An Interactive Chinese to English Retrieval System

Yan Qu, Hongming Jin, David A. Evans
CLARITECH Corporation, 5301 Fifth Ave., Pittsburgh, PA 15232, USA
Email: {yqu,hongming,dae}@claritech.com

Abstract. In this paper, we describe an interactive Chinese to English retrieval system that takes in Chinese queries and retrieves relevant documents in Chinese and English from document collections of both languages. We first describe the components for Chinese language processing and for Chinese-to-English query translation. Then we present the complete interactive system, through which a user is able to select and edit query terms manually in both Chinese and English, review the retrieval results in different presentation modes, and conduct retrieval using enhanced queries based on relevant documents or document clusters in either language. The system integrates CLARIT advanced technologies of natural language processing and information management, and is intended as a prototype and evaluation environment for identifying effective strategies to facilitate users with different levels of language proficiency in retrieving relevant documents from multilingual data collections.

1. Introduction

With the increasing amount of online information in Chinese, Chinese information retrieval is gaining growing importance. Since English is the most widely used language in the world, much information is available in English texts. Thus the ability to retrieval relevant English documents using Chinese queries should also be considered as a potential necessary part of information access for Chinese users, who desire to identify or monitor the developments around the world.

In this paper, we describe an interactive Chinese-to-English retrieval system that takes in Chinese queries and retrieves relevant documents in Chinese and English from document collections of both languages. We first describe the components for Chinese language processing and for Chinese-to-English query translation. Then we present the complete interactive system with an emphasis on the interface, through which a user is able to select and edit query terms in both Chinese and English, review the retrieval results in different presentation modes (such as ranked lists and document clusters), and conduct retrieval using enhanced queries based on relevant documents or document clusters in either language. Finally, we briefly present some results on evaluations of various components of the system. The system integrates CLARIT advanced technologies of natural language processing and information management. It is intended to be a prototype for identifying effective strategies to facilitate users with different levels of document language proficiency in retrieving information from multilingual document collections.

2. System Components

Chinese information retrieval and Chinese-to-English information retrieval modules are developed based on a suite of natural language processing and information management tools from CLARITECH Corporation. The CLARIT toolkit provides advanced modules such as automatic indexing, vector-based retrieval, thesaurus term discovery, summarization, filtering, and document clustering (Evans and Lefferts, 1994; Milic-Frayling et al., 1998; Evans et al., 1999). In this section, we restrict our description to resources and modules for Chinese text retrieval and Chinese-to-English retrieval.

2.1 Chinese text retrieval

We adapted the CLARIT English retrieval system for Chinese text retrieval. An important feature of CLARIT retrieval is the use of concepts identified by robust linguistic analysis coupled with specialized statistical routines. This allows the system to work with the ideas in text. For Chinese text retrieval, we developed Chinese word segmentation and Chinese NLP so that linguistically meaningful text units can be extracted from unrestricted Chinese text.

Chinese word segmentation

Unlike written English where spaces are used as word delimiters, Chinese texts do not use spaces to mark word boundaries. As a result, given a string of Chinese characters, it is not orthographically clear where a word starts or ends. Various approaches have been proposed to address the Chinese word segmentation problem (Chen et al., 1997; Kwok, 1997). In our work, we follow the longest match word segmentation strategy employed by many systems. Specifically, we treat each Chinese character in the input as a single token, and use a morphological processor to group characters together to identify lexemes. Whether or not a sequence of characters is a lexeme is determined by (automatically) consulting lexicons, which define lexemic token sequences and their associated categories and root forms. The Chinese lexicon is described in more detail in the next subsection.

Chinese NLP

The Chinese NLP module supports analysis and identification of linguistically meaningful units from Chinese text. The major linguistic resources include lexicons and a parsing grammar.

The parsing grammar describes a set of lexical categories and a set of finite-state based parsing rules. The set of lexical categories includes major and minor syntactically based categories for various parts of speech. Major categories include adjectives, adverbs, nouns, verbs, etc. Minor categories distinguish subclasses of parts of speech, such as common nouns, proper nouns, etc. The categories can also indicate morphological variation, such as regular adjectives or reduplicated adjectives. The parsing rules are defined by lexical categories, which describe the syntax of Chinese, in particular, the structure of the noun phrases (NPs).

