with the same meaning but written in the two languages of interest) and use the models to translate one language to the other. Web users can find several online machine translation services such as SYSTRAN (http://www.systransoft.com/), Live Translation by Microsoft (http://www.windowslivetranslator.com/), and Google Language Tools.
Cross-Language Information Retrieval (CLIR) is a subfield of the traditional field of Information Retrieval (IR). It provides users with access to information that is in a different language from their queries (Chen, 2006). Oard and Diekema (1999) identified three basic transformation approaches to CLIR: query translation, document translation, and interlingual techniques. Query translation based CLIR systems translate user queries into the language in which the documents are written. This approach has been the most widely used by CLIR experimental systems because of its simplicity and effectiveness. Most CLIR experimental systems emphasize finding the relevant documents from the collections, but the systems do little with the translating of returned documents.

Google’s cross-language search integrates CLIR and MT to provide the full function of finding information in languages different from users’ queries. Figure 1 shows the screen shot of GLT homepage.

![Figure 1. GLT’s Cross-Language Search Page](image)

On the GLT homepage as illustrated in Figure 1, a user can type in a search query in its search textbox, specify the language of the query, specify the language of the result pages, and then click the button “Translate and Search”. Then GLT will conduct cross-language search and present the search results in both query language and the intended language in two separate columns. For each result, GLT provides the title, a short summary, and the URL of the page, just like the results presented by Google Web search. For example, to find information resources on alternate energy sources in Chinese, we type in “green energy” in the search box of GLT. We specify the query language as English, and search pages written as Simplified Chinese. GLT returns the translated pages (in English) and the original pages (in Simplified Chinese). Figure 2 shows the screen shot of the search result.
Google's cross-language search contains the following components or processes:

A. Search Interface: this allows a user to type in the search terms and to specify the language of these search terms and the language of the retrieved Web pages;
B. Query Translation: the system will translate the users’ queries into the languages of the Web pages so that matching between the queries and the pages can be conducted;
C. Web Search or Information Retrieval: the actual search for relevant pages based on a retrieval algorithm;
D. Machine Translation of Results: the retrieved Web pages are translated into the languages of the queries;
E. Result Interface: the translated pages are presented to the users. The system may also present the results in their original languages simultaneously.

The Search Interface and Result Interface are for interaction with users, while other components that are transparent to users handle the most difficult issues of cross-language search: query translation, search, and machine translation of result pages.

As of 30 April 2009, GLT supports cross-language search for 41 languages. This means one can use GLT to submit search terms in one of the above languages, to search pages written in the remaining 40 languages, and to read search results in the language he/she understands, i.e. the language of his/her search terms.

**Google’s Translation Mechanism and Performance**

Every language related service provided by GLT, including the cross-language search, involves machine translation. How does Google carry out the translation then? The Google Translate FAQ (http://www.google.com/intl/en/help/faq_translation.html) explains Google’s strategy for
translation. To build its machine translation system, Google feeds “…the computer billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages.” It then applies “…statistical learning techniques to build a translation model.” (Google Translate FAQ, 2008). Google continues to work on translation quality, attempting to improve the performance of MT by understanding the context of words.

Chen and Bao (2009) conducted a small-scale evaluation to understand the performance of the GLT. They adapted 50 topics that have been evaluated at NTCIR-5 Cross-Lingual Information Retrieval Task (http://research.nii.ac.jp/ntcir5-cfp-en.html). These topics were originally presented in four languages: Traditional Chinese, Korean, Japanese, and English (http://research.nii.ac.jp/ntcir-workshop/OnlineProceedings5/cdrom/CLIR/ntc5-CLIR-ntc5-CLIR-ntc5-CLIR-eval.html). Each topic has several attributes that describe the content, such as TITLE, DESC (descriptors), NARRATIVE, and CONC (concepts). The TITLE of a topic consists of short phrases or words separated by the punctuation “,”, and the DESC of a topic is a short sentence describing the information that needs to be found for the topic. Two queries were constructed from each topic. One query consisted of the texts in the TITLE attribute and one query consisted of texts in the DESC attribute. Below is an example of the two queries from one topic:

Query 001-1: 時代華納，美國線上，合併案，後續影響
Query 001-2: 查詢時代華納與美國線上合併案的後續影響.

In total, 100 queries were generated in traditional Chinese from the 50 CLIR NTCIR-5 topics. These queries were divided into two groups: The TITLE group included 50 queries from the TITLE of each topic while the DESC group included 50 queries from the DESC of each topic. Each query was then sent to GLT to search pages in English. The translation results were evaluated by the authors. The translation of a query was judged “Correct” if it was exactly the same as the English version of the topic, or was judged correct semantically and grammatically by the two authors. Otherwise the translation was judged “Other”, which included multiple situations such as false translation, incomprehensible translation, or partially correct translation.

