

# MULTILINGUAL INFORMATION ACCESS IN DIGITAL LIBRARY

**Chen, Hsin-Hsi**

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, Taiwan

E-mail: hh\_chen@csie.ntu.edu.tw

## **Abstract**

The trend toward information globalization has brought new challenges for information management. On the one hand, it is often necessary for a digital library to share its valuable resources with users of different languages. On the other hand, it is also necessary for a digital library user to utilize knowledge presented in a foreign language. This paper addresses several important issues which should be tackled by the global information village. Some important technologies including query translation and/or transliteration, named entity extraction, translingual transmedia information retrieval, and information fusion are listed. Evaluation of a multilingual information access system is also discussed.

## **1. Introduction**

The major characteristic of information dissemination at the new information era is that Internet breaks the distance of regions and sets up a global information village without border. The information distributed in any places is extremely easy to obtain, not only rich but also real time. Besides large scale, the languages used are many. Thus, how to share the valuable resources with users of different languages, and how to utilize knowledge presented in a foreign language are indispensable.

Digital library, which owns large scale digitalized resources, plays important roles in media-rich life. Multi-media, multi-linguality and multi-culture are the three major characteristics (Borgman, 1997). Digital library is an integration of content and technology. This paper will focus on the issue of multi-linguality in the technology part.

Several factors are important to design a multilingual information access system (Bian and Chen, 2000), including information input, representation and transmission, manipulation, and visualization. Manipulation, which concerns classification, retrieval, filtering, extraction, and summarization of multilingual data, is the major focus of this paper, retrieval in particular.

## **2. Research Issues**

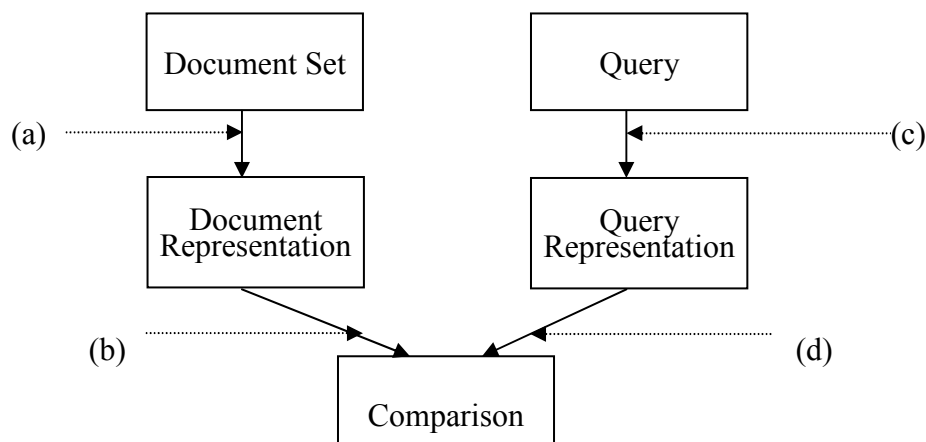
The research issues behind a multilingual information access system are many.

Some related to multilingual information retrieval are shown as follows.

- (1) Query and document belong to different languages, so that translation is required.
- (2) Query terms are ambiguous, so that translation ambiguity and target polysemy should be faced.
- (3) Query is usually short, so that to capture user's information need is important.
- (4) The boundary of basic tokens in some languages like Chinese is not clear, so that segmentation is necessary.
- (5) The rich content is in different languages and media, so that translingual transmedia is challenging.
- (6) The rich content is disseminated over different places, so that information fusion is needed.

### 3. Theories and Technologies

Figure 1 shows the possible enhancement of basic information retrieval model to deal with multilinguality. Four alternatives including (a) document translation, (b) document vector translation, (c) query translation, and (d) query vector translation are introduced (Chen, 2002).



**Figure 1.** Enhancing Basic Information Retrieval Model

Query translation is wide adopted because of its simplicity and practicability. Three approaches, i.e., dictionary-based, corpus-based, and integration-based, have been proposed before. These approaches employ different resources, e.g., dictionary and/or corpus, to select suitable translation terms. Thus how to set up bilingual resources (semi)-automatically is important. Chen, Lin and Lin (2002) proposed a method to integrate five linguistic resources, including English/Chinese sense-tagged corpora, English/Chinese thesauri, and a bilingual dictionary, and built a Chinese-English WordNet for translingual applications.