The Chinese lexicon was developed based on the LDC Chinese lexicon with 44,404 entries, enhanced by a set of wordlists from the Internet, and a collection of symbols and characters from the TREC Chinese corpus. The lexicon was manually cleaned and the entries were manually tagged with lexical categories for selected category classes. Due to time constraints, lexical tagging was concentrated on words of the closed-class categories (such as symbols and punctuation marks, auxiliary and modal verbs, conjunctions, prepositions, pronouns, and various types of functional particles). The version of the lexicon used for the experiments has a vocabulary of 136,570 entries including symbols and words.

Query: 世界妇女大会
(World Conference on Women)

Word segmentation output:
世界, 妇女, 大会

Phrase identified by CLARIT Chinese NLP:
(世界 妇女 大会)$_{NP}$

Figure 1: Example Output of Chinese NLP

In our model, Chinese word segmentation and Chinese NLP are closely integrated. The morphological processor that recognizes Chinese lexemes by consulting the Chinese lexicon assigns a lexical category to the token string when the word segmentation decision is made. The parser reads a stream of input lexemes received from the morphological processor, groups them into phrases, and adds delimiters to indicate the type and location of each word or phrase in the sentence. In the process, some lexemes such as those of the closed-class categories may be discarded. Figure 1 gives word segmentation output and parse output for the query string 世界妇女大会 (world conference on women). Note that the NLP module produces phrases as output rather than individual words.

2.2 Chinese-to-English Text Retrieval

In order to enable a user to query in one language but perform retrieval across multiple languages, we need to address one critical issue – how to bridge the language gap between the query language and the document language? Many resources have been exploited for crossing the language boundary between the query language and the document language, e.g., machine translation (Gachot et al., 1998; Oard and Hackett, 1998), machine-readable bilingual dictionaries or lexicons (Hull and Grefenstette, 1996; Ballesteros and Croft, 1998; Davis and Ogden, 1997), parallel or comparable corpora (Landauer and Litmann, 1990; Sheridan and Ballerini, 1996), and controlled languages (Deikema et al., 1999). In this work, we use bilingual lexicons for translating Chinese queries into English, because of the free availability of the online CEDICT dictionary (CEDICT, 1998).

Various types of linguistic resources are needed for bilingual lexicon-based cross-language information retrieval. Specifically, we need

1. Lexicons and grammars for the source language NLP;
2. Lexicons and grammars for the target language NLP;
3. Translation glossaries for term translation from the source language to the target language.

Linguistic resources in 1 and 2 are necessary to support monolingual NLP functionality, such as English NLP, Chinese NLP. We have already described Chinese NLP in section 2.1. CLARIT English NLP utilizes English lexicons and grammar to identify linguistic structures in English text. The interested reader is referred to Evans and Lefferts (1994) for details.

Translation Glossaries

We define translation glossaries as lexicons that only contain direct translations of the head words. Translation glossaries are necessary for term vector translation. The format we adopt is as follows:

Head word, translation-1, translation-2, translation-3, …

The availability of translation glossaries is very important for bilingual lexicon-based query translation. However, such resources are hard to come by. CEDICT is similar to many machine-readable bilingual dictionaries, which are developed mainly for human users. Machine-readable dictionaries present along with translations examples of usage and explanations, which are not needed for retrieval purposes. For example, in CEDICT, there is an entry:

"阿旺曲沛","Ngawang Choepel (Tibetan, Fullbright scholar)"

in which the translation of the name "阿旺曲沛" is contained in the translation entry "*Ngawang Choepel*

(Tibetan, Fullbright scholar)", in which the extra background informational terms *Tibetan, Fullbright scholar* appear. The expressions *Tibetan* and *Fullbright scholar* are part of an explanatory use, and should be identified as extraneous to the translations of the head word. Machine-readable dictionaries may also have the problem of missing word forms, due to lack of inclusion of morphological variants and different spelling convention.

We extracted from CEDICT translation glossaries for query translation. The automatic extraction and pre-processing processes include normalization of head words and translation entries, extraction of concept words from translation entries, and detection of term space discrepancies between monolingual lexicons and bilingual lexicons. Normalization of head words and translation entries is aimed at reducing the negative effect of missing word forms due to morphological variants and different spelling conventions. The extraction process is aimed at extracting direct translations from the translation entries. It should be noted that the extracted direct translations can be very noisy with extraneous translations. Therefore, we allow the generation of a text file of the glossary for human editing.

