Constructing Bilingual Resources for Digital Libraries

Hae-Chang Rim
Natural Language Processing Lab
Department of Computer Science, Korea University
1, AnamDong 5Ga, SungBuk-Gu
Seoul, Korea
rim@nlp.korea.ac.kr

1. Language Barriers at Digital Library

A digital library is a library on the Internet. Since the library is run on the Internet, its visitors will be from various regions and may use diverse languages. To consider these visitors, the digital library should be multi-lingual and offer information in multi languages. In order to construct a multi-lingual digital library properly, the language barriers should be surmounted. The language gap can be filled with methods such as machine translation and cross language information retrieval. These methods should be developed based on bilingual resources, therefore constructing bilingual resources is the first consideration for surmounting the language barriers.

2. Bilingual Resources for Surmounting Language Barriers

The first step to surmount language barrier is constructing linguistic resources that represent relations between different languages. As the linguistic resources, there is a bilingual dictionary which contains word relations between two languages, a bilingual corpus which shows bilingual word usage and expression, and a bilingual thesaurus which represents concept information between two different languages [1].

Bilingual Dictionary

A bilingual dictionary is the dictionary containing words and their translated words. It is an essential linguistic resource to cross language boundaries by language translation. To translate language, selecting appropriate translated words for every source words is required, and to do this, it is indispensable to know available target words for a source word. This information can be acquired from the bilingual dictionary. The target-source word information is the primary information in language translating process and can improve the system performance.

Bilingual Corpus

While the bilingual dictionary represents translated words for a word of source language, a bilingual corpus represents a translated sentence for a sentence in the source language. Since the bilingual corpus represents corresponding relation between bilingual sentences, bilingual word usage and expression can be acquired from the corpus. In the language translation process, the bilingual dictionary suggests translation word candidates for a source word, and the best target word is selected among the candidates. In selecting the best target word, it is insufficient to use bilingual dictionary only. Therefore, it is necessary to use word usage and expression information from the bilingual corpus. That is, the most appropriate translated word is selected based on lexical co-occurrence and expression information from the source sentence and co-occurrence words and expression information of the sentence in the target language. All of this information can be acquired from the bilingual corpus. The Canadian Hansard Corpus is the most well known bilingual corpus. This corpus is the bilingual corpus consisting of parallel texts in English and Canadian French [2].

Bilingual Thesaurus

A thesaurus is a semantic structure of words, which contains synonyms, hypernyms, and hyponyms of words. A bilingual thesaurus represents semantic structure of bilingual words. Namely, it contains semantic structures of two languages as well as translated words. The bilingual thesaurus is a higher-level linguistic resource than bilingual dictionary in a sense that the former contains semantic structure of words that are not represented by the latter. The bilingual thesaurus can be used for not only word translation as the bilingual dictionary, but also a query expansion in multilingual information retrieval. The query expansion is a method that can offer more information to the users, which can be done by using semantic structures in the bilingual thesaurus.
EuroWordNet is an example of the existing multi-lingual thesaurus. EuroWordNet, which is constructed by many European countries, uses a structure called *interlingual* index to represent the semantic structure of multilingual words [3].

3. Utilizing Bilingual Resources

Machine translation and multi-lingual IR systems are good examples for systems utilizing bilingual language resources. Machine translation, which translates a document to the other language, is the best way to surmount language barriers. To perform machine translation, the word-to-word conversion is required. In order to convert a word properly, first of all, we have to get information about words in another language that can be translated from the word in source language, and then select the most appropriate word among them. The information can be acquired from the bilingual dictionary, and the most appropriate translated word can be selected by using the information of linguistic expression. If the bilingual corpus is available, the example based machine translation method will improve the performance of MT [4].
Multilingual information retrieval is the method that can be used when the query language and the language used in documents to be retrieved are different. Namely, in multilingual information retrieval, the query is written in user-friendly language, but the retrieved documents are expressed in many languages. Therefore, the multilingual information retrieval is an appropriate model when the users of IR systems cannot express their opinion in another languages, but can read and understand them. Translating a query or documents to be retrieved into another language is required for multilingual IR. In this translation process, bilingual dictionary is needed, and the bilingual corpus can be used to improve the translation precision in the process. And bilingual thesaurus can be used in query expansion to offer more information to the users [5].

Reference

[1] Eduard Hovy, Nancy Ide, Robert Frederking, Joseph Mariani, Antonio Zampolli, *Multilingual Information Management: Current Levels and Future Abilities*, July, 1998.
[2] http://www.ldc.upenn.edu/
[3] Vossen, Piek. "Introduction to EuroWordNet". *Computers and the Humanities*, Vol 32, Nos. 2-3, pp. 73-89, 1998.
[45] Ralf D. Brown, "Example-Based Machine Translation in the Pangloss System". In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, p. 169-174. Copenhagen, Denmark, August 5-9, 1996.
[5] Julio Gonzalo, Felisa Verdejo, Carol Peters and Nicoletta Calzolari, "Applying EuroWordNet to Cross-Language Text Retrieval". *Computers and the Humanities*, Vol 32, Nos. 2-3, pp. 73-89, 1998.

HAE-CHANG RIM, PROF.
DEPARTMENT OF COMPUTER SCIENCE
KOREA UNIVERSITY
5KA,ANAM-DONG,SUNGBUK-KU
SEOUL136-701, KOREA
82-2-3290-3195(OFFICE),82-2-953-0771(FAX)
rim@nlp.korea.ac.kr

Education

- Ph.D., Computer Science, University of Texas at Austin, (Advisor: Robert F.Simmons) ,1990
- M.S., Computer Science, University of Missouri-Columbia, 1983
- B.A., German Linguistics & Literature, Korea University, 1979

Experience

- Professor, Department of Computer Science, Korea University, 1999-present
- Chairperson, Steering Committee, Special Interest Group of Korean Information Processing, Korea Information Science Society, 1999-present
- Editorial staff, Korea Information Science Society, 1994-present
- Director, Information Processing Lab., Research Institute of Language Information, Korea University, 1993-present
- Director, Natural Language Processing Lab., Korea University, 1991-present

Projects about DL

- Research for developing Korean bibliography information search system
- Research for constructing digital museum information system
- Research on automatic document clustering
- Developing a server construction of internet electronic store for e-commerce
- Constructing Korean thesaurus for multilingual information retrieval

Publications about DL

- Ching Y. Suen, Shunji Mori, **Hae-Chang Rim**,
  "Intriguing Aspects of Oriental Languages", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 12, No. 1, pp. 5-29, 1998
- Jin-Dong Kim, Heui-Seok Lim, **Hae-Chang Rim**,
  "Twoply Hidden Markov Model : A Korean POS tagging Model Based on Morpheme-unit with Word-unit Context",
  Computer Processing of Oriental Languages, Vol. 11, No. 3, pp. 227-290, 1998
- Bo-Hyun Yun, Yong-Jae Kwak, **Hae-Chang Rim**,
  "Resolving Ambiguous Segmentation of Korean Compound Nouns Using Statistics and Rules", *Computational Intelligence*, Vol. 15, No. 2, pp. 101-113, 1999
- Ho-Lee, **Hae-Chang Rim,** Jungyun Seo,
  "Word sense disambiguation using the Classification information Model",
  Computers and Humanities, Vol. 34, No. 1, pp.141-146, 2000
- Bo-Hyun Yun, Yong-Jae Kwak, **Hae-Chang Rim**,
  "Alleviating Syntactic Team Mismatches in Korean Text Retrieval",
  Information Processing and Management, Vol. 35, No. 4, pp. 481-500, 1999