f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Cross-language search: The case of Google Language Tools

by Jiangping Chen and Yu Bao

Abstract

This paper presents a case study of Google Language Tools, especially its cross–language search service. Cross–language search integrates machine translation (MT) and cross–language information retrieval (CLIR) technologies and allows Web users to search and read pages written in languages different from their search terms. In addition to cross–language search, Google Language Tools provides various language support services to multilingual information access. Our study examines the functions of Google Language Tools and the performance of its cross–language search. The results and analysis show that Google Language Tools are useful for Web users. Its cross–language search service provides quality query translation while the automatic translation of result pages needs further improvement. The paper suggests that cross–language search could be used by different types of Web users. The authors also discuss the strategies and important issues with regard to implementing multilingual information access services for information systems.

Contents

Introduction

The rapid growth of information and services on the Web provides many opportunities to users for accessing information. Web pages are available internationally in many different languages, although English remains the language of the Internet (Flammia and Saunders, 2007). The "multilinguality" of Web content provides opportunities for users to directly access and use previously incomprehensible sources of Web information; however, Web users find it difficult to take advantage of these opportunities when the online information access systems are monolingual.

Monolingual search engines only allow users to enter a search query in the language of the Web documents to conduct the search. Users who are unfamiliar with the language of Web documents retrieved are often unable to obtain relevant information from these documents. This restriction clearly limits the amount and type of information that an individual user can access. In a global community, users are looking for online information access systems or services that can help them find and use information presented in native or non–native

languages. For example, a Chinese student to whom English is her second language wants to search an English–language resource. The student might be able to use English for the query search, but needs the returned results to be translated into Chinese to facilitate her reading and use of the information.

Fortunately, in 2004 search engines began providing various language supports. Zhang and Lin (2007) investigated 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 [1]:

> "Google held an indisputable first position compared with other search engines. Google supported 37 different languages. Google's translation tool was the strongest one of all the search engines. The number of language translation pairs (one language can be translated into another language) reached eleven. Its translation mechanism allowed a direct translation of a Web page without using copy and paste in a search results list. It supported both Web page translation and paragraph translation."

On 23 May 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 a research field called cross–language information retrieval. 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 breakthrough because it signified the transition from cross–language information retrieval (CLIR) research to a real application. It was the first time that CLIR and machine translation (MT) were integrated to provide a real application on the Web. In this paper, we conducted a case study to examine the services provided by Google Language Tools (we use the acronym GLT to represent Google Language Tools in the remaining paper) and the performance of its cross–language search. Our purposes are two–fold:

1. To help Web users to appropriately evaluate and use GLT and other language support services on the Web.

Research has indicated that online information users may need to access information in non–native languages for a variety of purposes. For example, Oard (1997) pointed out that CLIR technologies would help commercial online services such as Dialog and Lexis/Nexis expand their markets. Individual government and international institutions also have the need to search and access large amounts of multilingual documents. Journalists look for news and stories in foreign language newspapers, and business analysts gather foreign business information and provide services to different countries. Petrelli, *et al.* (2002) specified that translators wanted to use CLIR systems to serve clients in other countries, and information professionals such as reference librarians searched as intermediaries for patrons. Gonzalo (2002) believed that bilingual users would more easily recognize the retrieved documents and grasp the context when using CLIR than arduously forming the queries and looking at the results in their second languages. A recent study (Cleveland, *et al.*, 2007) demonstrated that language is a serious barrier for Chinese communities in the Dallas–Fort Worth Metroplex area in Texas to search and use quality online medical information that is mostly in English.

This study aims to provide advice to Web users on how to appropriately use language support services through the analysis of Google Language Tools. Specifically, we plan to find answers to following questions:

> *What are the language support services provided by GLT?*
> *How useful is the cross–language search in GLT in terms of translation and retrieval?*
> *What are the alternatives to translating result pages in*

cross–language search?

Case studies are appropriate when researchers explore in depth a certain event or activity [2] in order to discover flaws in existing theories or to gain explanatory insights. In this study, we collected data from multiple sources: the Google Website including its FAQ pages, literature on MT and CLIR research, online discussions on Google's cross-language search, and some comparative studies of Google's search performance. We also conducted a small-scale evaluation of its query translation performance to gain insight into the performance of Google's translation technology and the reasons behind its performance.

