IASL System for NTCIR-6
Korean-Chinese CLIR

Yu-Chun Wang
Cheng-Wei Lee
Richard Tzong-Han Tsai
Wen-Lian Hsu *
Min-Yuh Day

Intelligent Agent Systems Lab. (IASL)
Institute of Information Science, Academia Sinica, Taiwan

NTCIR-6, Tokyo, Japan, May 15-18, 2007
Outline

- IASL CLIR System Architecture
  - Query Processing (Korean)
  - Term Translation (Korean - Chinese traditional)
    - Bilingual Dictionary Translation
    - Person Name Translation
    - Term Disambiguation
  - Document Indexing (Chinese)
  - Document Retrieval (Chinese)
- NTCIR-6 CLIR Evaluation Result
- Error Analysis
- Conclusion and Future Work
CLIR System Architecture

Korean

- Korean Query
  - Title
  - Description
  - Rule-based Term Processing
  - KLT Term Extractor
  - Key Terms
  - Bi-lingual Dictionary Translation
  - People Name Translation
  - Term Disambiguation
  - Translated Chinese Terms

Chinese (Traditional)

- CIRB 4.0
- CKIP AutoTag
- Sentence Index
- Lucene Indexing
- Document Index
- Lucene Query Transformer
- Lucene Query
- Lucene IR Engine
- IR Result

Daum Korean-Chinese Dictionary
Korean Wikipedia
Naver People Search
CLIR System Architecture

Korean

Query

Rule-based Term Processing

KLT Term Extractor

Key Terms

Translated Chinese Terms

IR Result

Document Retrieval

Chinese (Traditional)

Indexing

Lucene IR Engine

Sentence Index

Lucene Query Transformer

Lucene Query

Naver Korean-Chinese Dictionary

People Name Translation

People Search

Korean Wikipedia

Term Disambiguation

Translated Chinese Terms

Title
CLIR System Architecture

Korean

- Daum Korean-Chinese Dictionary
- Korean Wikipedia

Chinese (Traditional)

- Bi-lingual Dictionary Translation
- People Name Translation
- Naver People Search
- Term Disambiguation
- Translated Chinese Terms
- Term Translation

Lucene Indexing

Sentence Index

Document Index

Lucene IR Engine

IR Result Document Retrieval

Lucene Query Transformer

Lucene Query

Korean Chinese (Traditional) Daum Korean-Chinese Dictionary

Korean Wikipedia

Bi-lingual Dictionary Translation

People Name Translation

Naver People Search

Term Disambiguation

Translated Chinese Terms

Term Translation

Retrieval
CLIR System Architecture

- Korean
  - Korean Query
  - Query Processing
  - CIRB 4.0
  - Sentence Index
  - Translated Chinese Terms

- Chinese (Traditional)
  - Document Index
  - IR Result

- Language Translations
  - Daum Korean-Chinese Dictionary
  - People Name Translation

- System Components
  - Rule-based Term Processing
  - KLT Term Extractor
  - People Search
  - Term Disambiguation
  - Term Translation
  - CKIP AutoTag
  - Lucene Indexing
  - Lucene IR Engine
  - IR Result Document Retrieval
CLIR System Architecture

Korean

4

Chinese (Traditional)

Lucene Query Transformer

Lucene Query

Lucene IR Engine

IR Result

Bi-lingual Dictionary

Translation

Daum Korean-Chinese Dictionary

Korean Wikipedia

People Name Translation

Naver

KLT Term Extractor

Rule-based Term Processing

Key Terms

Query Processing

IR Result Document Retrieval

Lucene IR Engine

Sentence Index

Document Index

CKIP AutoTag

Lucene Query Transformer

Lucene Query

Lucene IR Engine

IR Result

NTCIR-6 IASL System for NTCIR-6 Korean-Chinese CLIR
CLIR System Architecture

Korean

- Korean Query
  - Title
  - Description
- Rule-based Term Processing
- KLT Term Extractor
- Key Terms
- Daum Korean-Chinese Dictionary
- Bi-lingual Dictionary Translation
- Korean Wikipedia
- People Name Translation
- Term Disambiguation
- Translated Chinese Terms

Chinese (Traditional)

- CIRB 4.0
- CKIP AutoTag
- Sentence Index
- Lucene Indexing
- Document Index
- Document Retrieval
- Lucene Query Transformer
- Lucene Query
- Lucene IR Engine
- IR Result
Query Processing

- Pre-defined rules for the title of query:
  - Chunk the sentence with spaces and punctuations.
  - Remove Josa at the end of the terms.

