A Lexical Knowledge Base Approach for English–Chinese Cross-Language Information Retrieval

Jiangping Chen
School of Library and Information Sciences, University of North Texas, P.O. Box 311068, Denton, TX 76203.
E-mail: jpchen@unt.edu

This study proposes and explores a natural language processing- (NLP) based strategy to address out-of-dictionary and vocabulary mismatch problems in query translation based English–Chinese Cross-Language Information Retrieval (EC-CLIR). The strategy, named the *LKB approach*, is to construct a lexical knowledge base (LKB) and to use it for query translation. In this article, the author describes the LKB construction process, which customizes available translation resources based on the document collection of the EC-CLIR system. The evaluation shows that the LKB approach is very promising. It consistently increased the percentage of correct translations and decreased the percentage of missing translations in addition to effectively detecting the vocabulary gap between the document collection and the translation resource of the system. The comparative analysis of the top EC-CLIR results using the LKB and two other translation resources demonstrates that the LKB approach has produced significant improvement in EC-CLIR performance compared to performance using the original translation resource without customization. It has also achieved the same level of performance as a sophisticated machine translation system. The study concludes that the LKB approach has the potential to be an empirical model for developing real-world CLIR systems. Linguistic knowledge and NLP techniques, if appropriately used, can improve the effectiveness of English–Chinese cross-language information retrieval.

Introduction

Cross-language information retrieval (CLIR) provides users with access to information that is in a different language from their queries. English–Chinese cross-language information retrieval (EC-CLIR) enables English native speakers to search for Chinese text information using English queries. The research interest in EC-CLIR is growing rapidly with economic development in China and the availability of more and more Chinese text information on the Internet. The basic strategies for CLIR include first translating queries into the language of the documents (query translation) or translating the whole document collection into the language of users' queries (document translation), and then conducting monolingual information retrieval. Query translation strategy has been widely applied by most EC-CLIR experimental systems because of its simplicity and effectiveness. Query translation-based EC-CLIR systems utilize various knowledge resources, such as bilingual dictionaries, machine translation (MT) systems, parallel texts, or a combination of them to translate English queries into Chinese, and then conduct Chinese information retrieval.

Current research has found three major problems that negatively affect the performance of CLIR and EC-CLIR systems: translation ambiguity (Ballesteros & Croft, 1998), out-of-dictionary (Wu, Huang, Guo, Liu, & Zhang, 2001), and vocabulary mismatch (Gao et al., 2001). Translation ambiguity is generally caused by lexical ambiguity when a single English term can be translated into multiple Chinese terms. Out-of-dictionary happens when the translation resource of the system has incomplete coverage, which leads to the failure of translating query terms, especially new words, domain-specific compositional phrases, and proper names. Vocabulary mismatch refers to the situation where the translations in the lexicon are not the terms used in the collection or "bad translation of key concepts" (Gao et al., 2001). For example, "livestock" can be translated as "牲畜" or "畜牧业". Documents using the term "畜牧业" will not be retrieved for livestock if the dictionary only contains "牲畜" as the Chinese equivalent of that term. Much more research has been conducted to handle translation ambiguity than the other two problems. However, error analysis shows that the unsatisfactory EC-CLIR performance of many queries results from the existence of out-of-dictionary and vocabulary mismatch problems (Lee, Oh, Huang, Kim, & Choi, 2001; Gao et al., 2001). Furthermore, these two problems are more prominent for domain-specific EC-CLIR systems because numerous domain specific terms and proper names are unlikely to appear in lexical resources available to EC-CLIR systems. They require more effective solutions in either generic or domain-specific EC-CLIR systems.

Received July 8, 2004; revised December 20, 2004; accepted December 22, 2004.

 ● Published online 23 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20273

This study aims at addressing out-of-dictionary and vocabulary mismatch problems in query translation-based EC-CLIR. It appears that the major cause for these two problems is because most translation resources, such as bilingual dictionaries or MT systems, are constructed independent of the document collection of the EC-CLIR system. The vocabularies in the translation resources are often not the same as the documents the system is designed to search. In other words, there is a vocabulary gap between the document collection and the translation resources. As a result, important Chinese terms in the documents are missing (out-of-dictionary), or the Chinese equivalents in the dictionary are not the exact terms used in the documents, even though they are correct in meaning or used in other contexts (vocabulary mismatch). A natural language processing-(NLP) based strategy is therefore proposed to remove the vocabulary gap. The strategy, called the *LKB approach*, constitutes constructing a lexical knowledge base (LKB) and using it for query translation-based EC-CLIR. Natural language processing consists of a series of automatic techniques to achieve human-like understanding of various linguistic phenomena (Liddy, 1998). It has been successfully applied to many other tasks such as machine translation and automatic question answering. Various NLP techniques can be used in LKB construction and query translation process.

The remaining sections are organized as follows. In the next section, I summarize related research on translation knowledge customization and Chinese NLP, after which I describe the research plan and experimental design. The results and findings from the evaluation are then presented, followed by a discussion of the research questions and answers and the generalized CLIR system design model. In the Conclusion, directions for future research are given.

Related Research

This study explores translation knowledge customization based on the EC-CLIR document collection. Related research focusing on broader translation resource coverage includes semi-automatic lexicon development (Melamed, 1998), Web parallel text mining (Chen & Nie, 2000), and statistics-based methods for automatic dictionary construction from raw Chinese texts (Jin & Wong, 2001). Among them, only Web parallel mining has been directly used for CLIR and EC-CLIR.

The study also involves applying Chinese NLP techniques to annotate Chinese documents. Chinese text segmentation is usually the first task for other Chinese information processing tasks because there is no word boundary in Chinese text. The N-gram approach (Kwok, 1997) and word segmentation have both been applied to segment Chinese texts. But Chinese word segmentation is more popular with various strategies including statistical approaches (Sproat, Shih, Gale, & Chang, 1996; Xue & Shen, 2003), lexical- or dictionary-based approaches (Nie, Gao, Zhang, & Zhou, 2000), and hybrid approaches (Palmer, 1997; Goh, Asahara, & Matsumoto, 2004). Chinese information extraction is another major Chinese NLP task that attempts to identify named entities and noun phrases from texts. Chen (2002) provided a summary of various Chinese information extraction techniques. Many systems applied statistical approaches to Chinese just as to English texts (Yu, Bai, & Wu, 1998; Sun, Zhou, & Gao, 2003). Other systems utilized linguistic heuristics to identify special Chinese terms such as proper names in texts (Peterson, 1998) and temporal information (Li, Wong, & Yuan, 2001).