2. To identify important issues as related to technological transfer in MT and CLIR.

Through a discussion of the implications of GLT's cross–language search, we would like to identify important issues in order to transfer research achievements in CLIR and MT to practical applications. GLT's cross–language search is one of the models that integrate MT and CLIR in order to provide multilingual information access. Information systems such as digital libraries could take advantage of cross–language search to provide or enhance their language support services. These services may greatly influence future information access on the Web. In other words, our study discusses the following question:

What are important issues in order to transfer research in CLIR and MT to practical services for information systems such as digital libraries?

The remaining paper is organized as follows. The Background section provides an overview of current research and progress in MT and CLIR, summarizing the major challenges for cross–language search. The next section reports functions and services provided by GLT, especially its cross–language search service. The Evaluation section examines the performance of GLT's cross–language search through literature review, document analysis and a query translation experiment. The Discussion section discusses concerns as related to the performance of GLT and issues regarding technological transfer of MT and CLIR research to practical systems. The paper concludes with a summarization and suggestions for future study on research and development in cross–language search.

Background

Cross–language search aims at facilitating information access across languages. It is built upon more than fifty years of research and development in machine translation (MT) and more than ten years of research in cross–language information retrieval (CLIR).

Machine translation (MT) has been a field in artificial intelligence. MT aims to automate the process of translation, which normally includes analyzing and understanding information in one language and expressing it in another language. Translation is difficult because the process involves interpretation of meaning in one language and its expression in a target language using correct terminology and syntax. Automatic translation by a computer system, *i.e.*, machine translation (MT), is even more difficult since the computer has not achieved a human–like understanding of languages. Machine translation (MT) systems apply various translation strategies to automatically convert text or speech from one language to one or more other languages. Manning and Schütze [3] summarized four different levels of translation strategy for machine translation. The simplest approach 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 methods 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 explores various statistical modeling based approaches for translation. MT systems build statistical models automatically "learned" from parallel corpora (texts with the same meaning but written in 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. Google Language Tools will be discussed in this paper.

Cross–language information retrieval (CLIR) is a subfield of the traditional information retrieval (IR). It provides users with access to information that is in a different language

from their queries (Chen, 2006). Research in CLIR has been significantly advanced by three major evaluation forums: a cross–language information retrieval track at TREC (Text Retrieval Conference: http://trec.nist.gov/) from 1997–2002, the Cross–Language Evaluation Forum (CLEF: http://clef-campaign.org/) evaluating many European languages, and the NTCIR Asian Language Evaluation (http://research.nii.ac.jp/ntcir/) that covers Chinese, Japanese, and Korean. These forums provide CLIR researchers and system developers with infrastructures for algorithmic developing, system testing, and resource sharing.

The basic strategy for information retrieval is to match documents to queries. A transformation on either side or both is necessary if the queries and documents are not written in the same language, as in the case of CLIR, since the match cannot be directly conducted. 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 to the language that the documents are written. Document translation is the reverse of query translation where documents are translated into the query language. The interlingual approach translates both documents and queries to a third representation. Among the three approaches, the query translation approach is the generally accepted approach and has been applied by most CLIR experimental systems because of its simplicity and effectiveness. Query translation based CLIR systems use various knowledge resources, such as bilingual dictionaries, MT systems, parallel texts, or a combination of these to translate queries from one language into another language; and then conduct monolingual search to retrieve relevant documents. Most CLIR experimental systems emphasize finding the relevant documents from the collections, but do little with the translating of returned documents.

Cross–language search integrates CLIR and MT to provide the full function of finding information in languages different from users' queries. Figure 1 illustrates the basic processes involved in a cross–language search system.

