

# Introduction to the Special Topic Section on Multilingual Information Systems

**Christopher C. Yang and Wai Lam**

*Department of Systems Engineering and Engineering Management, William M. W. Mong Engineering Building, The Chinese University of Hong Kong, Shatin, Hong Kong, People's Republic of China.*

*E-mail: {yang, wlam}@se.cuhk.edu.hk*

The information available in languages other than English on the World Wide Web and global information systems is increasing significantly. According to some recent reports, the growth of non-English speaking Internet users is significantly higher than the growth of English-speaking Internet users. Asia and Europe have become the two most-populated regions of Internet users. However, there are many different languages in the many different countries of Asia and Europe. And there are many countries in the world using more than one language as their official languages. For example, Chinese and English are official languages in Hong Kong SAR; English and French are official languages in Canada. In the global economy, information systems are no longer utilized by users in a single geographical region but all over the world. Information can be generated, stored, processed, and accessed in several different languages. All of this reveals the importance of research in multilingual information systems.

There are several essential components in multilingual information systems as depicted in Figure 1. These components are namely multilingual resources, machine translation, cross-lingual information retrieval, multilingual information extraction and summarization, and user evaluations and studies.

Multilingual resources include corpora, lexicons, and ontology. Parallel and comparable corpora are important for generating a statistical translation model to overcome the limitations of a manually generated dictionary. In addition, annotated corpora and lexicons have been widely used for many natural language processing tasks. Unfortunately, the development of these resources requires much human intervention. Ontology is an inventory of concepts organized in some internal structuring principle, which is important in organizing and managing information.

Machine translation has over 50 years of history. It is defined as an automated process to transform written text from one language to another. One approach is to convert the source text into an abstract semantic representation. This semantic representation is used for producing the translated text in the target language. Another approach is mainly based on a statistical model for word translations and word re-orderings. The model parameters can be learned from a large parallel corpus. Recently, research on translating named entities has become popular because it is useful in different information access applications for which named entities play an important role. Automatic generation of transliteration rules is also actively explored.

*Multilingual information retrieval* is defined as the process that takes queries in any language, searches a collection of objects—including text, images, sound clips—and returns the most relevant objects. It involves several major tasks, namely, query translation, indexing, and retrieval methods. One can employ common information retrieval techniques for conducting indexing and retrieval. Query translation is performed in a separate effort. Another approach is to develop a framework that can deal with all these tasks in a more integrated manner.

*Information extraction and summarization* refer to extracting the portion of text that is most relevant to user's tasks. Information extraction from documents aims at extracting short text fragments with certain semantic information. A common approach to information extraction is to develop a set of extraction rules. In some applications, the rules can be designed manually. Recently, some work on learning extraction knowledge from annotated training examples has been done. After the training phase, the learned knowledge or rules can be used to conduct automatic information extraction. The goal of automated summarization is to generate a summary from the document. Sentence selection using salient features extracted from documents is the most popular approach. Because of grammatical and lexical differences between languages, some of these techniques may not be applied directly.

---

Accepted March 9, 2005

© 2006 Wiley Periodicals, Inc. • Published online 1 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20325

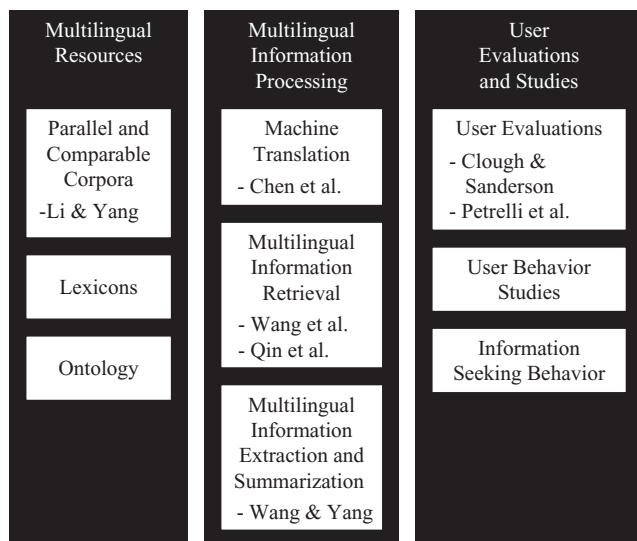


FIG. 1. Framework of multilingual information systems and charting of articles in this special topic section.

Whereas there are many studies related to multilingual information processing, there is limited work that focuses on the evaluation and assessment of multilingual information systems. Typical evaluation methods for monolingual systems may not be applicable to multilingual information systems. User behaviors may also be different when users deal with multilingual information systems.

The theme of this special topic section focuses on multilingual information systems. There are seven articles in this section presenting recent advances in different components of the field. Although the articles cannot cover all aspects of each component, they largely represent the latest research topics and state-of-the-art developments.

### Multilingual Resources

Li and Yang in their article, “Conceptual Analysis of Parallel Corpus Collected From the Web” (2006) present a conceptual analysis of the parallel corpus that they collected from the Web based on their automatic alignment technique using longest common subsequences. Concept equivalence and conceptual alternation are also discussed. In their analysis, translation of concepts are classified into six categories (a) loan word, (b) cultural substitute, (c) synonym (d) antonym, (e) reciprocal, and (f) descriptive phrase. Omission, addition, and redundancy are considered in complex concepts. They found that omission and addition were the major problems in unsuccessful alignments although their alignment technique achieved 100% of precision and 87% of recall.