Methodology

The proposed LKB approach constitutes constructing a lexical knowledge base (LKB) by customizing available lexical resources based on the document collection of the EC-CLIR system, and employing the LKB for query translation. Two research questions were explored with regard to the LKB approach: (a) What are the effects of the LKB approach on the out-of-dictionary and vocabulary mismatch problems in query translation? (b) How does the LKB approach affect the performance of English–Chinese cross-language information retrieval?

A three-phase research plan was carried out to answer the research questions. The first phase was the construction of the LKB in which Chinese documents in a collection were extensively processed applying NLP techniques. In the second phase, the established LKB was employed to translate English queries into Chinese. In the third and final phase, the translated queries from phase two were used in EC-CLIR experimentation to assess the impact of the LKB on EC-CLIR performance.

In the context of this study, a lexical knowledge base is defined as a well-organized structure for holding information about lexical items and their associations and usages for a particular type of application. For example, WordNet (Miller, 1990) is considered as a lexical knowledge base as it is used for word sense disambiguation. An LKB can be built on the basis of one or more pre-existing LKBs for a particular purpose.

If not otherwise specified, the document collection mentioned in this study refers to any document collection of the EC-CLIR system for which the LKB is constructed. Different EC-CLIR systems may have the same or different document collections, and different document collections may use the same or different vocabulary sets, subject fields, or discourse structure. Therefore, an LKB constructed with one document collection can be a resource for another collection, but may need to be customized again to the content of the new collection to perform query translation in an EC-CLIR system that searches against that new collection.

Phase One: Lexical Knowledge Base Construction

The LKB construction process has gone through four stages: LKB design, linguistic resource selection, Chinese information extraction, and translation knowledge collection. The outcome of the first phase is a lexical knowledge base which can be used for EC-CLIR query translation.

DOI: 10.1002/asi

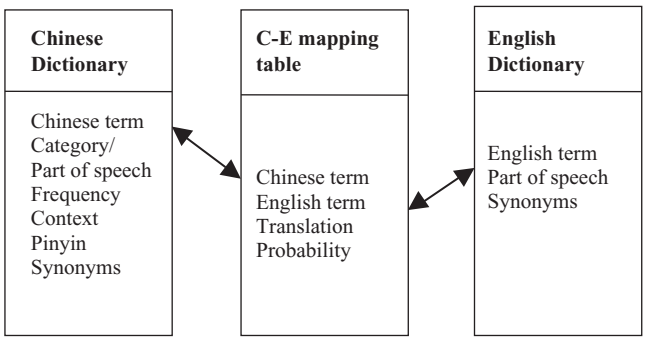

FIG. 1. The lexical knowledge base and the lexical information in the lexical knowledge base.

Lexical knowledge base design. Lexical knowledge base design aims to determine the structure and content, specifically, the components of the LKB and the types of lexical information that should be captured and stored for each component. The EC-CLIR literature was analyzed to derive the attributes of lexical items in the LKB. Fourteen systems that participated in the TREC-9 cross-language track (Voorhees & Harman, 2001) were selected as sample EC-CLIR systems for analysis. Based on the results, the LKB is designed to include three components: a Chinese dictionary, an English dictionary, and a Chinese–English (C–E) mapping table as shown in Figure 1. This simple structure has the advantages of (a) easy to use for translation, and (b) the flexibility of expansion.

As shown in Figure 1, each entry in the Chinese dictionary contains the following attributes where appropriate: Chinese term, category or part-of-speech, Pinyin (transliteration of written Chinese characters into the Roman alphabet), frequency, context, and synonym. For each Chinese term, its *category* or *part-of-speech* indicates whether this term is a Chinese character, a common Chinese word (noun, verb, adjective), or a Chinese named entity (a person's name, an organization's name, or a geographical name). The information will help in the translation knowledge collection stage to choose the appropriate indirect translation strategy as described later. For Chinese characters, their *Pinyin* representations can be used to transliterate Chinese person names and geographical names. *Frequency* refers to the number of occurrences of a term in the document collection. It can help the system to identify important Chinese terms. *Context* is defined here as one of the sentences in which the term appears. *Synonym* refers to the possible semantic alternatives of the term that may help in finding translations for terms that have no direct link with an entry in the English dictionary.

Lexical information for each entry in the English dictionary is simpler. Each entry contains the following three fields: English term, part-of-speech, and synonym. *Part-of-speech* is used for selecting correct translations, and *synonym* can help to find the Chinese equivalents for a term when there is no direct mapping between them.

The C–E mapping table serves as a bridge between the Chinese dictionary and the English dictionary. It contains three attributes: Chinese term, English term, and translation probability. *Translation probability* refers to the likelihood that the two terms are translations of each other. Its value can be assigned according to the mapping strategy described in the translation knowledge collection stage.

Linguistic resource selection. Linguistic resource selection is the second stage of LKB Construction. The research made use of three types of freely available lexical resources: (a) Dictionaries: including the LDC dictionary—a bilingual wordlist from the Linguistic Data Consortium, a Chinese character table, and a list of 3822 Chinese person names harvested from the Internet; (b) a Chinese-annotated corpus—the PKU corpus, and the document collection of the EC-CLIR system (The PKU corpus was downloaded from the Institute of Computational Linguistics, Beijing University; www.icl.pku.edu.cn.); and (c) WordNet (www.cogsci.princeton.edu/~wn/), the English ontology widely used in many NLP applications.