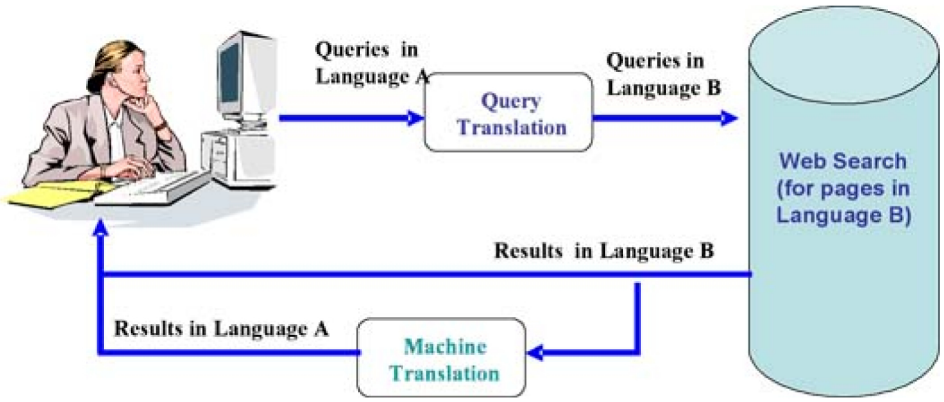

Figure 1: Architecture of Google's Cross–Language Search Service.

In a system illustrated in Figure 1, a user submits a query in one language, the system conducts several steps to find out the relevant Web pages written in a different language that the user may not understand (CLIR process). At last, the system translates the results into a language of the user's query (MT process). For example, an English–speaking user who needs information on alternative medicine written in Chinese can submit search terms in English to a system described in Figure 1. The system will translate the query into Chinese, and return the search results in English back to the user. In this example, translation between the two involved languages happens on query level (translate the English query into Chinese) and on the document level (translate the relevant Chinese documents into English).

The key function provided by Google Language Tools (GLT) is cross–language search. And the process illustrated in Figure 1 is exactly what GLT's cross–language search does. Next we examine the functions and performances of GLT.

Google Language Tools (GLT)

GLT is an integrated interface that provides access points for cross–language search and other language support services.

Cross–language search

GLT presents its language support service on the top of its home page: http://www.google.com/language_tools. A user can type in a search query in its search textbox, specify the language of the query, specify the language of the documents, 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 autism treatment in Chinese, we type in "autism treatments" 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.

Figure 2: A GLT search result page.

How does GLT do cross–language search? Google integrates Web search and machine translation to provide the cross–language search service. The architecture illustrated in Figure 1 depicts the process involved in GLT's cross–language search. In general, a cross–language search system consists of following components:

A. Search interface: It allows a user to type in search terms and to specify a language for these search terms and a language for retrieved Web pages;
B. Query translation: The system will translate the users' queries into the languages of Web pages so that matching between queries and 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 languages of the queries; and,
E. Results interface: the translated pages are presented to users. The system may also present results in their original languages simultaneously.

The search interface and results 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 28 October 2008, GLT supports cross–language search for 35 languages: Arabic, Bulgarian, Catalan, Chinese (simplified), Chinese (traditional), Croatian, Czech, Danish, Dutch, English, Filipino, Finnish, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, and Vietnamese. This means one can use GLT to submit search terms in one of the above languages, to search pages written in the remaining 34 languages, and to read search results in the language she understands, *i.e.*, the language of her search terms.

Other language support services

Other language support services provided by GLT include:

Monolingual search in your preferred language or country. GLT provides the choices of searching the Web in your preferred language. Google will return Web pages in that language. Also, you can visit Google's site in your country or region. In this case, Google will return results from that country or region. For example, you can use Google Canada to search Web pages in Canada.

Machine translation of Web pages. In GLT, users can type some text in one language and get it translated into the targeted language. GLT can also translate a Web page — users type in the URL of a Web page — and get the Web page translated into the targeted language.

User's feedback on translation. Google allows users to correct its translation through a small feedback textbox. On the top of the result pages as shown in Figure 2, a hyperlink in small font says "not quite right? Edit". This link leads users to a popup window that allows users to provide their own translation to their queries.

Online Dictionary lookup. This function allows users to enter a word in one language and get it translated in another language (http://translate.google.com/translate_dict?hl=en).

Other services. Google also provides certain services that allow users to convert their homepages into other languages. Users just need to select the language of their Web pages, copy and paste the HTML code Google provides to include the gadget on their Web pages (http://translate.google.com/translate_tools?hl=en). Another service provided by Google is the option of adding a list of buttons of languages that users can add to their browser's toolbar. Then whenever users want to translate a Web page, they can just click a button. Users can translate any part of the page by selecting that part before they click. Google also provides a free, downloadable toolbar that can translate any words from English to other languages, such as Chinese (traditional or simplified), Japanese, Korean, French, Italian, German, or Spanish.

