Experiments on Chinese-English Cross-language Retrieval at NTCIR-4

Yilu Zhou[1], Jialun Qin[1], Michael Chau[2], Hsinchun Chen[1]
[1]*Department of Management Information Systems*
The University of Arizona
Tucson, AZ 85721
yiluz@eller.arizona.edu, qin@u.arizona.edu, hchen@eller.arizona.edu
[2]*School of Business*
The University of Hong Kong
Hong Kong
mchau@business.hku.hk

Abstract

The AI Lab group participated in the cross-language retrieval task at NTCIR-4. Aiming at a practical retrieval system, our applied a dictionary-based approach incorporated with phrasal translation, co-occurrence disambiguation and query expansion techniques. Although experimental results were not as good as we expected, our study demonstrated the feasibility of applying CLIR techniques in real-world applications.

1. Introduction

Cross-language information retrieval (CLIR) involves finding documents in languages other than the query language. Many techniques have been proposed to improve CLIR retrieval performance. The NTCIR workshop, which was begun in 1998, studies CLIR among Asian language covering Chinese, Japanese and Korean. At the NTCIR-4 workshop, the AI Lab group participated in the Cross-language Retrieval Task. We worked on Chinese-English BLIR task and focused on effective and efficient means for CLIR that could be adopted in real-world, interactive Web retrieval applications.

In the remainder of this paper, we discuss related work in section 2. Section 3 presents our approaches and section 4 discusses our experimental results in NTCIT-4. The results include official runs we submitted and the additional runs after submission. Finally, in section 5 we conclude our work and suggest future directions.

2. Related Work

Most research approaches in CLIR translate queries into the document language, and then perform monolingual retrieval [9]. There are three major query translation approaches: using machine translation, a parallel corpus, or a bilingual dictionary. Machine translation-based (MT-based) approach uses existing machine translation techniques to provide automatic translation of queries. The MT-based approach is simple to apply, but the output quality of MT is not always satisfying and MT systems are only available for certain language pairs. A corpus-based approach analyzes large document collections (parallel or comparable corpus) to construct a statistical translation model. Although the approach is promising, the performance relied largely on the availability of the corpus. In a dictionary-based approach, queries are translated by looking up terms in a bilingual dictionary and using some or all of the translated terms. This is the most popular approach because of its simplicity and the wide availability of machine-readable dictionaries.

By using simple dictionary translations without addressing the problem of translation ambiguity, the effectiveness of CLIR can be 60% lower than that of monolingual retrieval [1]. Various techniques have been proposed to reduce the ambiguity and errors introduced during query translation. Among these techniques, phrasal translation, co-occurrence analysis, and query expansion are the most popular. Phrasal translation techniques are often used to identify multi-word concepts in the query and translate them as phrases [2]. Co-occurrence statistics help select the best translation(s) among all translation candidates by assuming that the correct translations of query terms

tend to co-occur more frequently than the incorrect translations do in documents written in the target language [2, 3, 4, 8]. Query expansion assumes that additional terms that are related to the primary concepts in the query are likely to be relevant, and by adding these terms to the query, the impact of incorrect terms generated during the translation can be reduced [1]. Most research has focused on the study of technologies that improve retrieval precision on large-scale evaluation collections. There is a need to explore a set of techniques to be integrated into real-world, interactive Web retrieval applications [5, 12].

3. Proposed Approach in Chinese-English Cross-language Retrieval

Chinese-English retrieval task is to search Chinese topics against the English document collection. Aiming to apply an integrated set of CLIR techniques in a practical system, we propose architecture for CLIR system which consists of four major components: (1) document and query indexing (2) term translation (3) post-translation query expansion (4) document retrieval. These four components were integrated as a one-stop retrieval in our system for CLIR.

3.1. Document and Query Indexing

Both Chinese queries and English documents need to be indexed in Chinese-English retrieval.

Indexing techniques for Chinese language has been studied in much research. Overlapping character n-grams, multi-word phrases and simple words are often used. Our system used phrase-based indexing for Chinese topics and descriptions. The Chinese phrase lexicon was a combination of two: Chinese phrases in LDC bilingual lexicon and Chinese phrases extracted by Mutual Information program. LDC lexicon is a bilingual English-Chinese lexicon available through the Linguistic Data Consortium (LDC). It includes two specific lists: the English-to-Chinese wordlist ("ldc2ec") and the Chinese-to-English wordlist ("ldc2ce"), each contains around 120,000 entries.

