English-Chinese CLIR using a Simplified PIRCS System

K.L. Kwok, N. Dinstl and P. Deng
Computer Science Department, Queens College, CUNY
65-30 Kissena Blvd.
Flushing, N.Y. 11367

kwok@ir.cs.qc.edu

ABSTRACT
A GUI is presented with our PIRCS retrieval system for supporting English-Chinese cross language information retrieval. The query translation approach is employed using the LDC bilingual wordlist. Given an English query, different translation methods and their retrieval results can be demonstrated.

1. INTRODUCTION
The purpose of cross language information retrieval (CLIR) is to allow a user to search, retrieve, and gain some content understanding of documents written in a language different from the one that the user is familiar with. This is to be accomplished automatically without expert linguist assistance. CLIR is of growing importance because it can literally open up a whole world of information for the user, especially with the ease and convenience of access and delivery of foreign documents provided by Internet logistics nowadays. Searching and retrieving Chinese documents via English is a major sub-problem within CLIR because many people in the world use these two languages. For example, one would expect trade between China and the U.S. (and other countries) to grow significantly in the near future because of the impending WTO membership for China. Monitoring trends and status information from Chinese sources may be an essential operation for organizations interested in these affairs. Chinese is a language completely different from English, and it is conceived to be difficult for foreigners to learn. This paper describes some of the methods that we employ to deal with this problem, and presents a demonstrable system to illustrate the workings of cross language document retrieval. In Section 2, techniques for the query translation approach to CLIR are discussed. Section 3 contains a description of our simplified PIRCS retrieval system that is the basis for monolingual retrieval. Section 4 describes the GUI supporting interactive query input, document output and other implementation issues, and Section 5 contains our conclusion and future work.

2. STRATEGY FOR CROSS LANGUAGE INFORMATION RETRIEVAL
When faced with the situation of a language mismatch between the target documents and the query (information need statement) of a user, one could reduce them to a common representation language for retrieval purposes by automatically translating the query to the document language, by translating the documents to the query language, or by converting both to a third representation language [1]. By far the simplest and most common approach seems to be the first method, and probably as effective as the others, and we have also taken this route. The question is what tools to use for query translation.

It is well known that machine translation is generally fuzzy and inaccurate [6]. This is particularly true when translation output are judged by humans, who tend to be unforgiving. However, translation for machine consumption (such as for information retrieval (IR)) may not be so bad because IR can operate with a bag of content terms without grammar, coherence or readability. What IR needs is that important content terms are correctly covered, even at the expense of noise translations. For this purpose, we have combined two different methods of query translation to hedge for errors and improve coverage, viz. dictionary translation and MT software.

2.1 Translation Using LDC Bilingual Wordlist
One method we employ is dictionary translation using the LDC Chinese-English bilingual wordlist (www.morph.ldc.edu/Projects/Chinese) which we label as ldc2ce. It has about 120K entries. Each entry maps a Chinese character sequence (character, word or phrase) into one or more English explanation strings delimited with slashes. Sample entries are shown below:

1) 集会 /gather/assembly/meeting/convocation/
2) 部件 /parts/components/assembly/..
3) 礼堂 /assembly hall/auditorium/..
4) 议院 /legislative assembly/
5) 立法局 /legislative council/..

When an English word from a query is looked up in the ldc2ce wordlist, it will usually be mapped into many Chinese terms and reduction of the output is necessary. For this disambiguation purpose, we employ several methods in succession as tabulated below:

- Dictionary structure-based: ldc2ce format is employed to select the more correct mappings among word translations. For example, when the word to translate is 'assembly', we would pick line 1) and 2) only, rather than the additional 3) or 4) because in the latter two, 'assembly' appears in context with other words.
- Phrase-based: ld2ce can also be regarded as a phrase dictionary by matching query strings with English explanations of Chinese terms, giving much more accurate phrase translations. For example, if 'legislative assembly' appears in a query, it would match line 4) exactly and correctly, and would supersede all other single word translations such as those from lines 1), 2), 3) and 5).
- Corpus frequency-based: for single word translations with many candidates, those with higher occurrence frequency usually have higher probability of being correct.
- Weight-based: a Chinese term set translated for one English word can be considered as a synonym set, so that each individual Chinese term is weighted with the inverse of the sum of the collection frequencies, and generally gives more effective retrieval.