As a result of pre-processing, the head words and translations are normalized by their respective NLP modules. The term space discrepancies between the monolingual lexicons and the bilingual lexicons are detected; new words and phrases are detected and exported into a text file. Such a file can then be manually edited and used to enhance the original monolingual or bilingual lexicons.

Sample entries from the resulting Chinese-English translation glossary are as follows:
中国: China, Chinese
知识: intellectual, knowledge
产权: property right
政策: policy
立法: legislation
执法: law enforcement
情况: circumstance, state of affair, situation
知识产权: intellectual property right

It should be noted that morphological variations have been normalized to their root forms in translation glossaries. Therefore, we find "*state of affair*" instead of "*state of affairs*", and "*intellectual property right*" instead of "*intellectual property rights*".

Query Translation

During query translation, the system looks up each term in the Chinese query vector in the translation glossaries and builds a translated query vector by concatenating all translated terms. Figure 2 shows the generation of English query terms from a Chinese query. Line 5 in Figure 2 illustrates the result of word segmentation of the original Chinese query string. Line 6 shows the query terms generated by the CLARIT Chinese query processing module. The terms 有关 (concerning) and 以及 (and) do not appear as final query terms because they belong to the categories of preposition and conjunction, respectively, which are not generally considered for indexing purposes. The term 执法 (law enforcement) does not become a final query term because it was recognized as a verb by Chinese NLP, which was not selected for indexing in this example. Verbs can be included as indexing terms if so desired. Line 7 shows the English translations of the query terms generated by the CLIR query translation module. It is also possible that the target corpus filters out some terms if the terms do not appear in the corpus.

3. The Complete Interactive Retrieval System

The general framework for query translation based CLIR using bilingual lexicons is illustrated by Figure 3. The framework contains the following steps:

- Process the document collections using CLARIT NLP systems. Multilingual document collections can be parsed using their respective NLP systems and linguistic resources. As a result, words are normalized to their normal forms and phrases are identified.
- Use CLARIT indexing modules to create indexes of normalized terms (words and phrases) for the document collections.
- Pass the query string in the source language to the CLARIT NLP module, which makes use of source language lexicons and grammars. Each query word is morphologically analyzed and normalized to its root form; phrases are identified.
- During query processing, the query words or phrases are processed against some pre-indexed reference corpora (e.g., the test corpus). Statistics such as term frequencies and distributions are assigned to the query terms. In addition, the system supports various levels of term granularity, including single-word terms, attested sub-phrases, all possible sub-phrases, etc. The output is a vector of such terms with statistics.
- The system looks up each term of the query vector in the translation glossaries and builds a translated query vector by combining all the possible translations.
- The translated query vector together with the original query vector is sent to the CLARIT retrieval engine for similarity computation against the document indexes.
- Relevant documents are returned from the retrieval process.

In the case of the work reported here, the source language is Chinese, and the document languages are Chinese and English.

The GUI for the system was designed to be a testing environment for conducting and evaluating general-purpose interactive retrieval, for conducting TREC experiments, and for evaluating different presentation methods such as ranked list and document clustering. Within the interface, the user is able to access the source text, enter or edit queries, build or open an indexed database, cluster documents, produce summaries for documents, select documents or clusters for feedback retrieval, and set different options for these capabilities. In this section, we examine the following major features, including

- A query window for entering and editing query text, query term vectors, or constraints
- A retrieval result ranked list display window
- A clustering result display window
- Feedback options using either selected documents or clusters

These features are illustrated in Figures 4 and 5. In Figure 4, the main window has a query window to the left, where the user can type in a query string and optional constraints. The "Generate Query" button extracts CLARIT terms from the query string and displays them in a list box along with their coefficients, distributions, and frequencies. The user can edit term coefficients or remove terms before performing text retrieval.

When retrieval is set to be cross-language, each query term in the source language is also looked up in the bilingual lexicons for its translation equivalents. The resulting translations are parsed using target language NLP. The extracted terms in the target language are also included in the term vector list box, together with their coefficients, distributions, and frequencies.

The "New Search" button enables the user to perform text retrieval. This will bring up a retrieval window, displayed on the right side in Figure 4. In the retrieval window, the document titles appear in a scrolled list.

With the document titles, the user can mark relevance judgments through a pop-up menu. The user can open a document and view the text with the query terms highlighted. The "Augment Query" button prompts the system to perform automatic thesaurus extraction over document selections and suggest terms for enhancing the original query. After a set of documents has been retrieved, the "Augment Query" button in the query window will be enabled. The user can also press this button to prompt the system to perform automatic thesaurus extraction over the current retrieval results and provide term suggestions for query enhancement.

