Cross-Language Information Retrieval System for Korean-Chinese-Japanese-English Languages

Yong-Seok Choi, Juho Lee, Jin-Xia Huang, Key-Sun Choi
KORTERM
Department of Computer Science
Korea Advanced Institute of Science and Technology
Taejon 305-701, Korea,
{angelove, mywork, hgh, kschoi}@world.kaist.ac.kr

Introduction

Cross-Language Information Retrieval (CLIR) (Oard and Dorr, 1996; Oard, 1997) refers to the retrieval when the query and the document database are in different languages. A user who uses one language can retrieve the documents in another language. Query translation is the least expensive and more practical approach to CLIR when compared to full document translation.

This CLIR system is the current status of the project Xirch (Hayashi *et al.*, 1999), a cross-language information retrieval System. The proposed CLIR system has two parts. One is *receiver* that gets the query from other servers; the other is *sender*, which sends the query to other servers. Our server uses meta-searcher method and enhanced STARTS (Stanford Agreement for internet ReTrieval and Search) protocol (Gravano, *et al.* 1997).

Figure 1 shows the receiver modules. When the other requesting server sends the English query to our server, the English query is translated into Korean. The translated Korean query is used to retrieve adequate documents from Korean database. The adequate documents are translated into English, if required from the requesting server. The results are returned to the server that sent the English query to our *receiver* server.

Our server sends queries to other servers for retrieving their documents, too. Figure 2 depicts about how to deal with a Korean query that will be sent to English-handling retrieval servers. When a Korean query is accepted, there are two kinds of actions: one is Korean-to-English query translation and the other is to retrieve Korean documents for the given query. The translated English query is sent to other English-handling servers that will send the results. Finally, the user gets combining results that are written in Korean, English, Japanese, and Chinese.

Configuration

The overall structure of the system is as follows. Korean users can access the Korean, Japanese and Chinese database through the intermediate manager. When Korean users make a query in Korean, the intermediate manager translates this into Korean, Japanese, Chinese and transmits each one to Korean, Japanese, Chinese database for retrieval. Then, documents in each language are received as results. These documents are translated into Korean for Korean users.

At present, the translation techniques among these three languages are not satisfactory. So, We use English as a interlingua to translate each languages. For example, when Korean is needed to be translated into Japanese, Korean is

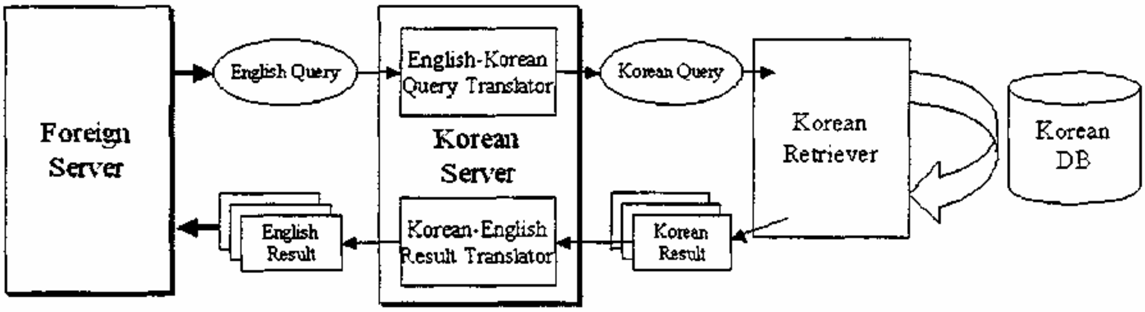

Figure 1. Receiver modules

translated into English at first and then this result is translated into Japanese.

When a Korean query is accepted, there are two kinds of actions: one is Korean-to-English query translation and the other is to retrieve Korean documents for the given query. The translated English query is sent to other English-handling servers that will send the results. Finally, the user gets combining results that are written in Korean, English, Japanese, and Chinese.