Besides the translation ambiguity from source query to target query, target

polysemy in target query also introduces noise in query translation. Two monolingual balanced corpora are employed to learn word co-occurrence for translation ambiguity resolution, and augmented translation restrictions for target polysemy resolution (Chen, Bian and Lin, 1999).

Named entities (MUC, 1998) form fundamental tokens in documents. They are usually the targets that users are interested in. That is, users often issue queries to retrieve those documents with some specific proper names. However, proper names are open sets. For example, new organizations are set up continuously, and old organizations may be renamed or even dismissed. A lexicon cannot capture all the named entities. Chen, Yang and Lin (2003) distinguish which part in a query term should be translated and which part should be transliterated. Two alternatives, grapheme-based and phoneme-based approaches (Chen, Huang, Ding, and Tsai, 1998; Lin and Chen, 2002), are proposed to backward-transliterate named entities.

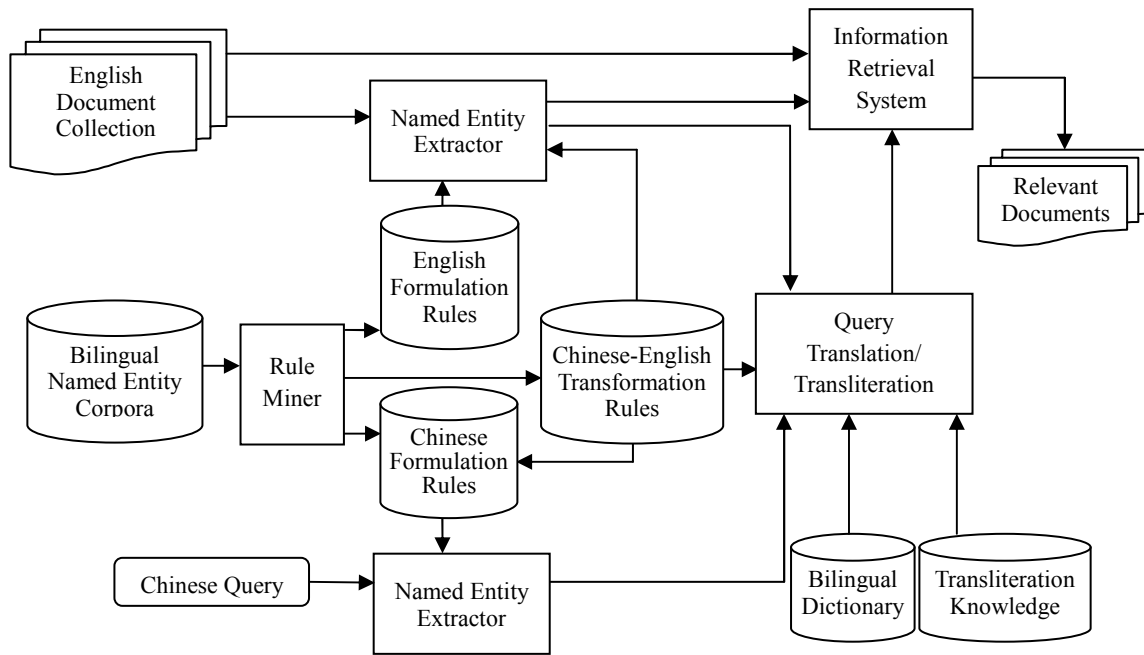
Contents are disseminated from different sources in different languages and media. The typical example is to employ Chinese speech to access an image database with English text captions. Three media, i.e., speech signals, images and text captions, are involved. The query translation mechanism is more complex in translingual transmedia information retrieval. Which are important cues for translation or transliteration, and what their semantic roles are should be dealt with. Lin, Lin and Chen (2004) present a pilot study on this problem.

Figure 2 shows a typical multilingual information retrieval system. The formulation rules and the transformation rules are mined from English-Chinese named entity corpora. From English documents, a set of index terms are extracted. Named entities form part of index terms. When a Chinese query is submitted, named entities are recognized by using Chinese formulation rules. Then query translation and transliteration is performed to transform the Chinese query into an English one. Finally, the result is sent to an English information retrieval system.

Chen (2001) deals with the translingual issue on the design of National Palace Digital Museum, which is a cultural showcase of Taiwan. A cross language information retrieval system is proposed to support English access to Chinese materials. Users can select English input and Chinese output when they are neither familiar with Chinese input, nor lack of Chinese input device, but can read Chinese. Images or videos are transparent to those users that cannot read/write Chinese.