These dictionary disambiguation techniques have been implemented and tested with TREC collections. In general, they accumulatively lead to successively more accurate retrievals [4]. Their output can be demonstrated in our system.

2.2 Translation Using MT Software

COTS MT software for English to Chinese (or vice versa) are now quite readily available on the market. They cost from scores to about a thousand dollars for a single license. These software mostly operate on the PC Windows platform. Their codes are proprietary and usually do not come with an API. Interfacing them with a UNIX and C platform thus becomes quite difficult and perhaps impossible. However, if one runs retrieval from a Windows environment, one can 'cut and paste' from their translation results. We investigated several and found that one from Mainland China called HuaJian (www.atlan.com) performs quite well. A number of other such packages can also be demonstrated within our system.

Once an English query has been translated into Chinese, we can perform monolingual Chinese IR using our PIRCS system described in the next section. The two translation outcomes, from dictionary and MT software, can be combined for retrieval and the final result is usually more effective than single translation method alone [3].

3. A SIMPLIFIED PIRCS RETRIEVAL SYSTEM

PIRCS (Probabilistic Indexing and Retrieval – Components – System) is our in-house developed document retrieval system that has participated in all previous TREC large-scale blind retrieval experiments with consistently good results. It supports both English and Chinese languages. PIRCS retrieval approach is based on the probabilistic indexing and retrieval methods, but extended with the ability to account for the influence of term frequencies and item length of documents or queries. PIRCS can best be viewed as activation spreading in a three- layer network,

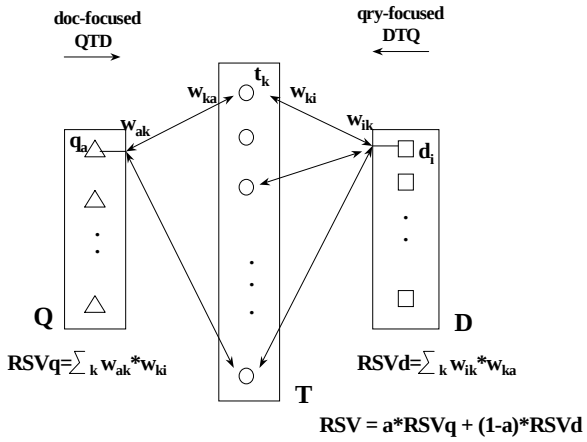

Figure 1. 3-Layer PIRCS Network

Figure 1, that also supports learning from user-judged or pseudo-relevant documents. The details of our model are given in [4, 5]. As shown in Figure 1, PIRCS treats queries and documents as objects of the same kind. They form a Q and a D layer of nodes connecting to the middle layer of term nodes. Retrieval means spreading activation from a query node via common term nodes to a document node and summed into a retrieval status value RSVd for ranking. This operation is gated by intervening edges with weights that are set according to the PIRCS model. An analogous operation is to spread activation from document to query nodes, resulting in another RSVq that has been shown to have similarity to a simple language model [2, 4]. The final retrieval status value RSV is a linear combination of the two.

Documents are pre-processed to create a direct file, an inverted file and a master dictionary that contains all the content terms and their usage statistics extracted from the collection. After appropriate processing, the master dictionary helps construct the middle layer T nodes of Figure 1. The direct file facilitates obtaining the terms and statistics contained in a given document, and helps construct the D node and D-to-T edges with weights. The inverted file facilitates obtaining the posting information of a given term, and helps construct the T-to-D edges with weights. At query time, a Q layer of one node is formed and the query terms are located on the T layer and linked in to define the Q-to-T and T-to-Q edges with weights.

Once the 3-layer network is defined, ranking of documents for the query is achieved by activation spreading Q-T-D realizing the document-focused retrieval status value RSVd, and vice versa for the query-focused RSVq. They are then linearly combined. This crosslingual PIRCS demonstration runs either on a SUN Solaris or Linux platform. The current implementation is a simplification of our batch PIRCS system and does not support automatic two-stage retrieval for pseudo-relevance feedback. However, users can interactively modify their queries to perform manual feedback.