Chinese lexical information extraction. At the Chinese lexical information extraction stage, several NLP procedures such as Chinese word segmentation, part-of-speech tagging, and phrase and named entity categorization were performed. As discussed in the Related Research section, the major strategies for these tasks can be categorized into three types: statistical, lexical and/or rule-based, and hybrid. To avoid expensive computing and facilitate fast implementation, this study applied mainly lexical and rule-based strategies in combination with frequency information derived from the document collection. Four subprocedures were performed to ultimately extract Chinese lexical information from Chinese documents. First, the Chinese document collection was automatically segmented by applying a strategy combining both dictionary and collection frequency information. The segmentation dictionary was derived from the PKU corpus and the LDC dictionary; it included about 52,410 items. Second, segmented words were assigned parts-of-speech utilizing the above dictionary and linguistic rules identifying special categories, such as number, time, and English words. Third, noun phrases and certain named entities including person names, geographical locations, and organizations were identified using linguistic heuristic rules. Finally, important Chinese lexical terms along with their parts-of-speech or entity type, frequency, and context information were extracted and stored in the Chinese dictionary of the LKB. Table 1 presents sample lexical entries in the Chinese dictionary. As a result, 89,742 Chinese words and 22,827 noun phrases were extracted and stored in the Chinese dictionary. The process also identified 68,639 words which belong to measure words, function words, and words representing time, number, and mathematical symbols.

Translation knowledge collection. The final stage of the LKB construction is translation knowledge collection, which creates the Chinese–English mapping table by linking lexical items in the Chinese dictionary with their translation

DOI: 10.1002/asi

TABLE 1. Entries in the Chinese dictionary of the lexical knowledge base.

Term	Category/Part of speech	Frequency	Context	Synonyms
孟庄	ns	35 4	开拓"孟庄之路"	
生物工程	n	22 5	我国最早报道生物工程进展的两种刊物	风能#29816 水产业#48831
铁杉	n	5 3	这里是资江、漓江的发源地，有铁杉、冷杉等珍贵树种。	冷杉#57809
乌多文科	nr	7 16	乌多文科强调，乌克兰新政府的政策是基于国家的独立和稳定。	
破产	v	39 527	不少企业大量裁员，破产倒闭。	面临#4322 负#2557 兼并#26509 停产#4686

equivalent in the English dictionary. The English dictionary was composed of English terms collected from WordNet and the LDC dictionary with a total number of 211,528 terms including words and phrases. A computer program was developed to automatically carry out the mapping process. Two approaches were applied to collect translation knowledge. One is called *direct translation*, which performs a look-up of the LDC dictionary for translations; the other is called *indirect translation*, which employs linguistic knowledge or resources to estimate translations for a word or a phrase if the direct translation approach fails. The two approaches were used in three mapping procedures: mapping Chinese words to their English equivalents, mapping English words and/or phrases to their Chinese equivalents, and phrase translation.

The mapping of Chinese words to their English equivalents involved evaluating each Chinese word in the Chinese dictionary and determining its English equivalents. Strategies for locating the translations for Chinese words included (a) LDC dictionary look-up, and (b) transliteration using Pinyin. If the word was not found in the LDC dictionary and the program determined that the term was a Chinese person name or a geographical name, its translation was produced by putting together the Pinyins of the individual characters. For geographical names, such as 东门区 (Dongmen district), the strategy also considered a possible translation which combined Pinyins with the English equivalent of the last part of the word. The third strategy for locating the translations for Chinese words was decomposition and dictionary look-up. If the above two steps failed, the translation of this term was assumed to be that of its componential characters or words. A lower probability was assigned to the term indicating the translation was obtained indirectly. Figure 2 illustrates the mapping process. If the word still was not translated, it remained in the mapping table without a translation, and was recorded into a file called *No-translation* for possible postprocessing.

The mapping English terms to their Chinese equivalents was simple. Terms in the English dictionary and their translations were recorded in the mapping table if they were found in the LDC dictionary. Otherwise, WordNet was employed to find synonyms with the first word sense (synset). The translations for the synonyms, if any, were regarded as approximate translations for this term. Terms for which no translations were found were also recorded in the no-translation file for postprocessing.

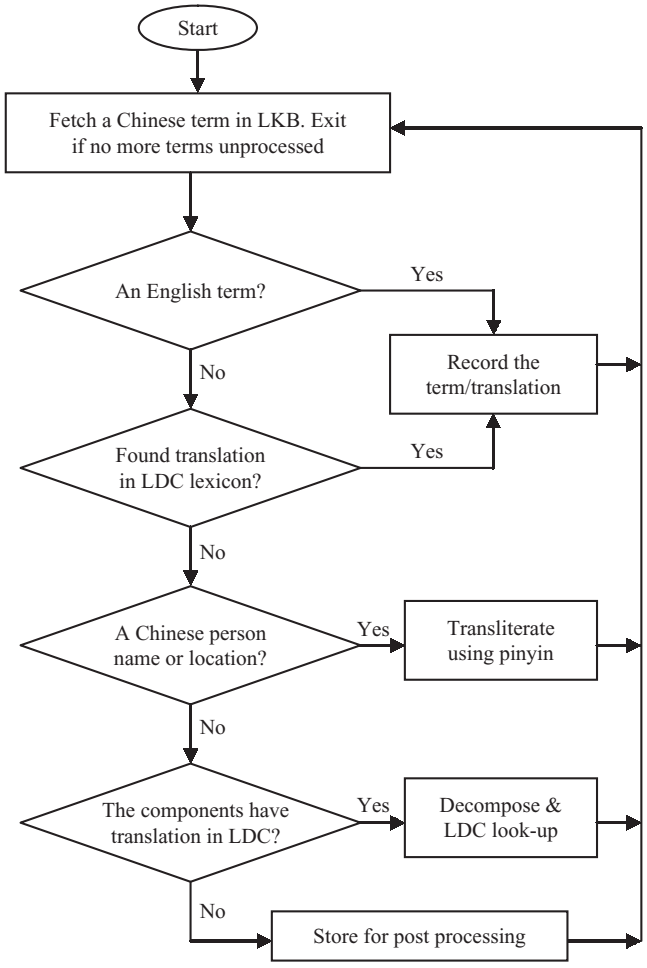

FIG. 2. Mapping Chinese terms to their English equivalents.

DOI: 10.1002/asi

TABLE 2. Results of translation knowledge collection.

Mapping steps	Total terms processed	Direct translation	Indirection translation	No translation
Chinese words to English	89,742	39,132 (43.6%)	28,931 (32.2%)	21,679 (24.2%)
English to Chinese	138,272	47,295 (34.2%)	26,680 (19.3%)	64, 297 (46.5%)
Chinese noun phrases to English	22,827	721 (3.1%)	15,927 (69.8%)	6,179 (27.1%)

The phrase translation procedure was similar to decomposition and dictionary look-up, as described above. The system first attempted to find the phrase as a whole in the LDC dictionary. If that was not successful, the program then searched for the translations of the individual words and put them together in different orders as the estimated translation for the phrase. This was also a type of *indirect translation*. Examples of the indirect phrase translation results were: 工程 技术 (engineering technology), 清 王朝 (Dynasty clear), 地方 铁路 (Local railroad), and 新闻 办公室 (news office). Note the indirect translation may produce false translations.