Google's translation mechanism

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 and attempts to improve the performance of MT by understanding the context of words.

Translation performance: An evaluation

What is the performance of GLT's cross–language search? Can Web users rely on GLT to find information they previously could not access? These questions are important to whoever wants to use GLT.

GLT's cross–language search consists of three major processes: query translation, search, and machine translation of result pages. Both query translation and machine translation of result pages are conducted by Google's machine translation service. Google is the most used Web search engine (Dudek, *et al.*, 2007). Its famous page rank search algorithm returns Web pages that are most hyperlinked by other pages. User studies show that Google

performs better searches than other search engines (Liu, 2007) and even better than some digital libraries' search functions (McCown, *et al.*, 2005). As for machine translation, GLT applies a statistical machine translation tool and has achieved top results in NIST's machine translation evaluation (NIST, 2005).

However, comparatively good performance does not mean the performance is adequate for practical use. In this paper, we report a small–scale evaluation of GLT's query translation in order to achieve an understanding of the quality of its machine translation. We sent to GLT 50 topics that have been evaluated at NTCIR–5 Cross–Lingual Information Retrieval Task (http://research.nii.ac.jp/ntcir-ws5/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-eval.html). Each topic has several attributes that describe the content, such as TITLE, DESC (descriptors), NARRATIVE, and CONC (concepts). Table 1 shows one of the topics in both Traditional Chinese and English in XML format. 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.

In our evaluation experiment, we constructed two queries from each topic. One query consisted of the texts in the TITLE attribute and one query consisted of texts in the DESC attribute. For example, we extracted two queries from the sample topic in Table 1:

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, we generated 100 queries in Traditional Chinese from the 50 CLIR NTCIR–5 topics. We divided the 100 queries into two groups: TITLE group and DESC group. The Title group included the 50 queries from the TITLE of each topic, and the DESC group included the 50 queries from the DESC of each topic. We considered that queries in TITLE group were easier than those in DESC group in terms of translation and retrieval. Queries in TITLE group are composed of Chinese words or short phrases that do not need word segmentation. However, queries in the DESC group are short sentences that need word segmentation in most cases. Word segmentation is an important and complicated task for MT systems and segmentation errors may greatly affect the accuracy of machine translation (Nie and Ren, 1999). In our study, we wanted to see whether there was any significant difference between the translation results of the two groups.

Table 1: A sample topic extracted from NTCIR–5 CLIR Task.

Chinese version	English version
<TOPIC> <NUM>001</NUM> <SLANG>CH</SLANG> <TLANG>CH</TLANG> <TITLE>時代華納, 美國線上, 合併案, 後續影響</TITLE> <DESC> 查詢時代華納與美國線上合併案的後續影響 </DESC> <NARR> <BACK>時代華納與美國線上於2000年1月10日宣佈合併, 總市值估計為3500億美元, 為當時美國最大宗合併案 </BACK> <REL>評論時代華納與美國線上的合併對於網路與娛樂媒體事業產生的影響為相關。敘述時代華納與美國線上合併案的發展過程為部分相關。內容僅提及合併的金額與股權結構轉換則為不相關。</REL> </NARR> <CONC>時代華納, 美國線上, 李文, Gerald Levin, 合併	<TOPIC> <NUM>001</NUM> <SLANG>CH</SLANG> <TLANG>EN</TLANG> <TITLE>Time Warner, American Online (AOL), Merger, Impact</TITLE> <DESC>Find reports about the impact of AOL/Time Warner merger.</DESC> <NARR> <BACK>Time Warner and American Online (AOL) announced a merger on January 10th, 2000. The market value was estimated at $US350 billion making it the biggest merger in the US.</BACK> <REL>Comments on AOL/Time Warner merger's effects on Internet and entertainment media businesses are relevant. Descriptions of the development of the AOL/Time Warner merger are partially relevant. Information about the total amount

案, 合併及採購, 媒體業, 娛樂事業</CONC> </TOPIC>	and the transformation of ownership structure are irrelevant.</REL> </NARR> <CONC>Time Warner, American Online, AOL, Gerald Levin, merger, M&A, Merger and Acquisition, media, entertainment business</CONC> </TOPIC>

We then sent each query to GLT manually. We instructed GLT to search pages in English. The result pages returned by GLT, as shown in Figure 2, included the translation of the query, and the summaries of pages that satisfied the query in both query language (Traditional Chinese in this evaluation) and the intended languages (English). We then compared the translation results of the queries with their English version in order to assess whether the translation was correct. The two authors served as evaluators in this experiment. 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.

