IASL Korean-Chinese CLIR System

Query Translation CLIR System based on Bilingual Dictionary and Co-occurrence Method

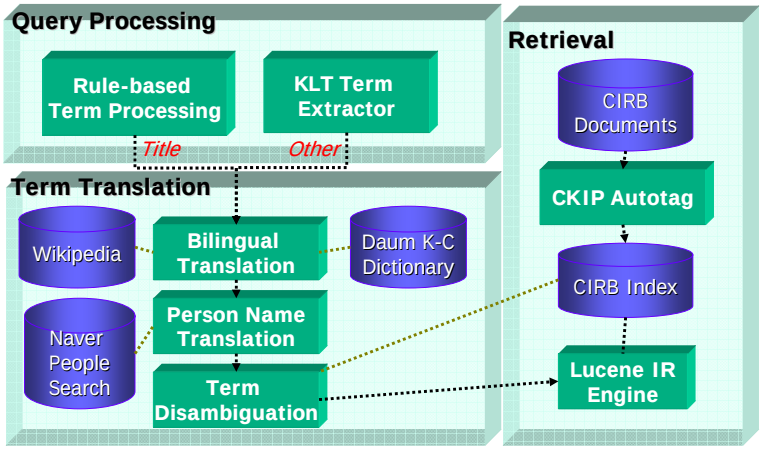
Yu-Chun Wang, Cheng-Wei Lee, Richard Tzong-Han Tsai, Wen-Lian Hsu*

***hsu@iis.sinica.edu.tw**

Academia Sinica, Taiwan, R.O.C.

We propose an architecture for retrieving Chinese documents based on Korean queries in NTCIR CLIR K-C Task. Our system uses a bilingual dictionary to perform query translation. We expand our bilingual dictionary by extracting words and their translations from the Wikipedia site, an online encyclopedia. To resolve the problem of translating Western people's names into Chinese, we propose a transliteration mapping method. We translate queries form Korean query to Chinese by using a co-occurrence method.

Architecture



Query Processing

We use two different segmentation methods, one for the title of the query and the other for other parts.

Predefined Processing Rules for Title part

Our Rules:

- ◆Split the title into several eojeols by the space characters
- ◆Remove Korean postpositions at the end of each eojeols

KLT Term Extractor for other parts

We use the KLT Term Extractor to extract vital key words and remove stop words.

KLT Term Extractor is developed by Kookmin University, Korea.

Chinese Information Retrieval

Document Indexing

- ◆CIRB 4.0 documents are pre-processed to remove noise and then segmented by CKIP AutoTag to obtain words and part-of-speech (POS).
- ◆We adopt **Lucene**, an open source information retrieval engine.
- ◆Our index is based on **Chinese characters**.

Lucene Query

One Query Example

日本 or 韓國 or 漁業 or (協定 or 條約^0.25 or 合約^0.25 or 合同^0.25)

Query Translation

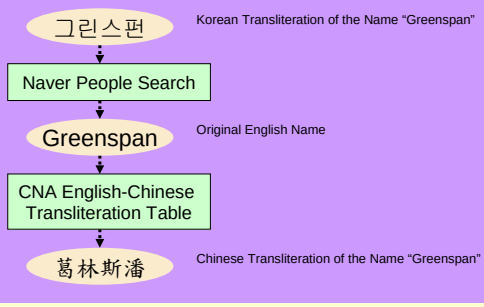
Bilingual Dictionary

We use the free online Korean-Chinese bilingual dictionary provided by the Daum Korean web site.

Wikipedia

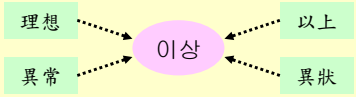
We use Wikipedia to expand our dictionaries for the proper nouns. The following is the procedure.

Person Name Translation



Term Disambiguation

Many different Chinese loanwords have the same pronunciation when written in the Hangul alphabet.



Mutual Information

$$\text{MI score}(te_{ij} \mid Q) = \sum_{x=1, x\neq i}^{|Q|} \sum_{y=1}^{Z(qt_x)} \frac{\Pr(tc_{ij}, tc_{xy})}{\Pr(tc_{ij})\Pr(tc_{xy})}$$

Performance

IASL CLIR K-C Performance

Run	Rigid		Relax	
	MAP	R-prec	MAP	R-prec
IASL-K-C-T-01	0.1118	0.1420	0.1392	0.1781
IASL-K-C-D-01	0.1022	0.1331	0.1274	0.1760

Acknowledgement

We would like to thank CKIP for providing us AutoTag for Chinese word segmentation.