- For descriptive part of a Korean query:
  - Use KLT Term Extractor (by Kookmin University) to extract vital key words and remove stop words.
CLIR System Architecture

Korean

- Korean Query
  - Title
  - Description
  - Rule-based Term Processing
  - KLT Term Extractor
  - Key Terms
  - Daum Korean-Chinese Dictionary
  - Korean Wikipedia
  - Bi-lingual Dictionary Translation
  - People Name Translation
  - Term Disambiguation
  - Translated Chinese Terms

Chinese (Traditional)

- CIRB 4.0
- CKIP AutoTag
- Sentence Index
- Lucene Indexing
- Document Index
- Lucene Query Transformer
- Lucene Query
- Lucene IR Engine
- IR Result
- Document Retrieval
Bilingual Dictionary Translation

- **Dictionary-based** translation method:
  - Daum Chinese-Korean online dictionary

- **Mapping table** to convert simplified Chinese characters to traditional Chinese ones.
The Rules for Splitting Korean Terms

- Apply the rules (based on the properties of Korean morphemes) to split a long term into several shorter terms.

<table>
<thead>
<tr>
<th>Number of Character</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ABC→A, BC</td>
</tr>
<tr>
<td></td>
<td>ABC→AB, C</td>
</tr>
<tr>
<td>4</td>
<td>ABCD→AB, CD</td>
</tr>
<tr>
<td></td>
<td>ABCD→A, BCD</td>
</tr>
<tr>
<td></td>
<td>ABCD→ABC, D</td>
</tr>
<tr>
<td>5</td>
<td>ABCDE→AB, CDE</td>
</tr>
<tr>
<td></td>
<td>ABCDE→ABC, DE</td>
</tr>
<tr>
<td>6</td>
<td>ABCDEF→AB, CD, EF</td>
</tr>
<tr>
<td></td>
<td>ABCDEF→ABC, DEF</td>
</tr>
<tr>
<td>7</td>
<td>ABCDEFG→AB, CD, EFG</td>
</tr>
<tr>
<td></td>
<td>ABCDEFG→AB, CDE, FG</td>
</tr>
<tr>
<td></td>
<td>ABCDEFG→ABC, DE, FG</td>
</tr>
<tr>
<td>8</td>
<td>ABCDEFGH→AB, CD, EF, GH</td>
</tr>
<tr>
<td>9</td>
<td>ABCDEFGHI→AB, CD, EF, GHI</td>
</tr>
<tr>
<td>10</td>
<td>ABCDEFGHIJ→AB, CD, EF, GH, IJ</td>
</tr>
</tbody>
</table>
Person Name Translation

- **Transliteration methods are not appropriate** for Korean-Chinese CLIR (Unlike Korean-English or Korean-Japanese CLIR)
  - Many Chinese characters have the same pronunciation in Korean.
  - Korean uses Japanese pronunciation to translate Japanese personal names.
  - Chinese uses Japanese Kanji characters directly.

- **Naver People Search** for person name translation processing.
  - Naver People Search is a database containing the basic profiles of famous people, including their original names.
  - If the original name is composed of Chinese characters, it will be sent to the next stage directly. (CJK person names)
  - If the original name is in English, we use the English name translation/transliteration table provided by Taiwan’s Central News Agency (CNA) to translate it into Chinese.
Term Disambiguation

- **Ambiguity** in translating Korean to Chinese
  - Since Hangul is an alphabet writing system, many different Chinese characters are written in the same Hangul characters.
  - For example

- Apply **Mutual Information** to measure correlation to choose the best translation term among translation candidates.

\[
\text{MI score}(te_{ij} \mid Q) = \sum_{x=1, x \neq i}^{n} \sum_{y=1}^{Z(qt_x)} \frac{\Pr(te_{ij}, te_{xy})}{\Pr(te_{ij}) \Pr(te_{xy})}
\]
CLIR System Architecture