#### **4. Evaluation**

Besides theory and technology, evaluation is an important step in a system development cycle. TREC, CLEF and NTCIR are three famous information retrieval evaluation forums. TREC focus on strategic languages like Arabic, while CLEF and NTCIR touch on European and Asian languages respectively (Chen, 2002; Chen and



**Figure 2.** A Chinese-English Information Retrieval System

Chen, 2001). A test bed for multilingual information retrieval consists of topic descriptions, multilingual document sets, and answer keys. Take NTCIR as an example (Chen and Chen, 2001). It is jointly organized by Japan, Korean and Taiwan researchers (Kishida, *et al.*, 2004). The topic descriptions and the document sets are in four languages, i.e., Chinese, English, Japanese and Korean. In this framework, we can simulate the use of Chinese queries to access documents in Chinese, English, Japanese and Korean. Systems of different query construction strategies, translation/transliteration strategies, and result merging strategies can be evaluated and improved.

## 5. Conclusion

This paper gives an overview of multilingual information access in digital library. Some technologies for query translation/transliteration, named entity extraction, translingual transmedia information retrieval, and information fusion are listed. An application to National Palace Digital Museum is also touched on. Evaluation is important for performance improvement. Three major multilingual information retrieval evaluation forums are discussed.

## References

- Bian, Guo-Wei and Chen, Hsin-Hsi (2000) "Cross Language Information Access to Multilingual Collections on the Internet." *Journal of American Society for Information Science*, **51**(3), 2000, 281-296.

- Borgman, C.L. (1997) "Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries: How Do We Exchange Data in 400 Languages." *D-Lib Magazine*, <http://www.dlib.org/dlib/june97/06borgman.html>.
- Chen, Hsin-Hsi (2001) "Cross-Language Information Retrieval for Digital Museums." *Global Digital Library Development in the New Millennium*, Ching-chih Chen (Editor), Tsinghua University Press, Peijing, China, 33-40.
- Chen, Hsin-Hsi (2002) "Cross-Language Information Retrieval: Theories and Technologies." *Journal of Library and Information Science*, **28**(1), 19-32.
- Chen, Hsin-Hsi, Bian, Guo-Wei and Lin, Wen-Cheng (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval." *Proceedings of 37<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, 215-222.
- Chen, Kuang-Hua and Chen, Hsin-Hsi (2001) "Cross-Language Chinese Text Retrieval in NTCIR Workshop – Towards Cross-Language Multilingual Text Retrieval." *ACM SIGIR Forum*, **35**(2), 12-19.  
<http://www.acm.org/sigir/forum/F2001-TOC.html>
- Chen, Hsin-Hsi, Huang, Sheng-Jie, Ding, Yung-Wei and Tsai, Shih-Chung (1998) "Proper Name Translation in Cross-Language Information Retrieval." *Proceedings of 17<sup>th</sup> International Conference on Computational Linguistics and 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, 232-236.
- Chen, Hsin-Hsi, Lin, Chi-Ching and Lin, Wen-Cheng (2002) "Building a Chinese-English WordNet for Translingual Applications." *ACM Transactions on Asian Language Information Processing*, **1**(2), 103-122.
- Chen, Hsin-Hsi, Yang, Changhua and Lin, Ying (2003) "Learning Formulation and Transformation Rules for Multilingual Named Entities." *Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, 1-8.
- Kishida, Kazuaki, Chen, Kuang-hua, Lee, Sukhoon, Chen, Hsin-Hsi, Kando, Noriko Kuriyama, Kazuko, Myaeng, Sung Hyon and Eguchi, Koji (2004) "Cross-Lingual Information Retrieval (CLIR) Task at the NTCIR Workshop 3." *ACM SIGIR Forum*.
- Lin, Wei-Hao and Chen, Hsin-Hsi (2002) "Backward Machine Transliteration by Learning Phonetic Similarity." *Proceedings of 6<sup>th</sup> Conference on Natural Language Learning*, 139-145.
- Lin, Wen-Cheng, Lin, Ming-Shun and Chen, Hsin-Hsi (2004) "Cross-Language Image Retrieval via Spoken Query." *Proceedings of RIAO 2004: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*.
- MUC (1998) *Proceedings of 7<sup>th</sup> Message Understanding Conference*, [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/index.html](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html).