We also submitted the 100 Chinese queries in the exact format to SYSTRAN (http://www.systransoft.com/), and conducted the same evaluation. Table 2 is the result of our evaluation of the 100 queries.

Table 2: Results of GLT's translation evaluation.

	Queries in TITLE group		Queries in DESC group	
	Google	**SYSTRAN**	**Google**	**SYSTRAN**
Correct	38 (76%)	26 (52%)	9 (18%)	13 (26%)
Other	12 (24%)	24 (48%)	41 (82%)	37 (74%)
Total queries	50	50	50	50

Table 2 demonstrates that both Google and SYSTRAN did much better on TITLE group than 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. Assuming the queries were random samples of all queries submitted to these two systems, we did a Chi–square test on Google's query translation performance and compared it with SYSTRAN. For TITLE queries, Google did significantly better (χ = 11.54, ρ value < 0.001) than SYSTRAN. But it didn't significantly differ from SYSTRAN on translating DESC queries (χ = 1.66, ρ value >0.1).

This small–scale evaluation on query translation assessed two types of machine translation: machine translation of words or phrases, and machine translation of sentences. Queries in TITLE group are more similar to queries a Web user submits to a search engine and are composed of words and short phrases. Google's translation service performs quite well on these queries. Queries in the DESC group are basically 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 less of a concern than its machine translation of results pages. Machine translation of results pages can be a bottleneck for cross–language search.

Discussion

In this section, we discuss the implication of GLT to Web users, researchers in CLIR and MT, and developers of information systems.

User's opinions of Google's Cross–language Search

We identified a data source that contained comments posted by users on GLT's cross–language search. On 16 May 2007, Michael Arrington posted a very short announcement on TechCrunch (http://www.techcrunch.com/) about Google's launch of its cross–language search engine. The message included a very fuzzy screen shot which displayed English results on the right side and the Arabic equivalents on the left.

In response to this announcement, there were 25 messages posted by 22 individuals from 16 May to 5 June 2007. These messages provide valuable information on Web users' reaction to this event. A simple content analysis of the postings resulted in the following categories as shown in Table 3.

Table 3: Comments on Google's cross–language search.

Category	Summary of comments	Frequency
Positive comments	The idea is interesting It is neat Very useful or useful for certain population Google is doing the right thing In the right direction, but more work needs to be done This is a big event with potentials	15
Negative comments	Google is not doing the right thing — bad translation, wrong query redirections… Google is poor in Arabic (MT) and not multilingual	4
Irrelevant comments	Advertisement for other Web services Suggestions on what a search engine should do Other similar development elsewhere Differences between multi–lingual search and cross-language search	6

Fifteen postings reflected a positive reaction to the original announcement. Essentially, users thought that Google was doing the right thing. The cross–language search is interesting and especially useful for non–English speakers. Even though a given translation is not perfect, this tool would certainly help many access information in various languages. In particular, one posting mentioned that cross–language information retrieval would help him in Japan since he didn't know Japanese.

Only four postings expressed negative opinion on this event. Two of them thought machine translation (MT) was not mature enough for practical use. One believed Google poorly translated Arabic. One posting thought cross–language search was not helpful as users were usually interested in searching in more than two languages.

Six of the postings were not directly related to Google's cross–language search. Three of these postings discussed multilingual search projects, forums, or differences between multi–lingual and cross–language searches.

In summary, the majority of postings welcomed Google's launch of the cross–language search service with a concern over the quality of machine translation.