The mutual information approach is a statistical method that identifies as meaningful phrases significant patterns from a large amount of text in any language [10]. The approach is an iterative process of identifying significant lexical patterns by examining the frequencies of word co-occurrences in a large amount of text. Three steps are involved: tokenization, filtering and phrase extraction. First, in the tokenization step, each word (or token) in the text is identified by recognizing the delimiter separating it from another word. In Chinese (or many other oriental languages), in which the smallest meaning-bearing unit is a character, the delimiter is identified as the boundary of each word (or character). Second, in the filtering step, a list of stop words is used to remove non-semantic-bearing expressions and a list of included words is used to retain good expressions (words or phrases). Regular expressions can be used in the two lists to specify patterns of words. Third, in the phrase extraction step, statistics of patterns of the words extracted from the above steps are computed and compared against thresholds to decide whether certain patterns are extracted as meaningful phrases. The mutual information (MI) algorithm is used to compute how frequently a pattern appears in the corpus, relative to its sub-patterns. Based on the algorithm, the MI of a pattern c (MIc) can be found by

$$MI_c = \frac{f_c}{f_{left} + f_{right} - f_c}$$

where *f* stands for the frequency of a set of words. Intuitively, MIc represents the probability of co-occurrence of pattern c, relative to its left sub-pattern and right sub-pattern. Phrases with high MI are likely to be extracted and used in automatic indexing. Chinese document collection in NTCIR was sent to MI to build the Chinese lexicon and around 97,000 phrases were extracted. While indexing Chinese queries, functional phrases were removed from description.

English documents were indexed using a combined word-based and phrase-based approach. To support document retrieval, English documents were indexed using a word-based indexing approach. The positional information on the words or characters within a document was captured and stored such that when the query was a phrase, documents containing the exact phrase could be retrieved and given higher ranking than documents with separated words. The English words were stemmed using Porter stemmer [11] and stopwords were removed. Because word-based indexing did not capture phrases during our general indexing process for English documents, Arizona Noun Phraser (AZNP), developed by our research group, was used to extract phrases from the English collection [14]. AZNP has three components: a word tokenizer, a part-of-speech tagger, and a phrase generation module. Its purpose is to extract all noun phrases from each document based on linguistic rules. The indexed terms are potential translations from bilingual dictionaries, and would be used in co-occurrence calculation for translation disambiguation purposes and post-translation query expansion.

3.2. Term Translation

The Translation component is the core of the system. It is responsible for translating search queries in the source language into the target language. Among the three translation approaches, the dictionary-based approach seems to be most promising for practical systems for two reasons. First, compared with the parallel corpora required by the corpus-based approach, MRDs used in dictionary-based CLIR are much more widely available and easier to use. The limited availability of existing parallel corpora cannot meet the requirements of practical retrieval systems in today's diverse and fast-growing information environment. Second, compared with MT-based CLIR, the dictionary-based CLIR approach is more flexible, easier to develop, and easier to control. Therefore, we used a dictionary-based approach combined with phrasal translation and co-occurrence analysis for translation disambiguation.

Query term translations were performed using the LDC (Linguistic Data Consortium) English-Chinese bilingual lexicon as dictionaries. LDC Chinese-to-English wordlist could be used as a comprehensive word dictionary as well as a phrase dictionary. Taking advantage of the phrasal translations, Kwok [7] reported that using the Chinese-to-English wordlist alone improved the effectiveness of CLIR by more than 70%. LDC bilingual lexicon was encoded in GB code that is used in mainland China, while the document collection was encoded in Big5 that is used in Hong Kong and Taiwan. Encoding conversion was performed on LDC lexicon to match the encoding with document collection.

In the dictionary lookup process, the entry with the smallest number of translations will be preferred over other candidates. In addition, we also conducted maximum phrase matching. Translations containing more continuous key words will be ranked higher than those containing discontinuous key words.

Co-occurrence analysis also was used to help choose the best translation among candidates. All possible definition pairs {D1, D2} in the dictionary were extracted such that D1 is a definition of a term 1 in the source language and D2 is a definition of a term 2 in the target language. Each pair was used as a query to retrieve documents in the indexed collections. The co-occurrence score between two definitions D1 and D2 then could be calculated as follows:

$$Co-occur(D_1, D_2) = \frac{N_{12}}{N_1 + N_2}$$

where N_{12} is the number of Web pages returned where performing an "AND" search using both D_1 and D_2 in the query and N_1, N_2 are the numbers of documents returned respectively when using only D_1 or D_2 in the query. Our method is similar to that of [8] in which they sent definition pairs to other search engines and used the number of returned documents to calculate the co-occurrence scores. We calculated co-occurrence scores in advance to avoid affecting run time efficiency.