When the other requesting server sends the English query to our server, the English query is translated into Korean. The translated Korean query is used to retrieve adequate documents from Korean database. The adequate documents are translated into English, if required from the requesting server. The results are returned to the server that sent the English query to our receiver server.

In this way the CLIR system provides useful information to users integrating many techniques.

Acknowledgements

This work was supported by KOSEF through the "Multilingual Information Retrieval" project at the AITrc and was supported by Ministry of Information and Communication through the " Cross-language Information Retrieval System for Korean-Chinese-Japanese-English Languages" project. And Many fundamental researches was supported by the R&D fund of Ministry of Science and Technology under the project "On Development of Deep-Level Processing and Quality Management Technology for Very Large Korean Information Base", a project of plan STEP2000.

This system is the result of the efforts of all of our project members, Sang-Heon Lee of Orom info., Yang Cho of LnC Corp., Zong-Cheol Zhoo of K4M and we are deeply grateful to them.

References

Gravano, Luis, Chen-Chuan K. Chang, Hector Garcia-Molina, Andreas Paepcke (1997) *"STARTS: Stanford Proposal for Internet Meta-Searching"*, Sigmod'97, http://www-db.stanford.edu/~gravano/starts_home.html.

Hayashi, Y. and H. Kuraishi (1999) *"Effectiveness of Automatically Extracted Translation Correspondences in Cross-Language Text Retrieval"*, Workshop on Cross-lingual Information Retrieval in MT Summit, Singapore.

Oard, D.W. (1997) *"Alternative Approaches for Cross-Language Text Retrieval."* Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval, 131-139.

Oard, D.W. and Dorr, B.J. (1996) *A Survey of Multilingual Text Retrieval.* Technical report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

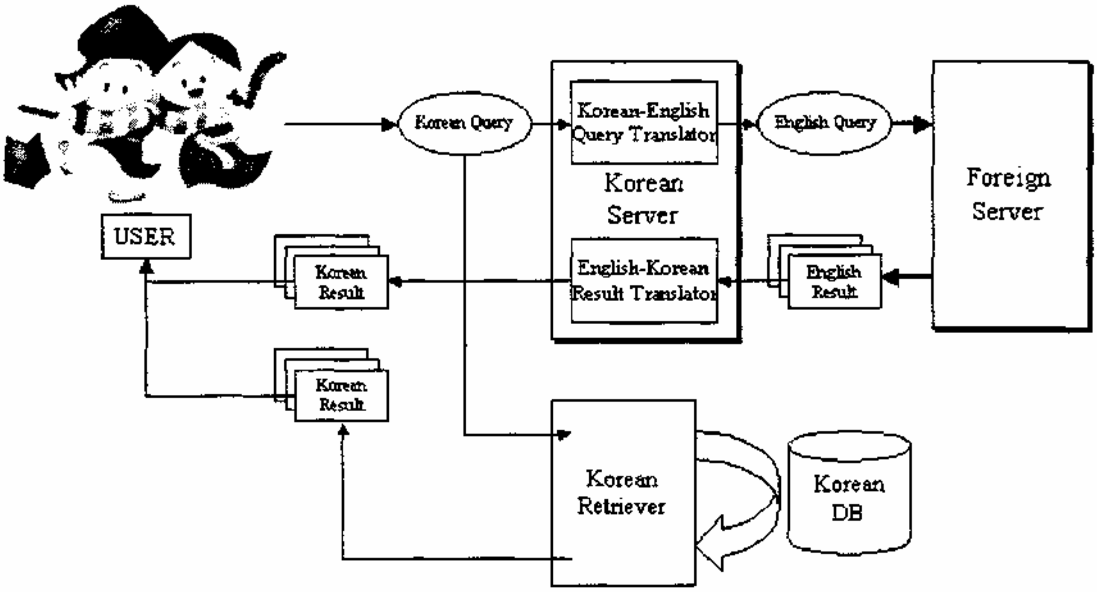

Figure 2. Sender modules