By pressing the "Cluster Results" button, the user can cluster the selected documents, or all retrieved documents if no document selection has been made. Clustering options related to feature selection methods, similarity methods, weighting methods, cluster display methods, etc. can be pre-selected. By default, clustering results are displayed in the form of a "forest" – a truncated or grouped hierarchy – as shown in the focused window in Figure 5. Each cluster title is composed of the first N terms of the cluster centroid, followed by its distribution and weight in parentheses. The number of documents within a cluster appears before the cluster title.

With a set of clusters, a user can edit the display in numerous ways. The user can view the representative terms in the clusters, view the documents in the clusters, delete, move and copy documents across clusters, and delete, merge, create or re-cluster clusters. In particular, for enhancing the original query, the user can use the "Augment Query" button to operate on selected clusters or selected documents within clusters that have been judged by the user as relevant. The button prompts the system to perform automatic thesaurus extraction over these selected documents or the documents in the selected clusters and suggest terms to enhance the original query.

In the case of multilingual databases, documents in the same language are clustered together. For example, in Figure 5, we obtained Chinese document clusters and English document clusters. Depending on the user's proficiency in the document languages, the usability of the clusters may differ. For example, a Chinese user who has no knowledge of English would choose to view the clusters and documents in Chinese while making relevance judgments. While a user who can read both languages may choose to view documents and clusters in both languages to make relevance judgments, and rely on the system to extract thesaurus terms automatically for expanding the original query.

4. System Evaluation

Various components of the system have been evaluated using large-scale TREC evaluations (cf. Voorhees and Harman, 1997). The CLARIT English retrieval performance has been evaluated in TREC English ad hoc tasks. Our efforts in the past few years have been targeted at identifying effective technologies to maximize a user's performance interactively. For example, Evans et al. (1999) conducted user studies in TREC-7 to compare the effectiveness of two modes of document organization supported by the system: ranked lists and document clusters. For each topic, users began their interactions by being presented with the initial query and the corresponding initial search results in, alternatively, ranked or clustered format. Users were instructed to identify as many relevant documents per topic as they could find within a given (brief) time period, and, along

the way, to mark any non-relevant documents that could be useful for negative feedback. Evans et al. (1999) observed that users who interacted with the system in cluster mode rendered more "positive" relevance judgments than users who interacted in ranked mode, and that overall retrieval performance by combining two modes of interaction outperforms either ranked-list or clusters used individually.

Chinese text retrieval has also been evaluated using TREC-5 Chinese track data. Chinese manual ad hoc runs produced average precision (N=1000 documents) above 0.31. Such retrieval performance is comparable with state-of-the-art monolingual retrieval systems that we are aware of (Voorhees and Harman, 1997).

We have not yet evaluated the usability of the interactive Chinese-to-English retrieval interface. Evaluation of an interactive CLIR system presents special problems that are not present in monolingual retrieval scenarios, and has to date not drawn much interest in the research community. It would be interesting to evaluate the effectiveness of feedback using relevance judgments by users with different levels of language proficiency in English, and it would be interesting to determine to what extent different modes of presentation matter for different types of users. We hope to investigate such issues in our future work.

5. Summary and Future Work

In this paper, we have described an interactive Chinese-to-English retrieval system that takes the user's query in Chinese and retrieves relevant documents in both Chinese and English. We have presented the interface that supports the user in making relevance judgments and in query refinement. The interactive system is an ideal test environment for investigating the contributions and limitations of monolingual retrieval techniques for users with different levels of language proficiency in the target language. In our future work, we would like to explore the following questions:

- When relevant documents from different languages are retrieved, how should the results be ranked in a combined list?
- When document clustering is used for presenting the retrieved documents, how should the clusters of different languages be ranked?
- In terms of retrieval performance, how effective is feedback based on relevance judgments by users with different levels of language proficiency?
- To what extend do different presentation modes matter to users with different levels of language proficiency?
- How do we support users with different levels of language proficiency?
- Should a CLIR system include machine translation of retrieved target-language documents?

References

Ballesteros, L., & Croft, W. B. Statistical methods for cross-language information retrieval. 1998. In G. Grefenstette (ed.), *Cross-Language Information Retrieval* (Chapter 3). Boston, MA: Kluwer Academic Publishers.