Table 2 summarizes the results of translation knowledge collection from the above three procedures. Among the 89,742 words in the Chinese dictionary, 39,132 or 43.6% had translations in the LDC dictionary; 28,931 or 32.2% were assigned translations using the indirect translation approach, including transliteration or componential translation; and 21,679 or 24.2% remained untranslated. The Chinese dictionary and English dictionary were both expanded because the indirect translation added translation equivalents obtained through the three procedures to the dictionaries accordingly. For instance, the transliterations for Chinese person names were added to the English dictionary, and the indirect translation for Chinese phrases were added to the Chinese dictionary.

Phase Two: Query Translation Using the Lexical Knowledge Base

This study took the query translation approach and the LKB was used as the translation resource in the query translation process. The translation process involved the following four steps. First, the English queries were processed by an NLP system developed at the Center for Natural Language Processing (CNLP) at the School of Information Studies, Syracuse University (Syracuse, New York). Each query term was stemmed and part-of-speech tagged. Second, the LKB mapping table was loaded and the system carried out a look-up for each word and phrase in the query. If a term or its stemmed form were found in the LKB mapping table, its Chinese translations along with the translation probability were obtained. The part-of-speech (POS) information of a query term was used to filter out translations that were not the same POS as the English term if this term had too many translations (larger than 50). Next, only m ($m = 10$ in the study) translations with highest probability were kept for each term. The dictionary look-up approach was similar to that of Kwok (1999). Terms that had high frequency (>300) in the mapping table were considered stop words and were filtered out. Finally, the Chinese translation results were ranked according to their frequencies in the document collection. The top N ($N = 1 - 3$ in the study) Chinese translations in the document collection were chosen to form the final query representation and sent to the Chinese information retrieval system.

To evaluate the results of query translation using the LKB comparatively, query translation using the LDC bilingual dictionary was conducted since the LDC dictionary is the only translation resource employed to construct the LKB. The study also used a third translation resource—the Huajian machine translation (MT) system for comparison purposes. The query translation results using Huajian were kindly provided by Professor Kwok (Queens College, City University of New York).

The outcomes of query translation using the LKB and the LDC dictionary were manually evaluated. The results using the Huajian system were not manually evaluated due to the difficulty in specifying translation(s) for each individual term. The evaluator was given the original bilingual texts of the 54 queries, the query translation results for experiments using the LKB and the LDC dictionary, and instructions about how to conduct the evaluation. The instructions asked the evaluator to classify each translation for each term into one of the four categories: correct, incorrect, missing translation, or unsure. Also, she was asked to identify and mark terms that were named entities and noun phrases. The query translation results were provided to the evaluator in a table format, which displayed the query number, the English query term, and its translations. The original bilingual texts of the queries were also provided to the evaluator to help her judge the translations of each term based on the context of the topic. The evaluation was completed in one week. At the end of the evaluation, the researcher met with the evaluator and went through the evaluation results. Any terms in the "unsure" category were discussed and reclassified to one of the first three categories: correct, incorrect, or missing translation. The results are reported in the Results and Findings section below.

Phase Three: English–Chinese Cross-Language Information Retrieval Experiments

An experimental EC-CLIR system has been built to measure the usefulness of the constructed lexical knowledge base (LKB). In addition to the LKB construction subsystem and the query translation subsystem described above, the EC-CLIR system contains a Chinese information retrieval (CIR) subsystem that retrieves relevant documents from the collection for each test topic.

DOI: 10.1002/asi

The CIR subsystem applies the traditional vector-space IR model (Salton & McGill, 1983) and the *Lnu.ltu* weighting scheme (Singhal, Buckley, & Mitra, 1996). The indexing approach was Chinese short terms combining non-stopword Chinese characters, which was similar to Kwok and Grunfeld (1997). The same procedures as described to extract Chinese lexical information in the LKB construction phase segmented the Chinese texts and extracted Chinese words, phrases, and named entities as index terms. Terms longer than three characters were converted into overlapping shorter ones or individual characters and used as index terms as well.

The test collection for the experiments is the combination of the ones used in the TREC-5 and TREC-6 Chinese Track, which includes a document collection, 54 topics in both Chinese and English, and relevance judgments. Each of the 54 topics contains three portions: title, description, and narrative in both Chinese and English. Figure 3 shows an example of the TREC-6 topics.

The major goal of the EC-CLIR experiments was to evaluate the effects of the LKB. As described before, two other translation resources, the LDC dictionary and Huajian machine translation (MT) system were used in comparison with the LKB. This study applied average precision to measure the performance for all conducted experiments. The results are reported in the next section.

Results and Findings

Information Retrieval

Four sets of information retrieval experiments were conducted using the experimental EC-CLIR system built by the researcher. Among them, one set was monolingual Chinese information retrieval whose performance would serve as the baseline. The other three were CLIR experiments using the LKB, the LDC dictionary, and the Huajian MT system, respectively.

Chinese monolingual retrieval experiments were conducted applying different portions of the 54 TREC queries. The run (a *run* is an execution of the retrieval program) using all the portions achieved a higher mean average precision (MAP) score (0.4171) than the other three runs. The score was significantly higher than the runs using either the title portion (0.3181) or the description portion (0.3618) portion in terms of the nonparametric test—the paired Wilcoxon signed ranks test (Conover, 1999), the *p-values* were both less than 0.001. But that run was not significantly better than the run using both title and description portions (0.3962).