The same queries were also submitted to SYSTRAN (http://www.systransoft.com/). Table 1 shows the results of the evaluation. It demonstrates that both Google and SYSTRAN did much better on the TITLE group than the DESC group. And the difference on performance for TITLE and DESC groups were significant for both systems. Also, Google could correctly translate more TITLE queries but fewer DESC queries than SYSTRAN.

Table 1. Results of GLT’s Translation Evaluation (Chen & Bao, 2009)

<table>
<thead>
<tr>
<th>Queries in TITLE group</th>
<th>Queries in DESC group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Google</td>
</tr>
<tr>
<td>Correct</td>
<td>38 (76%)</td>
</tr>
<tr>
<td>Other</td>
<td>12 (24%)</td>
</tr>
<tr>
<td>Total queries</td>
<td>50</td>
</tr>
</tbody>
</table>

This small-scale evaluation on query translation has actually assessed two types of machine translation: machine translation of words or phrases, and machine translation of sentences. Queries in the TITLE group are more similar to queries a Web user submits to a search engine. They are composed of words and short phrases. Google’s translation service can perform quite well on these queries. Queries in the DESC group are sentences that are similar to those constituting Web pages. Google, like SYSTRAN, failed to translate many of those correctly. Due to the fact that most Web pages include text composed of sentences in natural languages, we consider GLT’s query translation as less of a concern than its machine translation of result pages. However, still 24% of queries were not correctly translated. It is expected the correctness rate will decrease for domain-specific queries as they are unlikely included in general-purpose lexicons used by machine translation systems.
Digital Library with Bilingual or Multilingual Information Access

Due to the fact that Machine Translation usually produces translations that are difficult to understand, many organizations and information systems still rely on human translators for translating documents or files from one language to other languages. As for digital libraries, very few digital libraries have implemented multilingual information access (Chen, 2007).

We analyzed about 150 US digital libraries that we found through DL literature and search engines, yet only five of them could be accessed by using more than one language. Table 2 lists these five digital libraries.

Table 2. Digital Libraries with Multilingual Information Access

<table>
<thead>
<tr>
<th>Library Name</th>
<th>URL</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting of Frontiers</td>
<td><a href="http://frontiers.loc.gov/intldl/mfhtml/mfsplash.html">http://frontiers.loc.gov/intldl/mfhtml/mfsplash.html</a></td>
<td>English/Russian</td>
</tr>
<tr>
<td>France in America</td>
<td><a href="http://international.loc.gov/intldl/fiahome.html">http://international.loc.gov/intldl/fiahome.html</a></td>
<td>English/French</td>
</tr>
<tr>
<td>Parallel Histories</td>
<td><a href="http://international.loc.gov/intldl/eshtml/">http://international.loc.gov/intldl/eshtml/</a></td>
<td>English/Spanish</td>
</tr>
<tr>
<td>The Perseus Digital Library</td>
<td><a href="http://www.perseus.tufts.edu">http://www.perseus.tufts.edu</a></td>
<td>Greek, English, Latin</td>
</tr>
</tbody>
</table>

Meeting of Frontiers is a “bilingual, multimedia English-Russian digital library that tells the story of the American exploration and settlement of the West, the parallel exploration and settlement of Siberia and the Russian Far East, and the meeting of the Russian-American frontier in Alaska and the Pacific Northwest” (The Library of Congress, 2002, About the Project, 1). It is intended for use in U.S. and Russian schools, libraries and by the general public in both countries. The bilingual collection includes Books and Other Printed Materials, Manuscripts, Maps, Photographs and Prints, Mixed Format Collections, Sheet Music, Motion Pictures and Recorded Sound, and Exhibitions. Users can search and browse either in English or Russian.

France in America is “a bilingual, multi-format English-French digital library that tells the story of the French presence in America and the interactions between the French and American peoples from the early 16th to the late 19th centuries” (The Library of Congress, n.d., About the Site, 1). It is part of the Library of Congress’s Global Gateway project which aims to establish cooperative digital libraries with national libraries from around the world. The collection has been designed for students, scholars, and researchers worldwide. Users can browse the collections in the original languages of the collections. In terms of searching functions, the text of the descriptive information records, the full text transcriptions, and the themes in the Web site are presented using Latin 1 character encoding - this encoding includes diacritic marks commonly found in western European languages (i.e., accent marks).

Parallel Histories: Spain, the United States, and the American Frontier is a bilingual, multi-format English-Spanish digital library site that explores the interactions between Spain and the United States in America from the fifteenth to the early nineteenth centuries” (The Library of Congress, n.d., About the Project, 1). It is also part of the Library of Congress’ Global Gateway project. The digital library aims at “making available to students, researchers, and lifelong learners unique documents from the cultural heritage of Spain and the United States.” (http://international.loc.gov/intldl/eshtml/) Users can browse the collections in the original languages of the collections applying the same method as in France in America.
The above three digital libraries are all developed by the Library of Congress in partnership with libraries in the respective countries.