4. GUI FOR INTERACTIVE CLIR

A simplified PIRCS system with first stage retrieval will be used for demonstrating English-Chinese CLIR. This system is based

on an applet-servlet model that runs on a UNIX operating system (such as Solaris or Linux). User's interaction with PIRCS is supported via a GUI based on the Netscape browser (Internet Explorer is a better browser for this GUI, but UNIX has Netscape only). The applet or HTML forms in the browser communicate with the servlet on the Apache server. The servlet works as a bridge between the front-end program (in HTML and applet) and the background programs that do the translation or retrieval. Based on the input from the user, it can dispatch calls to the retrieval system and then format the output and send results back to the applet or directly into user's browser through a customized applet-servlet channel via HTTP protocol.

A GUI software that is modeled on that of ZPRISE (www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html) but enhanced for CLIR will be demonstrated. The GUI supports five windows: one for English query input and editing, a translation window for displaying the Chinese terms mapped from the English query via the ldc2ce wordlist, a search-result window for displaying the fifty top-ranked document identities after retrieval, a document box for displaying the content of the current selected document in Chinese, and another index box showing the index terms used for retrieval together with their frequency statistics. This allows a user to do CLIR interactively.

If run in a Windows environment, the translation box also allows input and editing for those users who know some Chinese. In this test system, all Chinese are GB-encoded. A query of a few words currently takes a few seconds for translation and about 20 seconds or more for retrieval depending on the number of unique terms. This response time can be improved in the future.

A typical screen of the GUI is shown in Figure 2. A user starts by typing in an English query in free text. When the 'Convert to Chinese GB' button is clicked, translation via the LDC dictionary look-up based on a default (best) option will be displayed. Other options for translation such as using dictionary-structure only, add phrase matching, or include target collection frequency disambiguation, etc. (Section 2.1) can be chosen. If the user finds too many English words left un-translated, s/he can re-phrase the English query wordings and repeat the process. Otherwise, the retrieval button can be clicked and the top 50 document ID's will be displayed in the search-result box (below the translation) sorted by the retrieval status value shown next to each ID. Content of the top document is also displayed automatically in the large window with index terms high-lighted. Additional documents following the one displayed can also be brought in for browsing purposes.

