Supporting Multilingual Information Retrieval in Web Applications: An English-Chinese Web Portal Experiment*

Jialun Qin[1], Yilu Zhou[1], Michael Chau[2], Hsinchun Chen[1]

[1] Department of Management Information Systems, The University of Arizona, Tucson, Arizona 85721, USA
qin@u.arizona.edu, yilu@u.arizona.edu, hchen@eller.arizona.edu
[2] School of Business, The University of Hong Kong, Pokfulam, Hong Kong
mchau@business.hku.hk

Abstract. Cross-language information retrieval (CLIR) and multilingual information retrieval (MLIR) techniques have been widely studied, but they are not often applied to and evaluated for Web applications. In this paper, we present our research in developing and evaluating a multilingual English-Chinese Web portal in the business domain. A dictionary-based approach has been adopted that combines phrasal translation, co-occurrence analysis, and pre- and post-translation query expansion. The approach was evaluated by domain experts and the results showed that co-occurrence-based phrasal translation achieved a 74.6% improvement in precision when compared with simple word-by-word translation.

1 Introduction

The increasing diversity of the Internet has created a tremendous number of multilingual resources on the Web. Web pages have been written in almost every popular language in the world, including various Asian, European, and Middle East languages. It is difficult for a user to retrieve documents written in a language that is not spoken by him/her. Supporting cross-language information retrieval (CLIR) and multilingual information retrieval (MLIR) for Web applications has become a pressing issue. In this paper, we report our experience in designing and evaluating a Web portal called ECBizPort that supports cross-language retrieval for Web pages in different languages.

2 Literature Review

Most reported approaches translate queries into the document language, and then perform monolingual retrieval. There are three main approaches in CLIR and MLIR: using machine translation, a parallel corpus, or a bilingual dictionary. Machine translation-based (MT-based) approach uses existing machine translation techniques to

* This project was supported in part by an NSF DLI-2 grant, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, 1999/4–2002/3.

provide automatic translation of queries. The MT-based approach is simple to apply, but the output quality of MT is still not very satisfying, especially for western and oriental 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 quality of the corpus. Also, parallel corpus is very difficult to obtain, especially for western and oriental language pairs. 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 [3]. 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 [3, 6]. 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 [3, 5, 8, 11]. 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 discussed has focused on the study of technologies that improve retrieval precision on standard TREC collections rather than real-world, interactive Web retrieval applications. Only a few Web-based CLIR and MLIR systems exist, such as Keizai [9], Arabvista, and MULINEX [4]. However, no evaluation data was available for most of these systems, leaving their effectiveness uncertain.

3 Proposed Approach and System Prototype

To support multilingual retrieval on the Web, we apply an integrated set of CLIR and MLIR techniques in the Web environment and propose an architectural design for multilingual information retrieval on the Web. Our architecture consists of five major components: Web Spider, Pre-translation Query Expansion, Query Translation, Post-translation Query Expansion, and Document Retrieval. The Web Spider component is responsible for building our document collections for the Web portal. These documents are not only the information resources provided to users, but also a comparable corpus that can be used for translation disambiguation and query expansion. The Pre-translation Query Expansion component takes a search query and sends it to the local document collection to perform a search. To achieve higher efficiency, our approach followed the local feedback method [2]. The Query Translation component is responsible for translating search queries in the source language into the target language. The Post-translation Query Expansion component expands the query in the target language in a way

similar to Pre-translation expansion. Finally, the Document Retrieval component takes the query in the target language and retrieving the relevant documents from the text collection.

Based on the architecture, we implemented ECBizPort, an English-Chinese Web portal for business intelligence in the information technology (IT) domain. The AI Lab SpidersRUs toolkit (http://ai.bpa.arizona.edu/spidersrus/) is used to build the English and Chinese collections for the Web portal. Each collection consists of about 100,000 IT business Web pages. Query term translations are performed using the LDC (Linguistic Data Consortium) English-Chinese bilingual wordlists as our dictionaries. Phrasal translations [7] and co-occurrence analysis were also implemented.