Implications of GLT for Web users

We believe GLT is a useful tool for many Web users. Ogden and Davis (2000) summarized two types of users for CLIR: bilingual users who formulate their queries in their native languages to retrieve documents in their second languages; and, monolingual users who are interested in finding information in other languages. Similarly, GLT will benefit these two types of users, helping them to access information on the Web. We suspect that there will be a variety of uses for this and other language support tools by Web users. Here are some possible applications:

- Immigrants who know little English can search U.S. government Web pages as well as other documents for information about immigration;
- Investors interested in examining new markets can search news reports or Web documents about foreign companies;
- College students learning a foreign language can submit queries in English to find historic stories in languages that they are studying. They can also look up new words in online multilingual dictionaries;
- Patients or caregivers can search and find medical or treatment information from other countries or in other languages; and,
- U.S. travelers can search for local information and events using GLT *en route*.

In general, bilingual users who can read original result pages will greatly benefit from cross–language search services. Monolinguals, however, may need to seek other assistance to understand the content of results pages due to the less than satisfactory performance of machine translation. For example, they may need a human translator to access information on some retrieved pages. Current Web 2.0 technologies or online social interaction may provide an alternative to human translators — Web users could obtain help from other users who happen to know languages of specific documents.

Implication of GLT's cross–language search for research and development in MT and CLIR

The CLIR community has focused on improving retrieval performance for more than a decade. Numerous matching strategies have been explored to develop CLIR between various language pairs. However, CLIR services have been offered by very few information systems. There is some concern in actually transfering CLIR research to actual information access services (Gey, *et al.*, 2005). The launch of Google's cross–language search marked the first step towards practical use of CLIR in search engines. It is promising that other kinds of information systems, such as digital libraries and corporate information systems, could provide cross–language search functions for their users.

Our evaluation indicates that machine translation of Web pages is more problematic than translating users' queries which are comprised of words and short phrases. Current machine translation systems have not reached a level of accuracy in translating documents from one language to another. Solutions to the problems of assisting users to understand search results should be explored in order to apply CLIR to empirical systems. Better understanding of the needs of users could help in this investigation.

Implications of GLT to other information systems such as digital libraries

Searching has become an essential component of all information systems including digital libraries. A large amount of information stored in digital libraries is not accessible in a variety of search engines. It would benefit global users if digital libraries were to offer multilingual information access services so that more individuals could access unique information found in digital libraries.

Researchers in CLIR and MT could work in collaboration with digital library developers to explore solutions that are appropriate for 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. Moreover, digital library systems could implement social computing services allowing users to help each other to translate results found in digital libraries.

Conclusions

Google Language Tools provide multiple language support tools for Web users. These tools include a cross–language search service for 36 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 that have been investigated many years by a variety of research communities. Although systematic user evaluation of GLT's cross–language search needs 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 information that could not be accessible before. Different types of Web users such as immigrants, investors, students, patients, and travelers may benefit from this service.

The quality of machine translation is a major concern of Web users. The translation of a whole document or a Web page remains a challenge. Google's machine translation system could be improved as human knowledge for translation accumulates in resources such as parallel Web pages and users' feedback files. For Web users, the rapid development and expansion of Web 2.0 technologies may provide alternatives for understanding results pages.

Information systems, such as digital libraries or corporate information management systems, would better serve their users if language support services were integrated into their systems. Important issues of transferring research achievements of CLIR and MT to practical applications include the understanding of the needs of targeted users and careful design, identification, implementation, and evaluation of search and translation strategies. Google's strategy of cross–language search should not be the only model. The designers of different applications have to understand their users in order to choose appropriate strategies for implementation. For examples, an information system may choose to conduct machine translation for all documents before indexing instead of doing query translation at the time of searching. Digital libraries with stable collections can apply computer–assisted mechanisms 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, a log analysis of Google's cross–language search would help understand the behavior of its users. In addition, systematic user–oriented evaluation and testing of cross–language search performance, especially for those languages that are less common on the Web, should be conducted. These studies will provide constructive guidance to information system developers for implementing similar services.

About the authors

Jiangping Chen, Ph.D., is an Assistant Professor at the Department of Library and Information Sciences, College of Information, Library Sciences, and Technologies at the University of North Texas.
Direct comments to: jpchen [at] unt [dot] edu

Yu Bao is a doctoral student in the Interdisciplinary Ph.D. program at the Department of Library and Information Sciences, College of Information, Library Sciences, and Technologies at the University of North Texas.