CEDICT. 1998. Chinese-English machine-readable dictionary, freely available from http://www.mindspring.com/~paul_denisowski/cedict.html. The dictionary development project was initiated by Paul Denisowski.

Chen, A., He,, J., Xu, L., Gey, F. C., & Meggs, J. 1997. Chinese text retrieval without using a dictionary. In *SIGIR'97*, pp. 42-49.

Davis, M., & Ogden, W. 1997. QUILT: Implementing a large-scale cross-language text retrieval system. In SIGIR'97, pp. 92-98.

Diekema, A., Oroumchian, F., Sheridan, P., & Liddy, E. D. 1999. TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In E. M. Voorhees & D. K. Harman (eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication 500-242, pp. 169-180. Washington, DC: U.S. Government Printing Office.

Evans, D. A., & Lefferts, R. G. 1994. Design and evaluation of the CLARIT-TREC-2 system. In D. K. Harman (ed.), *Information Technology: The Second Text REtrieval Conference (TREC-2)*. NIST Special Publication 500-215, pp. 137-150. Washington, DC: U.S. Government Printing Office.

Evans, D. A., Huettner, A., Tong, X., Jansen, P., & Bennett., J. 1999. Effectiveness of Clustering in Ad-Hoc Retrieval. In E. M. Voorhees & D. K. Harman (eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication 500-242, pp. 143-148. Washington, DC: U.S. Government Printing Office.

Gachot, D. A., Lange, E., & Yang., J. 1998. The SYSTRAN NLP browser: An application of machine translation technology in cross-language information retrieval. In G. Grefenstette (ed.), *Cross-Language Information Retrieval* (Chapter 9). Boston, MA: Kluwer Academic Publishers.

Hull, D. A., & Grefenstette., G. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49-57.

Kwok, K.L. 1997. Comparing Representations in Chinese Information Retrieval. In *SIGIR'97*. pp. 34-41.

Landauer, T. K., & Littman., M. L. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Conference of University of Waterloo Center for the New Oxford English Dictionary and Text Research*, pp. 31-38.

Milic-Frayling, N., Zhai, C., Tong, X., Jansen, P., & Evans., D. A. 1998. Experiments in query optimization, the CLARIT system TREC-6 report. In E. M. Voorhees & D. K. Harman (eds.), The Sixth Text REtrieval Conference (TREC-6), pp. 415-454. NIST Special Publication 500-240. Washington, DC: U.S. Government Printing Office.

Oard, D. W., & Hackett, P. 1998. Document translation for cross-language text retrieval at the University of Maryland. In E. M. Voorhees & D. K. Harman (eds.), *Information Technology: The Sixth Text REtrieval Conference (TREC-6)*, NIST Special Publication 500-240, pp. 687-696. Washington, DC: U.S. Government Printing Office.

Sheridan, P., & Ballerini, J. P. 1996. Experiments in multilingual information retrieval using the spider system. In *SIGIR'96*, pp. 58-65.

Voorhees, E. M., & D. K. Harman (eds.). 1997. *Information Technology: The Fifth Text REtrieval Conference (TREC-5).* NIST Special Publication 500-238. Washington, DC: U.S. Government Printing Office.

1. Original Chinese query: 中国有关知识产权政策立法以及执法情况
2. Manual Chinese query terms: 中国，知识，产权，政策，立法，执法
3. Manual English translation of Chinese query: Regulations and Enforcement of Intellectual Property Rights in China
4. Manual English translation of query terms: China, intellectual, property rights, policy, regulations, enforcement
5. Word segmentation of Chinese query using CLARIT Chinese NLP: 中国，有关，知识产权，政策，立法，以及，执法，情况
6. CLARIT generated Chinese query terms with corpus statistics:

Term	Coefficient	Distribution	Frequency
中国	1	376	1
知识	1	38	1
产权	1	24	1
政策	1	82	1
立法	1	19	1
情况	1	65	1
知识产权	1	20	1

7. CLARIT generated English translations of query terms with corpus statistics:

Term	Coefficient	Distribution	Frequency
Property right	1	3	2
Intellectual	1	31	1
Intellectual property	1	5	1
Intellectual property right	1	17	1
Circumstance	1	21	1
Knowledge	1	21	2
Property	1	59	2
Legislation	1	20	1
Situation	1	80	1
Affair	1	94	1
Policy	1	159	1
State	1	244	1
Right	1	271	2
Chinese	1	321	1
China	1	484	2