A number of EC-CLIR experiments using the LKB were conducted to determine the best strategy for LKB usage. Corresponding to the monolingual runs, different portions of the same TREC topics were used to constitute the queries. Also, different numbers of translations for each query term were attempted for each situation. The identifier of each run reflects these choices. For example, "lkb_tdn_1" identifies the run that used the LKB as the translation resource, all three portions of a topic, and the one top ranked translation for each query term. Altogether, four sets, which are 12 runs, were carried out. Their results are summarized in Table 3 in which their run tags start with "lkb." Table 3 shows that lkb_tdn_1 achieved the highest MAP score among runs using the LKB. Notice that bringing more translations into the query as terms actually hurts the system performance, which conflicts with findings in other research (Kwok, 2000; Xu & Weischedel, 2001). This was a result of the problematic translation

```
<top>
<num> Number: CH29
<C-title> 信息高速公路的建设
<E-title> Building the Information Super Highway

<C-desc> Description:
信息高速公路，建设
<E-desc> Description:
Information Super Highway, building

<C-narr> Narrative:
相关文件应提到信息高速公路的建设，包括任何技术上的，或与信息基础设施有关的问题，
以及有关发达国家或发展中国家对国际网络的应用计划.

<E-narr> Narrative:
A relevant document should discuss  building the Information Super Highway, including
any technical problems, problems with the  information infrastructure, or plans for
use of the Internet by developed or developing countries.
</top>
```

FIG. 3. A TREC-6 topic.

DOI: 10.1002/asi

TABLE 3. English–Chinese cross-language information retrieval results.

Runs using LKB	Mean average precision	R-Precision	Runs using LDC dict	Mean average precision	R-Precision	Runs using Huajian	Mean average precision	R-Precision
lkb_t_1	0.161	0.2077	ldc_t_1	0.1461	0.1937	mt_t	0.25	0.2837
lkb_d_1	0.2301	0.2729	ldc_d_1	0.1703	0.2179	mt_d	0.2443	0.2756
lkb_td_1	0.2527	0.2963	ldc_td_1	0.2066	0.2554	mt_td	0.2928	0.319
lkb_tdn_1	0.2825	0.3265	ldc_tdn_1	0.2466	0.2934	mt_tdn	0.3062	0.3426
lkb_t_2	0.1511	0.1943	ldc_t_2	0.1478	0.1881			
lkb_d_2	0.218	0.262	ldc_d_2	0.1689	0.2238			
lkb_td_2	0.2467	0.2928	ldc_td_2	0.2082	0.2637			
lkb_tdn_2	0.2686	0.3171	ldc_tdn_2	0.2389	0.2851			
lkb_t_3	0.1452	0.1969	ldc_t_3	0.1313	0.1791			
lkb_d_3	0.2087	0.258	ldc_d_3	0.1569	0.215			
lkb_td_3	0.2375	0.2938	ldc_td_3	0.1912	0.2518			
lkb_tdn_3	0.2509	0.3033	ldc_tdn_3	0.2104	0.2728			

disambiguation strategy of the experimental system, which is one of the limitations of this study.

Table 3 also presents the EC-CLIR results of using the LDC dictionary for query translation. Their run tags start with "ldc." In total, 12 runs were conducted. It is observed that runs using the LKB have achieved higher MAP scores than the runs with the same query construction (applying the same portion(s) of the test topics and the same number of top translations) using the LDC dictionary. The EC-CLIR experiments using the Huajian MT system are also recorded in Table 3. Their run tags start with "mt." Runs using the Huajian MT system did not differentiate the number of translations because the translation results were not in word-by-word format.

To have a manageable set of data, one run from each of the four sets of IR experiments was selected for further analysis. The four runs were (a) "mono_tdn," Chinese monolingual IR; (b) "lkb_tdn_1," EC-CLIR using the LKB; (c) "ldc_tdn_1," EC-CLIR using the LDC dictionary; and (d) "mt_tdn," EC-CLIR using the Huajian MT system. They were the four top runs that received the highest MAP scores in their set of experiments. Table 4 presents the statistics of these four top runs in terms of MAP scores. The numbers in parentheses stand for the percentage of the MAP as compared to that of run mono_tdn. The results show that EC-CLIR using any of the translation resources received lower MAP scores than the monolingual information retrieval, as depicted in Figure 4. The results were consistent with those of other CLIR systems (Gao et al., 2001; Ruiz, Rowe, Forrester, & Sheridan, 2001).

As to the three EC-CLIR runs, Table 4 shows mt_tdn using the Huajian MT system received higher MAP than runs using the LKB (lkb_tdn_1) or the LDC dictionary (ldc_tdn_1). However, significance testing using the paired Wilcoxon test (Conover, 1999; Hull, 1993) showed that the difference in CLIR performance between the Huajian system and the LKB was not significant, while the performance using the LKB and the LDC dictionary was significantly different.

An examination of the average precision scores of individual topics shows that lkb_tdn_1 has achieved higher average precision than ldc_tdn_1 for the majority topics (38/54). Figure 5 illustrates the differences among the individual topics

TABLE 4. The top monolingual and cross-lingual information retrieval runs.

	mono_tdn	*mt_tdn*	*lkb_tdn_1*	*ldc_tdn_1*
Number of retrieved relevant documents	4427	3592	3702	3443
Mean average precision	0.4174	0.3062 (73.4%)	0.2825 (67.7%)	0.2466 (59.1%)
Median	0.4057	0.2549	0.2765	0.1940
SD	0.218	0.2191	0.2049	0.1967
Range	0.7777	0.7575	0.8142	0.7921
Minimum	0.0414	0.003	0.008	0.001
Maximum	0.8191	0.7605	0.8227	0.7933
p-value		0.425		<0.001

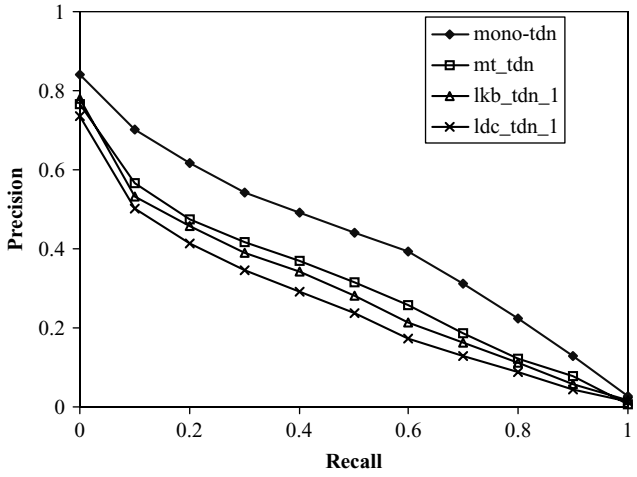

FIG. 4. P-R Curves of the four top runs.