Notes

1. Zhang and Lin, 2007, p. 530.

2. Creswell, 2009, p. 13.

3. Manning and Schütze, 1999, p. 464.

References

Jiangping Chen, 2006. "A lexical knowledge base approach for English–Chinese cross–language information retrieval," *Journal of the American Society for Information Science and Technology*, volume 57, number 2, pp. 233–243.

Ana Cleveland, Della Pan, Jiangping Chen, Xinyu Yu, Jodi Philbrick, Martin O'Neill and Lisa Smith, 2008. "Analysis of the health information needs and health related Internet usage of a Chinese population in the United States," *Journal of Library and Information Service*, volume 52, number 3, pp. 112–116.

John W. Creswell, 2009. *Research design: Qualitative, quantitative, and mixed methods approaches*. Third edition. Thousand Oaks, Calif.: Sage.

Debra Dudek, Anna Mastora, and Monica Landoni, 2007. "Is Google the answer? A study into usability of search engines," *Library Review*, volume 56, number 3, pp. 224–233.

Madelyn Flammia and Carol Saunders, 2007. "Language as power on the Internet," *Journal of the American Society for Information Science and Technology*, volume 58, number 12, pp. 1899–1903.

Fredric C. Gey, Noriko Kando, and Carol Peters, 2005. "Cross–language information retrieval: The way ahead," *Information Processing and Management*, volume 41, number 3, pp. 415–431.

Julio Gonzalo, 2002. "Scenarios for interactive cross–language retrieval systems," In: *Proceedings of SIGIR 2002 Workshop: Cross–Language Information Retrieval: A Research Roadmap*, at http://ucdata.berkeley.edu:7101/sigir-2002/sigir2002CLIR-13-gonzalo.pdf, accessed 12 November 2007.

Bing Liu, 2007. "Personal evaluations of search engines: Google, Yahoo! and MSN," at http://www.cs.uic.edu/~liub/searchEval/SearchEngineEvaluation.htm, accessed 30 September 2008.

Christopher D. Manning and Hinrich Schütze, 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Frank McCown, Johan Bollen, and Michael L. Nelson, 2005. "Evaluation of the NSDL and Google for obtaining pedagogical resources," *Proceedings of ECDL '05*, at http://www.cs.odu.edu/~mln/pubs/nsdl-google.pdf, accessed 26 February 2009.

National Institute of Standards and Technology (NIST), 2005. "NIST 2005 Machine Translation Evaluation Official Results," at http://www.nist.gov/speech/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html, accessed 26 February 2009.

Jian–Yun Nie and Fuji Ren, 1999. "Chinese information retrieval: Using characters or words?" *Information Processing & Management*, volume 35, number 4, pp. 443–463.

Greg R. Notess, 2008. "Multilingual searching: Search engine language tools," *Online*, volume 32, number 3, pp. 40–42, and at http://www.infotoday.com/ONLINE/may08/Notess.shtml, accessed 1 September 2008.

Douglas W. Oard, 1997. "Serving users in many languages: Cross–language information retrieval for digital libraries," *D–Lib Magazine*, volume 3 (December), at http://www.dlib.org/dlib/december97/oard/12oard.html, accessed 18 November 2007.

Douglas W. Oard and Anne R. Diekema, 1999. "Cross–language information retrieval," *Annual Review of Information Science and Technology*, volume 33, pp. 223–256.

Willam C. Ogden and Mark W. Davis, 2000. "Improving cross–language text retrieval with human interactions." *Proceedings of the Hawaii International Conference on System Science — HICSS–33*, at http://citeseer.ist.psu.edu/ogden00improving.html, accessed 10 November 2007.

Daniela Petrelli, Micheline Beaulieu, and Mark Sanderson, 2002. "User participation in CLIR research." *Proceedings of SIGIR 2002 Workshop: Cross–Language Information Retrieval: A*

Research Roadmap, at http://ucdata.berkeley.edu:7101/sigir-2002/sigir2002CLIR-12-petrelli.pdf, accessed 12 November 2007.

Jin Zhang and Suyu Lin, 2007. "Multiple language supports in search engines." *Online Information Review*, volume 31. number 4, pp. 516–532.

Editorial history

Paper received 19 December 2008; accepted 23 February 2009.

Cross–language search: The case of Google Language Tools
by Jiangping Chen and Yu Bao
First Monday, Volume 14, Number 3 - 2 March 2009
http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2335/2116