Korean

- Korean Query
  - Title
  - Description
- Rule-based Term Processing
- KLT Term Extractor
- Key Terms
- Bi-lingual Dictionary Translation
- People Name Translation
  - Naver People Search
  - Term Disambiguation
  - Translated Chinese Terms

Chinese (Traditional)

- CIRB 4.0
- CKIP AutoTag
- Sentence Index
- Lucene Indexing
  - Document Index
- Lucene Query Transformer
  - Lucene Query
    - Lucene IR Engine
      - IR Result
      - Document Retrieval

- Daum Korean-Chinese Dictionary
- Korean Wikipedia
- People Name Translation
Chinese Document Indexing and Lucene IR

- CIRB 4.0 documents are pre-processed to remove noise and then segmented by CKIP AutoTag.
- Lucene IR engine
  - Index Chinese documents based on Chinese characters.

- The translated Chinese query from the original Korean query will be transformed into Lucene query to proceed IR.
  - If a term has different translation candidates, the weight of the candidate with highest mutual information score will be increased by 1 by the boost operator ^.
## NTCIR-6 CLIR Evaluation Result of IASL’s Runs

<table>
<thead>
<tr>
<th>Run</th>
<th>Rigid MAP</th>
<th>R-prec</th>
<th>Relax MAP</th>
<th>R-prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>IASL-K-C-T-01</td>
<td>0.1118</td>
<td>0.1420</td>
<td>0.1392</td>
<td>0.1781</td>
</tr>
<tr>
<td>IASL-K-C-D-01</td>
<td>0.1022</td>
<td>0.1331</td>
<td>0.1274</td>
<td>0.1760</td>
</tr>
</tbody>
</table>
Error Analysis (1/3) – Problems of Bilingual Dictionaries

- The dictionaries do not always have the proper translation candidates of the words and terms in queries.

  - The word “암” (cancer) is translated as “岩” (rock), “庵” (nunnery), and “雌” (female), but no correct translation, i.e., “癌” (cancer).
Error Analysis (2/3) — Different Phraseology Used in Taiwan and China

- The Daum Korean-Chinese dictionary was written for people studying Mainland Chinese (Simplified Chinese).
  - The CIRB 4.0 document collection contains Taiwanese newspapers (Traditional Chinese).

- The characters, vocabulary and grammar used in Taiwan and China are slightly different.
  - The differences can make IR difficult.
    - The term “휴대폰” (mobile phone) is translated into Mainland Chinese word as “移動電話”; however, the correct word used in Taiwan is “手機”.
    - The word “유전자” (gene) is translated to “遺傳子”, not to correct word “基因” used in Taiwan.
Error Analysis (3/3) —
Different Expressions Used in Korean and Chinese

- **Different expressions** used in Korean and Chinese may cause translation problems.
  - The word “10대” refers to people aged between 10 and 19 in Korean.
  - The corresponding translation of the word “10대” in Chinese is “青少年” (teenager).
    - Our system translates to “10代” (ten generations).

- **Abbreviations** used in Chinese.
  - “왜국인 노동자” (foreign worker) is translated into “外國人勞工” (foreign worker) by our system.
    - In Taiwanese newspapers, the abbreviation “外勞” (foreign worker) is used more frequently.
Conclusion and Future Work

- IASL Korean-Chinese CLIR system: the only entry in the NTCIR-6 CLIR K-C task.
  - Query-translation approach
  - Using Naver People Search and CNA transliteration table

- Our K-C translation method is effective
  - Limitations of the dictionaries
  - Different phraseology used in Taiwan and China
  - Different expressions used in Chinese and Korean

- Future Work
  - Applying a Chinese thesaurus
  - Query expansion method
Q&A

IASL System for NTCIR-6
Korean-Chinese CLIR

Yu-Chun Wang (王昱鈞)
Cheng-Wei Lee (李政緯)
Richard Tzong-Han Tsai (蔡宗翰)
Wen-Lian Hsu* (許聞廉)
Min-Yuh Day (戴敏育)

Intelligent Agent Systems Lab. (IASL)
Institute of Information Science, Academia Sinica, Taiwan

NTCIR-6, Tokyo, Japan, May 15-18, 2007