### Machine Translation

Chen, Lin, Yang, and Lin in “Translating–Transliterating Named Entities for Multilingual Information Access” (2006) have developed a framework for handling multilingual-named

entities. They propose a principled way of dealing with formulation, transformation, translation, and transliteration of multilingual-named entities. They focus on person names, location names, and organization names. Formulation and transformation rules are mined from several aligned corpora. A similarity-based model is proposed for backward transliteration. The similarity measure is mainly conducted at the phoneme level. They also investigate an application of the proposed method for cross-lingual retrieval of a collection of images with captions.

### Multilingual Information Retrieval Techniques

Wang, Teng, Lu, and Chien in their study, “Exploiting the Web as the Multilingual Corpus for Unknown Query Translation” (2006) propose an approach for exploiting the Web as the multilingual corpus source for translating unknown query terms. Many user queries contain terms not found in an ordinary translation dictionary. The authors developed a novel technique to mine bilingual search-result pages obtained from a Web search engine for helping the translation of unknown query terms. This approach can also be used for improving a domain-specific bilingual lexicon. Experimental results demonstrate that the proposed approach is effective in finding translations for many unknown terms including proper nouns, technical terms, and Web query terms. This approach can also benefit multilingual access in a digital library.

Qin, Zhou, Chau, and Chen in their article, “Multilingual Web Retrieval: An Experiment in English–Chinese Business Intelligence” (2006) investigate the feasibility of employing various cross-lingual information retrieval techniques for developing and evaluating a bilingual Web portal. Their system consists of five major components, namely, Web spider and indexer, pretranslation query expansion, query translation, posttranslation query expansion, and document retrieval. The query translation is based on a dictionary-based approach that includes phrasal translation, co-occurrence analysis, and pre- and posttranslation query expansion. They apply the techniques to the development of a multilingual Web portal, ECBizPort, which is an English–Chinese Web portal for business intelligence in the information technology domain.

### Multilingual Information Processing and Applications

Wang and Yang in their work, “The Impact Analysis of Language Differences on an Automatic Multilingual Text Summarization System” (2006) develop a bilingual automatic text summarization system and investigate the impact of language difference on the system. Fractal theory is applied in their summarization technique in which the hierarchical structure of documents are considered. An English–Chinese parallel corpus was adopted in the study. The authors investigate the parallelism of the parallel corpus in terms of number of text blocks at different levels of

document structure, sentence mapping, term frequency, and length of sentences. In conducting a comparison of the English and Chinese summaries, they found that most of the sentences in the English and Chinese summaries were not matched but their precisions were close and their contents were similar. As the compression ratio increased, the percentage of direct match increased but the precision decreased. They also found that the precision of the matched sentences was relatively higher than the unmatched sentences.

### User Evaluation and Studies

Clough and Sanderson in “User Experiments With the Eurovision Cross-Language Image Retrieval System” (2006) present the user evaluation of the text-based cross-language image retrieval system, Eurovision. Eurovision provides Web-access to the St. Andrews University Library image archive and supports cross-language queries using an existing machine-translation system. The authors conducted user evaluations with know-item search and category search. Their key findings were that the overall performance of the cross-language system was relatively close to the monolingual system and the image categories assisted users in cross-language search. Users tended to browse through pages of image results rather than viewing the image captions; concept hierarchy was frequently used and bilingual searching was preferable.

Petrelli, Levin, Beaulieu, and Sanderson in their paper, “Which User Interaction for Cross-Language Information Retrieval? Design Issues and Reflections” (2006) present the user evaluations undertaken during the iterative design of the cross-language retrieval system, Clarity. The user interaction

of Clarity was divided into two phases (a) query translation and user checking, and (b) search. Supervised mode and delegated mode were investigated in the user study. User verification and/or modification in query translation was available in the supervised mode but not in the delegated mode. They found that the supervised mode performed better than the delegated mode but not significantly. However, the delegated mode is more preferable. The studies are important to the final user interface design of Clarity.

### References

- Chen, H.-H., Lin, W.-C., Yang, C., & Lin, W.-H. (2006). Translating-transliterating named entities for multilingual information access. *Journal of the American Society for Information Science and Technology*, 57, 645–659.
- Clough, P. & Sanderson, M. (2006). User experiments with the Eurovision cross-language image retrieval system. *Journal of the American Society for Information Science and Technology*, 57, 697–708.
- Li, K.W., & Yang, C.C. (2006). Conceptual analysis of parallel corpus collected from the Web. *Journal of the American Society for Information Science and Technology*, 57, 632–644.
- Petrelli, D., Levin, S., Beaulieu, M., & Sanderson, M. (2006). Which user interaction for cross-language information retrieval? Design issues and reflections. *Journal of the American Society for Information Science and Technology*, 57, 709–722.
- Qin, J., Zhou, Y., Chau, M., & Chen, H. (2006). Multilingual Web retrieval: An experiment in English–Chinese business intelligence. *Journal of the American Society for Information Science and Technology*, 57, 671–683.
- Wang, F.L., & Yang, C.C. (2006). The impact analysis of language differences on an automatic multilingual text summarization system. *Journal of the American Society for Information Science and Technology*, 57, 684–696.
- Wang, J.-H., Teng, J.-W., Lu, W.-H., & Chien, L.-F. (2006). Exploiting the Web as the multilingual corpus for unknown query translation. *Journal of the American Society for Information Science and Technology*, 57, 660–670.