**International Children’s Digital Library** was created by an interdisciplinary research team at the University of Maryland in cooperation with the Internet Archive." The ICDL collection has two primary audiences. The first audience is children ages 3-13, as well as librarians, teachers, parents, and caregivers who work with children in this age group. The second audience is international scholars and researchers in the area of children's literature" (International Children’s Digital Library, n.d., Frequently Asked Questions). The library can be accessed in 11 languages.

**The Perseus Digital Library** is in the Department of the Classics, Tufts University. Its collection includes Classics, Papyri, Renaissance, London, California, Upper Midwest, and Tufts History. It is for both general readers and specialists. The library provides many tools for accessing the collection, such as English to Greek and Latin Word Search, Greek and Latin Morphological Analysis, and a Greek and Latin Vocabulary Tool.

The above digital libraries share the following characteristics:
- They have been funded by various funding agencies, especially from the federal government;
- They are the products of collaboration. People from different countries work together to produce the bilingual or multilingual collections;
- They serve a broader or global user community in which users speak different languages;
- They do not employ cross-language search, cross-language information retrieval techniques or machine translation.

Now, we are asking the questions: Should digital libraries employ cross-language search? How can digital libraries achieve bilingual or multi-lingual information access with minimal cost and effort?

**Multilingual Information Access in Digital Libraries: Possible Strategies**

The CLIR community has been focusing on improving retrieval performance for more than 10 years. Numerous matching strategies have been explored to realize CLIR between various language pairs. However, CLIR service has been offered by very few information systems. Translation performance has been considered the major obstacle of applying the technologies to practical systems (Gey, Kando, and Peters, 2005). Also, the lack of knowledge of the users is among the major reasons for this situation. As pointed out by Petrelli, Beaulier and Sanderson (2002), little effort has been made to identify the users of CLIR systems and to fully understand how these users can make use of such systems.

Search has become an essential component of all information systems, including digital libraries. A large amount of information stored in digital libraries is not accessible to search engines. It would benefit global users if digital libraries were to offer multilingual information access services so that more people in the world could use the information in those digital libraries. Below are strategies that can be considered to develop multilingual digital libraries with limited funding.

Digital library (DL) developers should:

1. Collaborate with researchers in CLIR and MT to explore solutions that are appropriate for the specific digital objects while seeking funding to support cross-language search as a value-added service. As digital objects are more organized than Web pages crawled by search engines, it is possible that better performance of machine translation could be achieved through the construction of a customized knowledge base for machine translation software.
(2) Collaborate with DL developers in other countries to increase the languages that the users can access. Many digital libraries manage precious digital assets that can be attractive to people on the other side of the earth. Collaboration with colleagues in other countries would make information resources more available.

(3) Collaborate with the users. In the current digital age, even monolingual digital libraries are accessed by people who don’t know the language (Sorid, 2009). Social computing has been widely used on the Internet, and it can play a big role in involving users in the multilingual information access services: users may even volunteer to translate digital objects into another language. They may help to correct errors produced by machine translation systems. They may donate money to help the DL to offer the new service if they know the significance of the service, or the information needs from the other side of the earth.

(4) Take a step-by-step approach. DL developers can first implement a multilingual interface, then the metadata, and then the whole collection.

Conclusions

Google Language Tools provide multiple language support tools for Web users. These tools include a cross-language search service for 41 languages, monolingual search in user preferred language or country, machine translation of texts or Web pages, and online dictionary lookup. GLT’s cross-language search integrates MT and CLIR, which have been investigated for many years by separate research communities. Although systematic user evaluation of GLT’s cross-language search has yet to be conducted, our evaluation shows that GLT can do a reasonably good job translating short queries from Chinese into English. GLT’s cross-language search service enables Web users to access previously inaccessible information.

Information systems such as digital libraries would better serve their users if language support services were integrated as part of the systems. Very few digital libraries in the United States have implemented multilingual information access. However, digital libraries with less money should also consider adding bilingual or multilingual information access through collaboration with CLIR and MT researchers, colleagues in other countries, and the users. As for the specific implementation approach, the model of Google’s cross-language search should not be considered the only approach. A digital library may choose to conduct machine translation for all the documents before indexing instead of doing query translation at the time of searching; digital libraries with stable collections can apply a computer-assisted mechanism to build their translation knowledge base for query translation. The study of user needs under specific conditions will help developers to build efficient and effective systems for information access across languages.

As for future research, we would like to understand more about the needs and information behavior of bilingual users because bilingual users have been identified as the most possible users for CLIR systems. Also, we plan to collaborate with small digital libraries to investigate effective and efficient solutions to providing multilingual information access for the digital library users.
References