We built our own lexicon for the IT business domain in both English and Chinese in order to support query expansion using the Arizona Noun Phraser [12] and the Mutual Information technique [10]. The extracted terms were used as the lexicons for both pre- and post-translation query expansion. The last component, Document Retrieval, was supported by the AI Lab SpidersRUs toolkit.

4 System Evaluation

An experiment was designed and conducted to evaluate the performance of our system. Basically, we followed the standard TREC evaluation process in our experiment design. However, because there was no established relevance judgment available for precision and recall calculation, we recruited human judges to determine the relevance of each document. Three business school graduate students, fluent in both English and Chinese, served as the experts. They identified seven Chinese queries of interest in the business IT domain and translated these queries into English as the base queries. The original Chinese queries were used to get monolingual retrieval results. Such monolingual retrieval represents the performance of traditional information retrieval. The English base queries were used to get cross-lingual runs based on five settings: word-by-word translation, phrasal translation with co-occurrence analysis, phrasal translation with co-occurrence analysis and pre-translation expansion, phrasal translation with co-occurrence analysis and post-translation expansion, and phrasal translation with co-occurrence analysis and both pre- and post-translation expansion.

The results were compared with two standard benchmark settings: (1) monolingual retrieval (the best-case scenario), and (2) word-by-word translation (the worst-case scenario). In general, the results are encouraging. All methods except word-by-word translation achieved over 70% of the performance of the monolingual system.

We also observed that all four methods based on phrasal translation and/or query expansions perform significantly better than word-by-word translation, the baseline translation method. Phrasal and co-occurrence disambiguation performed much better than word-by-word translation, achieving a 74.6% improvement. However, pre-translation and post-translation expansion did not greatly improve the performance.

5 Future Directions

This study demonstrated the feasibility of applying CLIR and MLIR techniques in Web applications and the experimental results are encouraging. The proposed approach can be applied in other Web-based applications or digital libraries. In the future, we plan to integrate other CLIR and MLIR techniques into the Web portal to make it more robust. In addition, we plan to expand the Web portal to more languages, such as Spanish and Arabic. Such expansion will allow us to study whether the reported techniques will perform differently for a multilingual Web portal with more than two languages.

References

1. Ballesteros, L. & Croft, B. (1996). "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.
2. Ballesteros, L. & Croft, B. (1997). "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval," *SIGIR'97*, Philadelphia, PA, July 1997, pp. 84-91.
3. Ballesteros, L. & Croft, B. (1998). "Resolving Ambiguity for Cross-language Retrieval," *SIGIR'98*, Melbourne, Australia, August 1998, pp. 64-71.
4. Capstick, J., Diagne, A. K., et al. (1998). "MULINEX: Multilingual Web Search and Navigation," in *Proc. of Natural Language Processing and Industrial Applications*, Moncton, Canada, 1998.
5. Gao, J., Nie, J.-Y., et al. (2001). "Improving Query Translation for Cross-language Information Retrieval Using Statistical Models," *SIGIR'01*, New Orleans, Louisiana, 2001, pp. 96-104.
6. Hull, D. A. & Grefenstette, G. (1996). "Querying across Languages: a Dictionary-based Approach to Multilingual Information Retrieval," *SIGIR'96*, Zurich, Switzerland, 1996.
7. Kwok, K. L. (2000). "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. Maeda, A., Sadat, F., et al. (2000). "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. Ogden, W. C., Cowie, J., et al. (1999): "Keizai: An Interactive Cross-Language Text Retrieval System," in *Prco. of Workshop on Machine Translation for Cross Language Info. Retrieval*.
10. Ong, T.-H. & Chen, H. (1999). "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. Sadat, F., Maeda, A., et al. (2002). "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.
12. Tolle, K. M. & Chen, H. (2000). "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools," *Journal of the American Society for Information Science*, 51(4), 352-370.