DOI: 10.1002/asi

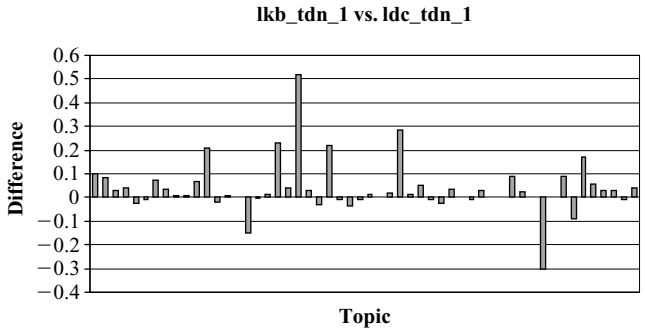

FIG. 5. Difference of lkb_tdn_1 from ldc_tdn_1 in average precision per topic.

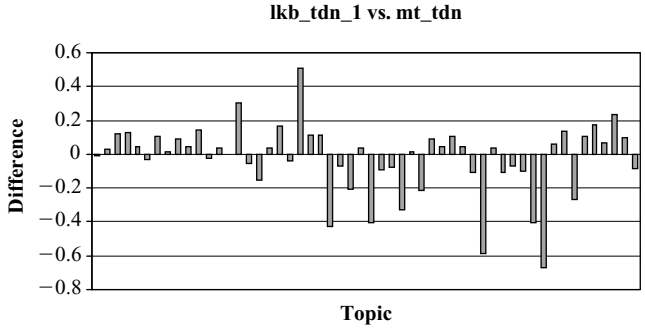

FIG. 6. Difference of lkb_tdn_1 from mt_tdn in average precision per topic.

between these two runs. The differences between lkb_tdn_1 and mt_tdn are depicted in Figure 6 in which mt_tdn has achieved much higher average precision scores for about six topics, which leads to a higher MAP score.

Query Translation

The three selected EC-CLIR runs (lkb_tdn_1, mt_tdn, and ldc_tdn_1) employed different translation resources using the same test collection. Query translation is considered the major cause of the differences in IR performance. The manual evaluation of the translation results is expected to provide insights to the process.

Table 5 summarizes the manual evaluation results on query translation of runs lkb_tdn_1 and ldc_tdn_1 in terms of frequency and relative frequency. Run lkb_tdn_1 translated more terms correctly (78.3%) than ldc_tdn_1 (73.6%), largely due to the increase in correct translations of noun phrases. Also, lkb_tdn_1 had fewer terms translated incorrectly or missing a translation. The Wilcoxon signed ranks test results demonstrates that the differences between the two runs on correct translations and missing translation are statistically significant. The *p-values* are both less than 0.001. The difference between incorrect translations is not significant.

TABLE 5. Summary of the evaluation of the query translation results.

	lkb_tdn_1	*ldc_tdn_1*
Total terms evaluated	1538	1610
Correct translation	1204 (78.3%)	1185 (73.6%)
Named entities	197	191
Noun phrases	84	39
Other terms	923	955
Incorrect translation	260 (16.9%)	282 (17.5%)
Named entities	10	14
Noun phrases	10	2
Other terms	240	264
Missing translation	74 (4.8%)	143 (8.9%)
Named entities	12	24
Other terms	62	119

The correct translation of named entities and phrases has been considered important for CLIR and EC-CLIR (Ballesteros & Croft, 1997; Gao et al., 2001). Table 5 shows that lkb_tdn_1 has fewer cases of missing translations for named entities than does ldc_tdn_1. Also, it correctly translates 84 noun phrases, nearly twice as many as correctly translated by the LDC dictionary in ldc_tdn_1. Examples of the new named entities and phrases that have been correctly translated by the LKB but have not been found in the LDC dictionary are listed in Table 6. The majority of new named entities in Table 6 are Chinese geographic names or person names that are expressed in Pinyin. The successful transliteration of these names resulted from the translation knowledge collection stage in LKB construction phase. In general, the correct translation of some important conceptual terms of some queries greatly improves the EC-CLIR performance.

Short queries. This study also examined the effects of the LKB on short queries that were formulated using only the title portion or description portion of the test topics, which were called *title-only queries* and *description-only queries*, respectively. These short queries might be more similar to real-world user queries (Crouch, Crouch, Chen, & Holtz, 2002) even though they might not fully represent users' information need.

The EC-CLIR runs of these short queries received lower MAP scores than the monolingual runs. Among the EC-CLIR runs, those using the Huajian MT system received the highest MAP scores. The runs using the LKB received the second highest MAP scores. Results of Wilcoxon signed ranks tests show that the difference between EC-CLIR using the Huajian MT system and that using the LKB is statistically significant at $\alpha = 0.05$ level for title-only queries, but is not significant for description-only queries. The difference between EC-CLIR performance using the LKB and that using the LDC dictionary is not significant for title-only queries, but is significant for description-only queries.

What caused the difference between using the LKB for title-only queries and description-only queries? The description portions have different characteristics from the title portions. The former are longer; they contain more missing phrases and named entities. The LKB approach provided better translations to both types of queries but the effect on description-only queries was bigger than on title-only queries. The Huajian system provided better translations to title portions than the LKB and the LDC dictionary.

DOI: 10.1002/asi

TABLE 6. Sample of correctly translated named entities and phrases by the lexical knowledge base.

Named entity in English	Chinese translation	English phrases	Chinese translation
Daya Bay	大亚湾	Most favored nation	最惠国
Qinshan	秦山	Accident reports	事故报告
Xisha	西沙	Territorial dispute	领土纠纷
PRC	中国	Economic situations	经济形势
Peng dingkang	彭定康	Economic strength	经济实力
Haihe	海河	Concrete measures	具体措施
Liaohe	辽河	Project Hope	希望工程
Songhua River	松花江	U.S. military	美国军事
Huaihe	淮河	Vietnamese government	越南政府

Discussion

Effects of the Lexical Knowledge Base Approach on Out-of-Dictionary and Vocabulary Mismatch Problems

As defined in the Introduction section, out-of-dictionary refers to the existence of new words and phrases that are not covered by the translation resource. To evaluate the effects of the LKB approach on the out-of-dictionary problem, two sets of query terms (words and phrases) were inspected: (a) words that had no translation in the LDC dictionary, but were correctly translated by the LKB, and (b) phrases that were correctly translated by the LKB, but were not included in the LDC dictionary. These two sets of terms represent the possible out-of-dictionary query terms. There were 143 words that had no translation in the LDC dictionary and actual out-of-dictionary query words were 140 (three words lost their translations in the process of translation disambiguation). Among the 140 words, the LKB correctly translated 59 or 42% of them; 72 or 51% remained untranslated; and 9 words were incorrectly translated. As for phrase translation, 36 new phrases were correctly translated by the LKB in lkb_tdn_1. Examples of the phrases are presented in Table 6. These phrases were distributed over 22 queries. Altogether, there were 95 out-of-dictionary query terms correctly translated by the LKB. These 95 terms were distributed over 41 test queries of which 30 received higher average precision from lkb_tdn_1 as compared to ldc_tdn_1.