3.3. Post-translation Query Expansion

The Post-translation Query Expansion component is responsible for expanding the query in the target language (English). The local feedback method was implemented for post-translation query expansion in our system. Our approach followed the method reported by Ballesteros and Croft [2]. The translated query was sent to the document collection in the target language to retrieve the relevant documents. All terms from the top 20 documents were extracted and ranked by tf*idf scores. The top 5 ranked terms were then combined with the translated query and reweighed to build the final query.

3.2. Document Retrieval

The Document Retrieval component is responsible for taking the query in the target language and retrieving the relevant documents from the text collection. After a target query had been built, it was passed to the search module of the system. The search module searched the document indexes and looked up the documents that were most relevant to the search query. The retrieved documents then were ranked by their tf*idf scores and returned to the user through the interface.

4. Evaluation Results

CLIR evaluation in NTCIR aims at testing the effectiveness, measured by precision and recall, of retrieval systems. In this section, we present both our official Chinese-English BLIR results and some post hoc experiments.

The English document collection provided by NTCIR contains 347,549 new articles in China, Taiwan, Hong Kong, Japan and Korea. Evaluation was based on 50 topic descriptions, and relevance judgments were developed using a pooled assessment methodology. NTCIR used four ranks of relevance, highly relevant (S), relevant (A), partially relevant (B) and irrelevant (C) [6]. In the case of "Rigid" documents judged S and A were regarded as correct

answers, while in the case of "Relax" documents judged B were also regarded as correct answers. For each topic, a ranked list of documents were produced and retrieval effectiveness were computed using NTCIR-4 released relevance judgments. We used the Chinese document collection of 381,375 news articles for Mutual Information training process.

For evaluation we submitted Bilingual Chinese-English runs and monolingual English runs. For BLIR, we submitted one result using title queries, AILab-C-E-T-01, and one result using description queries, AILab-C-E-D-01. Narrative part of the topics were not used in our runs. We did not apply query expansion techniques in our official runs. We submitted two official monolingual runs called AILab-E-E-T-01 (title only) and AILab-E-E-D-01 (description only). Table 1 shows non-interpolated average precision values for official runs averaged over all the test queries.

Table 1: Average Precision for Official Runs

	Assessment	Avg. Precision	% of Mono. IR
AILab-E-E-T-01	**Rigid**	0.0802	
	Relax	0.1032	
AILab-E-E-D-01	**Rigid**	0.0342	
	Relax	0.0483	
AILab-C-E-T-01	**Rigid**	0.0587	73%
	Relax	0.0729	70%
AILab-C-E-D-01	**Rigid**	0.0412	39%
	Relax	0.0520	50%

Our official runs did not achieve a high performance, which could be resulted from several factors. First, topics in NTCIR contains a lot of proper nouns that were not covered by LDC bilingual lexicon. Failure in translation there proper nouns dramatically affected the performance of bilingual retrieval. These proper names were mostly people's name, medicine names, organization names and etc. Second, some phrases were mistranslated. Special event titles and special names that contain general meaning nouns often resulted in an incorrect translation. This was often due to the wrong segmentation of Chinese phrases. We believe word-based indexing for Chinese queries brought an information loss because some meaningful phrases, especially new terminologies were not included in our phrase lexicon. We used Mutual Information approach to extract Chinese phrases from NTCIR official Chinese document collection as an addition to existing phrase lexicon. However, the training corpus for MI was not highly comparable to the English document collection for retrieval. Therefore phrases that did not appear often in the training corpus were missed. Third, there was an error in our document retrieval component which affected the performance of both monolingual and bilingual retrieval.

In our post hoc experiment, we corrected the error in English document retrieval process and topic title was used as query terms. In bilingual post hoc experiment, we used local feedback as our post-translation query expansion. The performance improved significantly after the error correction. Table 2 shows non-interpolated average precision values for post hoc runs averaged over all the test queries.

Table 2: Average Precision for Post-hoc Runs

	Assessment	Avg. Precision	% of Mono. IR
AILab-E-E-T-Post hoc	**Rigid**	0.2155	
	Relax	0.2664	
AILab-C-E-T-Post hoc	**Rigid**	0.1023	47%
	Relax	0.1345	50%
AILab-C-E-D-Post hoc	**Rigid**	0.0928	43%
	Relax	0.1120	42%

We observed that using description field yielded lower precision than using title field. We believe that because we used simple tf*idf in document ranking and treated all the query words/phrases equally, same weight was given to unimportant phrases as well as important phrases in description field. A balanced query formulation could improve the performance of document retrieval.