Figure 2: An Example of query translation

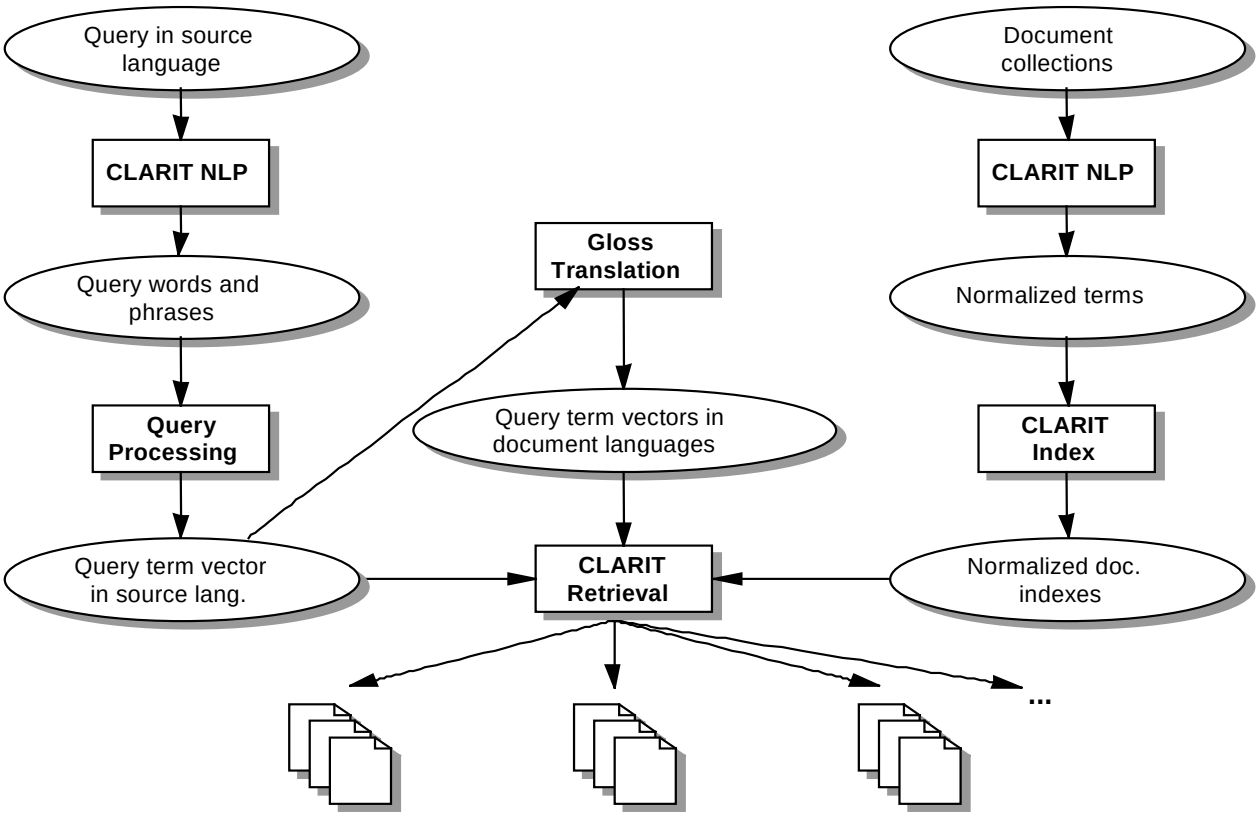

Figure 3:Diagram for CLIR using gloss-based query translation

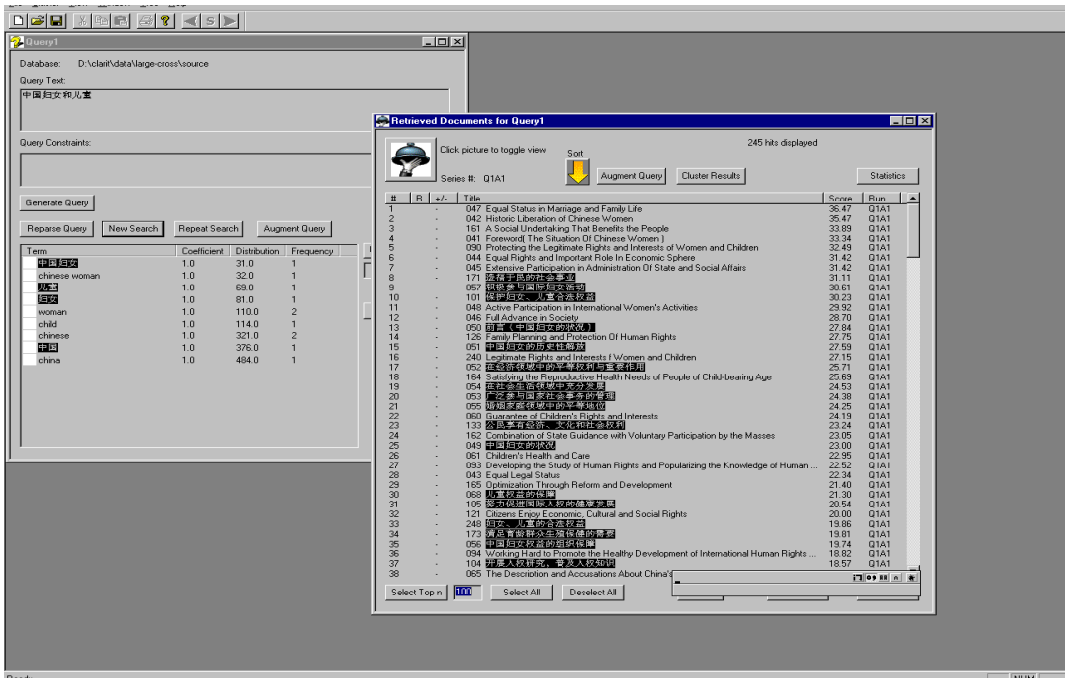