Vocabulary mismatch refers to the situation in which the translations of some words or phrases in the translation resource do not match the right terms in the document collection of the EC-CLIR system. Two types of terms might be vocabulary mismatch terms: (a) words that had no translation, and (b) terms that were not correctly translated. No vocabulary mismatch occurred among the words that had no translation. The reason for this may be the large size and the open domain nature of the document collection. Thirty terms were incorrectly translated by either the LKB or the LDC dictionary, but not by both. Among them, eight received incorrect translation due to translation disambiguation. Only 22 terms were vocabulary mismatch cases, and the LKB correctly translated 18 of them.

To summarize, the LKB approach significantly increased the percentage of terms that were correctly translated and decreased the percentage of terms that had no translation. The approach provided effective solutions to translating phrases and new named entities. The correctly translated named entities were mainly Chinese person names and geographic names. The effect of the LKB on vocabulary mismatch was rather limited although it was positive.

Additionally, the LKB approach identified Chinese terms in the document collection that had no English translation and English terms that had no Chinese equivalents at the translation knowledge collection stage. These terms were stored in two no-translation files, which made it possible for a real-world EC-CLIR system to employ either human efforts or system search of other translation resources to provide translations for some or all of these terms.

Effects of the Lexical Knowledge Base Approach on English–Chinese Cross-Language Information Retrieval Performance

Three sets of EC-CLIR experiments were conducted using the LKB, the LDC dictionary, and the Huajian MT system, respectively. Based on the analysis of the top runs using each of the three translation resources, EC-CLIR performance using the LKB was significantly better than that using the LDC dictionary. The experiment using the LKB achieved higher average precision scores for 70% (38/54) of the queries. The majority of the 38 queries benefited from correct translation of the important terms using the LKB. The difference between EC-CLIR performance using the LKB and the Huajian MT system was not significant. In other words, the LKB approach has achieved the same level of EC-CLIR performance as a sophisticated machine translation system.

This study also found that the constructed LKB had a limited effect on title-only queries. However, it produced significant improvement to EC-CLIR performance on description-only queries. The LKB approach provided better translations to description-only queries because of the effective mapping approaches employed in the LKB construction phase.

DOI: 10.1002/asi

The Lexical Knowledge Base Approach as a Model for Cross-Language Information Retrieval Systems Design and Development

The LKB approach explored in this study is essentially a customization approach which builds the translation resource for CLIR systems by adapting available translation resources to the targeted document collection. It provides a practical model for the design and development of CLIR systems, especially real-world CLIR systems that target specific groups of users such as researchers in R&D departments within a corporation or an organization. The real-world CLIR systems, in general, have subject-specific document collections and user requirements. Very possibly, there are no well-matched electronic bilingual dictionaries or MT systems for use within the CLIR system even though it is not very difficult to find some useful translation resources in related fields or for general purposes. The LKB approach can be employed in this context to construct a useful translation resource for the system. A CLIR system model based on the LKB approach, or we call it the LKB model, can include the following components:

1. The initial investigation of available translation resources.

   The translation resource constitutes one of the most important components of a CLIR system. The initial investigation involves the gathering and selection of appropriate available translation resources for customization, and an exploration on whether the translation resource matches the document collection. One possible strategy is to employ NLP techniques to extract the important terms in the document collection and compare them with those contained in the available translation resources.
2. Lexical knowledge base construction.

   If the available translation resources do not match the document collection, efforts will be made to build a lexical knowledge base by customizing available translation resources based on the document collection. The purpose is to ensure that terms in the document collection have accurate and complete translations in the other language. Depending on the specific domains and language pairs, the strategies employed in LKB construction may be quite different from what we have used in this study. Furthermore, in addition to the automatic approach for information extraction and translation knowledge collection, human effort may be needed to find translations for terms that are considered important in the document collection but have no translation in the available resources or cannot be found with automatic approaches.
3. Query translation using the constructed lexical knowledge base.

   Query translation using the LKB includes two processes. The first is to preprocess the original queries to ensure that the query terms are in the same format as in the LKB. The second process is to use the LKB for translating phrases and words in combination with an effective translation disambiguation strategy.

The LKB model has the potential for application to the design and development of real-world or large-scale cross-language retrieval systems. This study is the first step toward the thorough investigation and wide application of the LKB model. The LKB model will be continually evaluated and enriched to solve problems in real-world EC-CLIR and CLIR systems.

Summary and Future Research

The LKB approach proposed in this study aims to bridge the vocabulary gap between the document collections and the translation resources used by CLIR and EC-CLIR systems. The experimental results demonstrate that the LKB approach is very promising and has the potential to serve as a design model for developing CLIR systems where the appropriate translation resource is unknown. This study has investigated the general procedures for constructing the lexical knowledge base and using it for an EC-CLIR system. It can serve as a reference for developing effective EC-CLIR and CLIR systems.

Future research can be carried out in two directions. One is to investigate further improvements of the LKB approach, which will mainly focus on two aspects. The first is an effective strategy for finding translations for important terms in the no-translation files. Literature has shown that the Web is a valuable resource for collecting translation knowledge (Chen & Nie, 2000). A thorough exploration of the role of the Web in the context of the LKB model and other ways to collect translation knowledge will be an interesting topic for future research. The second focus in approving the LKB approach entails an investigation of an appropriate translation disambiguation strategy for the LKB model. Future research should also explore the application of the LKB model, as described in the previous section, to domain-specific, real-world EC-CLIR systems, as well as CLIR systems of other language pairs. The research will provide more evidence on the effectiveness of the LKB model and improve our understanding of the CLIR problems and solutions.