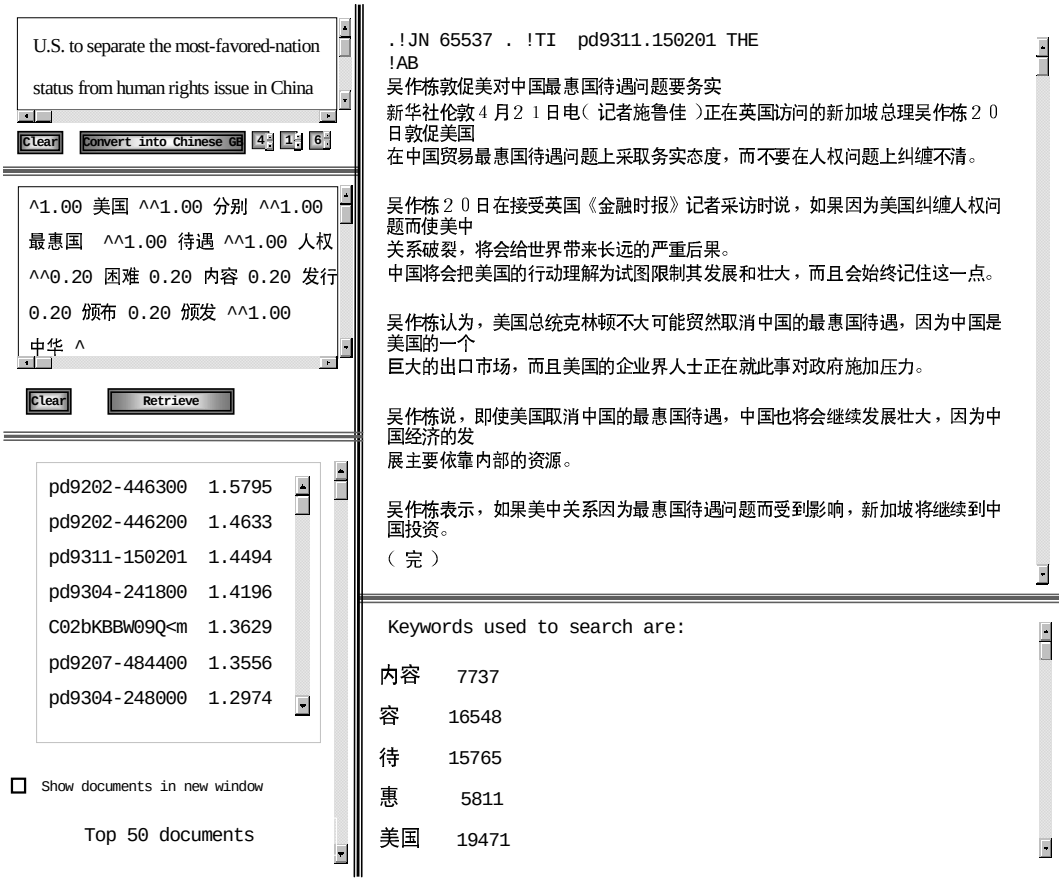

Figure 2. GUI for Cross Language Information Retrieval

If the user knows some Chinese, s/he can have more options for interaction. For example, the user can 'cut and paste' terms that s/he likes during perusal of the retrieved documents to do relevance feedback manually. In addition, the query index terms (in Chinese) and their document frequencies are also displayed at the right hand bottom of the screen. They can provide useful information about the query and can help the user make changes to it.

As discussed before, we also make use of COTS MT software for query translation. These can also be demonstrated separately. However, these packages are proprietary, run under Windows platform, and are not interfaced with our retrieval system that is Linux based. Another set-up that we can demonstrate is to use a Windows platform to run Internet Explorer that is also compatible with our GUI. Internet connection will have to be made to our home computers at Queens College. In this case, an MT software can be running in the background for query translation. The translation result can then be 'cut and paste' to the translation window of our GUI. Users can compare retrieval results based on our dictionary approach and the MT software. Alternatively, both translations can be combined to improve retrieval.

5. CONCLUSION AND FUTURE WORK

English-Chinese CLIR is an important topic in Human Language Technology and has great utility. This project demonstration combines simple translation with IR to provide a workable solution to CLIR. It is an ongoing project and eventually can help non-Chinese speaking users access Chinese text in a reasonable fashion. Our next step is to add capability to show gistings of a retrieved Chinese document in English to assist the user in understanding the document content. Faster machines and upgrading of the programs would also provide speedier response time.

6. ACKNOWLEDGMENTS

This work was partially supported by the Space and Naval Warfare Systems Center San Diego, under grant No. N66001-1-8912.

7. REFERENCES

[1] Grefenstette, G. *Cross Language Information Retrieval*. Kluwer, 1998.

[2] Hiemstra, D & Kraaj, W. Twenty-One at TREC-7:ad-hoc and cross language track. In: *Information Technology: The Seventh Text Retrieval Conference (TREC-7).* E.M.Voorhees & D.K. Harman, (eds.), NIST Special Publication 500-242, GPO: Washington, D.C, 227-238, 1999.

[3] Kwok, K.L, Grunfeld, L., Dinstl, N & Chan, M. TREC-9 cross-lingual, web and question-answering track experiments using PIRCS (Draft). Preliminary paper at TREC-9 Conference, Gaithersburg, MD, Nov, 2000.

[4] Kwok, K.L. Improving English and Chinese ad-hoc retrieval: a Tipster Text Phase 3 project report. Information Retrieval, 3(4):313-338, 2000.

[5] Kwok, K.L. A network approach to probabilistic information retrieval. ACM Transactions on Office Information System, 13:324-353, July 1995.

[6] Nirenburg, S, Carbonell, J, Tomita, M & Goodman, K. (Eds.) *MT: A Knowledge-Based Approach.* Morgan Kaufmann, 1994.