Figure 4: Query window and the retrieval result display window

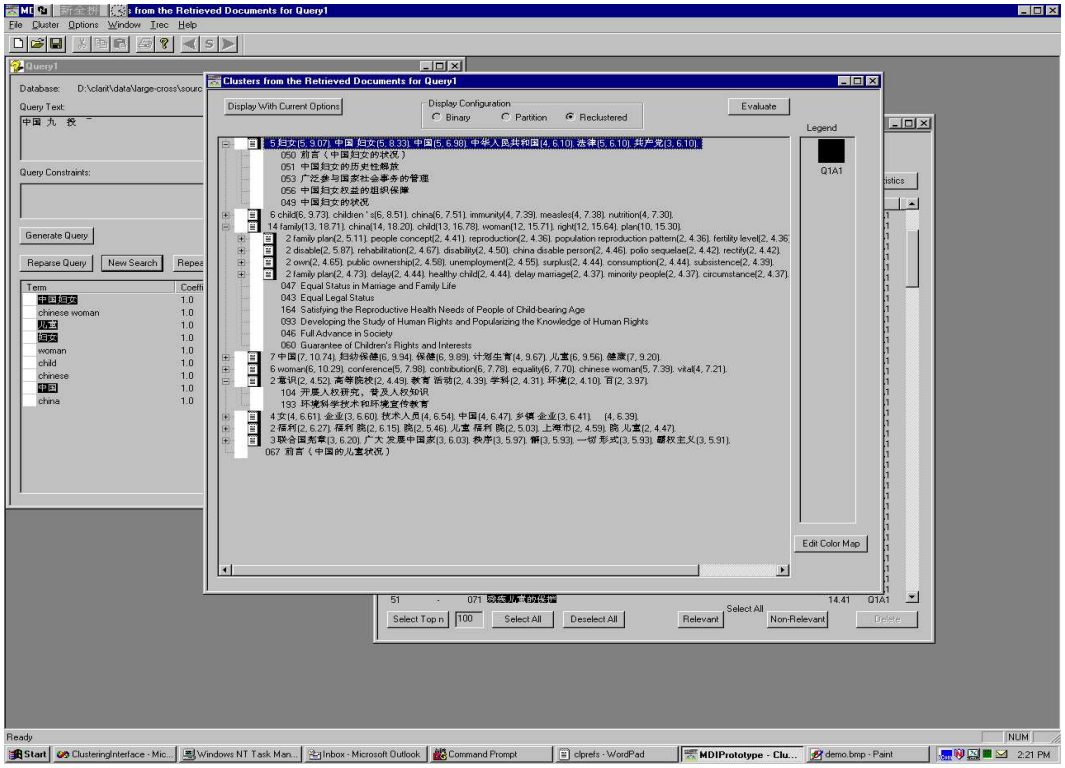

Figure 5: Query window and the cluster result display window