Acknowledgments

The author would like to thank Elizabeth D. Liddy, Barbara Kwasnik, Jian Qin, Kui-Lam Kwok, and Anne Diekema for their valuable advice and statistical assistance on this study. The research is supported by a 2003 ASIST/ISI dissertation proposal scholarship.

References

Ballesteros, L., & Croft, W.B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA. Retrieved October 4, 2005, from http://www.ee.umd.edu/medlab/filter/sss/papers/ballesteros.ps

Ballesteros, L., & Croft, W.B. (1998). Resolving ambiguity for cross-language retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 64–71). New York: ACM.

Chen, H.H. (2002). Chinese information extraction techniques. Lecture Notes at the 2002 Summer School on Intelligent Media and Information

DOI: 10.1002/asi

Processing (SSIMIP-2002), National University of Singapore, Singapore. Retrieved October 4, 2005, from http://www.comp.nus.edu.sg/~pris/summer_school/materials/hhchen-cie.pdf

Chen, J., & Nie, J.Y. (2000). Automatic construction of parallel English–Chinese corpus for cross-language information retrieval. In Proceedings of the 6th Conference on Applied Natural Language Processing (pp. 21–28).

Conover, W.J. (1999). Practical nonparametric statistics (3rd ed.). New York: Wiley.

Crouch, C.J., Crouch, D.B., Chen, Q., & Holtz, S.J. (2002). Improving the retrieval effectiveness of very short queries. Information Processing and Managements, 38, 1–36.

Gao, J., Nie, J.Y., Zhang, J., Xun, E., Su, Y., Zhou, M., et al. (2001). TREC-9 CLIR experiments at MSRCN. NIST Special Publication 500–249: The Ninth Text REtrieval Conference (pp. 343–353). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Goh, C., Asahara, M., & Matsumoto, Y. (2004). Chinese word segmentation by classification of characters. In Proceedings of 3rd ACL SIGHAN Workshop on Chinese Language Processing. Retrieved September 16, 2004, from http://www1.cs.columbia.edu/~rambow/master-cd-rom/sighan/PDF/13-Ling-2col.pdf

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 329–338). New York: ACM.

Jin, H.L., & Wong, K.F. (2001). A dictionary construction algorithm for information retrieval. ACM Transactions on Asian Language Information Processing, 1(4), 281–296.

Kwok, K.L. (1997). Comparing representations in Chinese information retrieval. In N.J. Belkin, A. Desai Narasimhalu, P. Willett, & W. Hersh (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 34–41). New York: ACM.

Kwok, K.L. (1999). English-Chinese cross-language retrieval based on a translation package. In Proceedings of the Workshop on Machine Translation for Cross Language IR–MT Summit VII (pp. 8–14).

Kwok, K.L. (2000). Exploiting a Chinese–English bilingual wordlist for English-Chinese cross language information retrieval. In Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (pp. 173–179).

Kwok, K.L., & Grunfeld, L. (1997). TREC-5 English and Chinese retrieval experiments using PIRCS. NIST Special Publication 500–238: The Fifth Text REtrieval Conference (pp. 133–142). Gaithersburg; MD: Department of Commerce, National Institute of Standards and Technology.

Lee, K., Oh, J., Huang, J., Kim, J., & Choi, K. (2001). TREC-9 experiments at KAIST: QA, CLIR and batch filtering. NIST Special Publication 500–249. The Ninth Text REtrieval Conference (pp. 303–316). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Li, W.J., Wong K.F., & Yuan, C. (2001). Toward automatic Chinese temporal information extraction. Journal of the American Society for Information Science and Technology, 52, 748–762.

Liddy, E.D. (1998). Enhanced text retrieval using natural language processing. ASIS Bulletin, 24(4).

Melamed, I.D. (1998). Empirical methods for MT lexicon development. In D. Farwell, L. Gerber, & E. Hovy (Eds.), Machine translation and the information soup, The Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA'98 (pp. 18–30). London: Springer-Verlag.

Miller, G. (1990). WordNet: An on-line lexical database [Special issue]. International Journal of Lexicography, 2(4).

Nie, J.Y., Gao, J.F., Zhang, J., & Zhou, M. (2000). On the use of words and n-grams for Chinese information retrieval. Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages (pp. 141–148).

Palmer, D.D. (1997). A trainable rule-based algorithm for word segmentation. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97).

Peterson, E. (1998). A Chinese named entity extraction system. Retrieved September 16, 2004, from http://epsilon3.georgetown.edu/~petersee/chinesene.html

Ruiz, M.E., Rowe, S., Forrester, M., & Sheridan, P. (2001). CINDOR TREC-9 English-Chinese evaluation. NIST Special Publication 500–249: The Ninth Text REtrieval Conference (pp. 379–388). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Salton, G., & McGill, M.J. (1983). The SMART and SIRE experimental retrieval systems (pp. 118–155). New York: McGraw-Hill.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In H.P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 21–29). New York: ACR.

Sproat, R.W., Shih, C.L., Gale, W., & Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. Computational Linguistics, 22, 377–404.

Sun, J., Zhou, M., & Gao. J. (2003). A class-based language model approach to Chinese named entity identification. The International Journal of Computational Linguistics and Chinese Language Processing, 8(2), 1–28

Voorhees, E.M., & Harman, D. (Eds.). (2001). TREC-9 results: Common evaluation measures. NIST Special Publication 500–249: The Ninth Text REtrieval Conference (pp. A-15–A-19). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Wu, L., Huang, X., Guo, Y., Liu, B., & Zhang, Y. (2001). FDU at TREC-9: CLIR, Filtering, and Q&A tasks. NIST Special Publication 500–249: The Ninth Text REtrieval Conference (pp. 189–202). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Xu, J., & Weischedel, R. (2001). TREC-9 cross-lingual retrieval at BBN. NIST Special Publication 500–249. The Ninth Text REtrieval Conference (pp. 106–115). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Xue, N., & Shen, L. (2003). Chinese word segmentation as LMR tagging. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. Retrieved September 9, 2004, from http://www.cis.upenn.edu/~libin/paper/sighan03.pdf

Yu, S.H., Bai, S.H., & Wu, P. (1998). Description of the Kent Ridge digital labs system used for MUC-7. Proceedings of the 7th Message Understanding Conference.

DOI: 10.1002/asi