5. Conclusions and Future Directions

NTCIR-4 provided large-scale test collections for CLIR experiments. In this paper, we presented our experience in an Chinese-English retrieval system in NTCIR-4. Aiming at a practical retrieval system, our applied a dictionary-based approach incorporated with phrasal translation, co-occurrence disambiguation and query expansion techniques. Our approach was relatively simple and all the components were integrated as a one-stop searching. However retrieval performance was not as good as we expected. Using description fields yields lower precision than using title fields. This reflected the impact of query length in our retrieval model. NTCIR-4 task is different from our previous experience in Web retrieval where short queries are involved. Overall, our study demonstrated the feasibility of applying CLIR techniques in real-world applications and the experimental results are encouraging.

We plan to expand our research in several directions. First, we plan to integrate more CLIR

techniques into our system to make it more robust. We are also investigating how the speed of the system can be improved to achieve faster response time, which is necessary for an interactive system.

6. Acknowledgement

This project was supported in part by an NSF Digital Library Initiative-2 grant, PI: H. Chen, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999-March 2002. We would also like to thank the AI Lab team members who developed the AI Lab SpidersRUs toolkit, the Mutual Information software and the AZ Noun Phraser.

7. References

[1] L. Ballesteros & B. Croft. Dictionary Methods for Cross-Lingual Information Retrieval. In *Proc. of the 7th DEXA Conference on Database and Expert Systems Applications*, Zurich, Switzerland, September 1996, pp. 791-801,1996.

[2] L. Ballesteros & B. Croft. Resolving Ambiguity for Cross-language Retrieval. *SIGIR'98*, Melbourne, Australia, August 1998, pp. 64-71, 1998.

[3] J. Gao, J. Y. Nie et al. Improving Query Translation for Cross-language Information Retrieval Using Statistical Models. *SIGIR'01*, New Orleans, Louisiana, 2001, pp. 96-104.

[4] D. A. Hull & G. Grefenstette. Querying across Languages: a Dictionary-based Approach to Multilingual Information Retrieval. *SIGIR'96*, Zurich, Switzerland, 1996.

[5] N. Kando. Evaluation - the Way Ahead: A Case of the NTCIR. In *Proceedings of the ACM SIGIR Workshop on Cross-Language Information Retrieval: A Research Roadmap*, Tampere, Finland, August 2002.

[6] K. Kishida, K. Chen et al. Overview of CLIR Task at the Forth NTCIR Workshop. In *Proc. of the 4th NTCIR workshop*, forthcoming.

[7] K. L. Kwok. Exploiting a Chinese-English Bilingual Wordlist for English-Chinese Cross Language Information Retrieval. In *Proc. of the Fifth Int'l Workshop on Information Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 173-179.

[8] A. Maeda, F. Sadat et al. Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine. In *Proc. of the Fifth Int'l Workshop on Info. Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 173-179.

[9] D. Oard. Cross-language Text Retrieval Research in the USA. In Proceedings of the 3rd ERCIM DELOS Workshop, Zurich, Switzerland, March 1997.

[10] T. H. Ong & H. Chen. Updateable PAT-Tree Approach to Chinese Key Phrase Extraction Using Mutual Information: a Linguistic Foundation for Knowledge Management. In *Proc. of the 2nd Asian Digital Library Conference*, Taipei, Taiwan, 1999.

[11] M. F. Porter. An Algorithm for Suffix Stripping *Program*, 14, 130-137, 1980.

[12] J. Qin, Y. Zhou, M. Chau and H. Chen. Supporting Multilingual Information Retrieval in Web Applications: An English-Chinese Web Portal Experiment. In *Proceedings of the International Conference on Asian Digital Libraries (ICADL 2003)*, Kuala Lumpur, Malaysia, December 8-11, 2003.

[13] F. Sadat, A. Maeda, et al. A Combined Statistical Query Term Disambiguation in Cross-language Information Retrieval," in *Proc. of the 13th Int'l Workshop on Database and Expert Systems Applications*, Aix-en-Provence, France, September 2002, pp. 251-255.

[14] K. Tolle & H. Chen. Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. *Journal of the American Society for Information Science*, 51(4), 352-370, 2000.