



# IASL System for NTCIR-6 Korean-Chinese CLIR

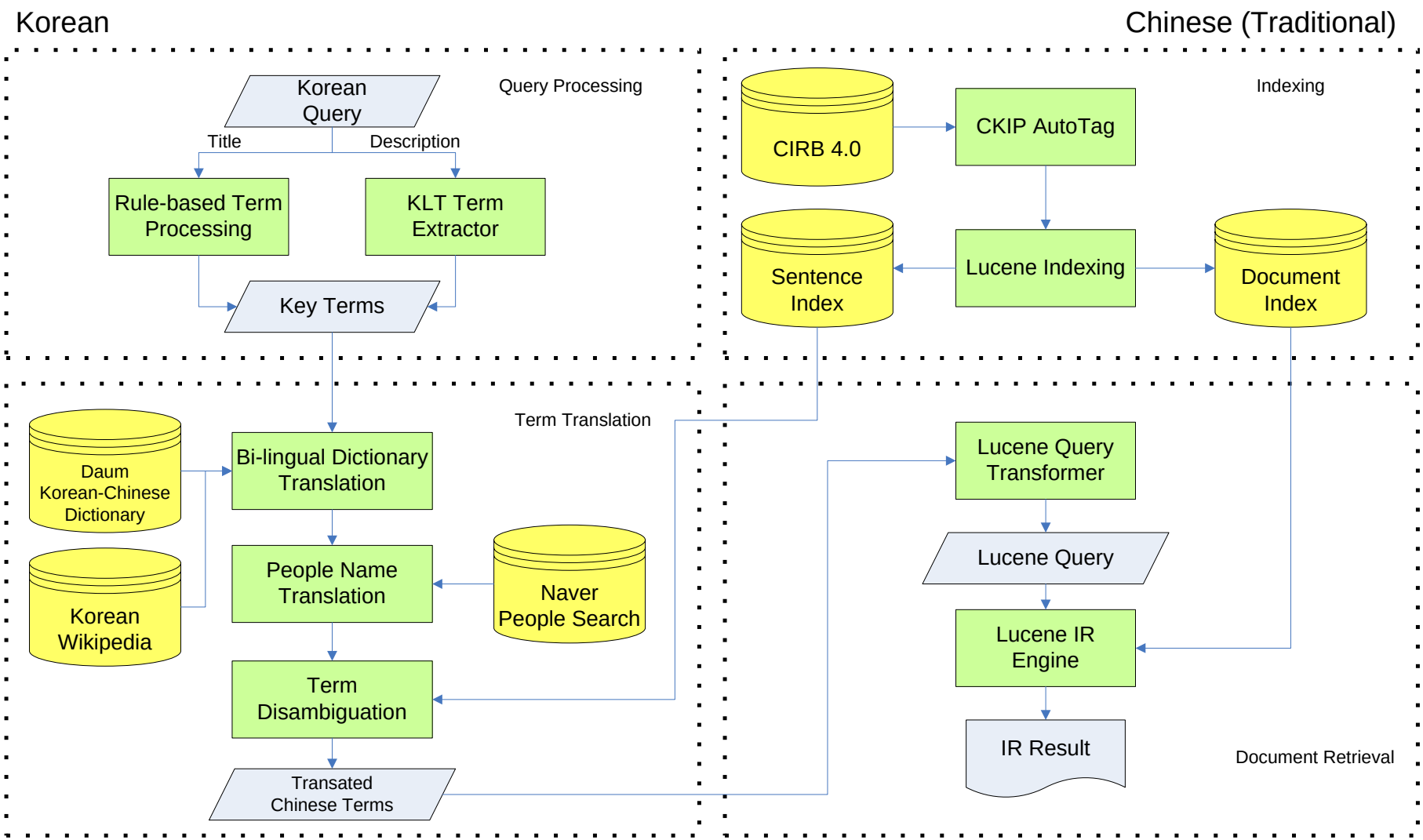
Yu-Chun Wang
Cheng-Wei Lee
Richard Tzong-Han Tsai
Wen-Lian Hsu *
Min-Yuh Day

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

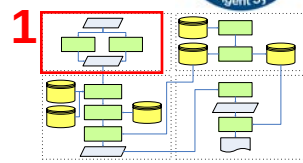
# Outline

- 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



# CLIR System Architecture




Korean

Chinese (Traditional)

Korean Query

Query Processing

Indexing

1

Korean Query

Rule-based Term Processing

KLT Term Extractor

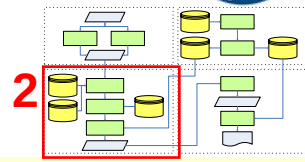
Key Terms

Transated Chinese Terms

IR Result

Document Retrieval

# CLIR System Architecture




Korean

Chinese (Traditional)

2

Term Translation

Daum Korean-Chinese Dictionary

Korean Wikipedia

Bi-lingual Dictionary Translation

People Name Translation

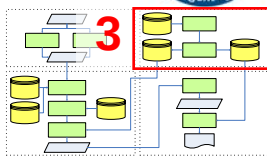
Naver People Search

Term Disambiguation

Transated Chinese Terms

Retrieval

# CLIR System Architecture




Korean

Chinese (Traditional)

Korean Query

Query Processing

Indexing

CKIP AutoTag

3

CIRB 4.0

CKIP AutoTag

Sentence Index

Lucene Indexing

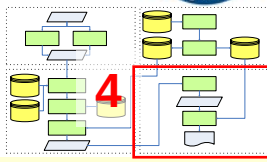
Document Index

IR Result

Document Retrieval

Transated Chinese Terms

# CLIR System Architecture




Korean

Chinese (Traditional)

4

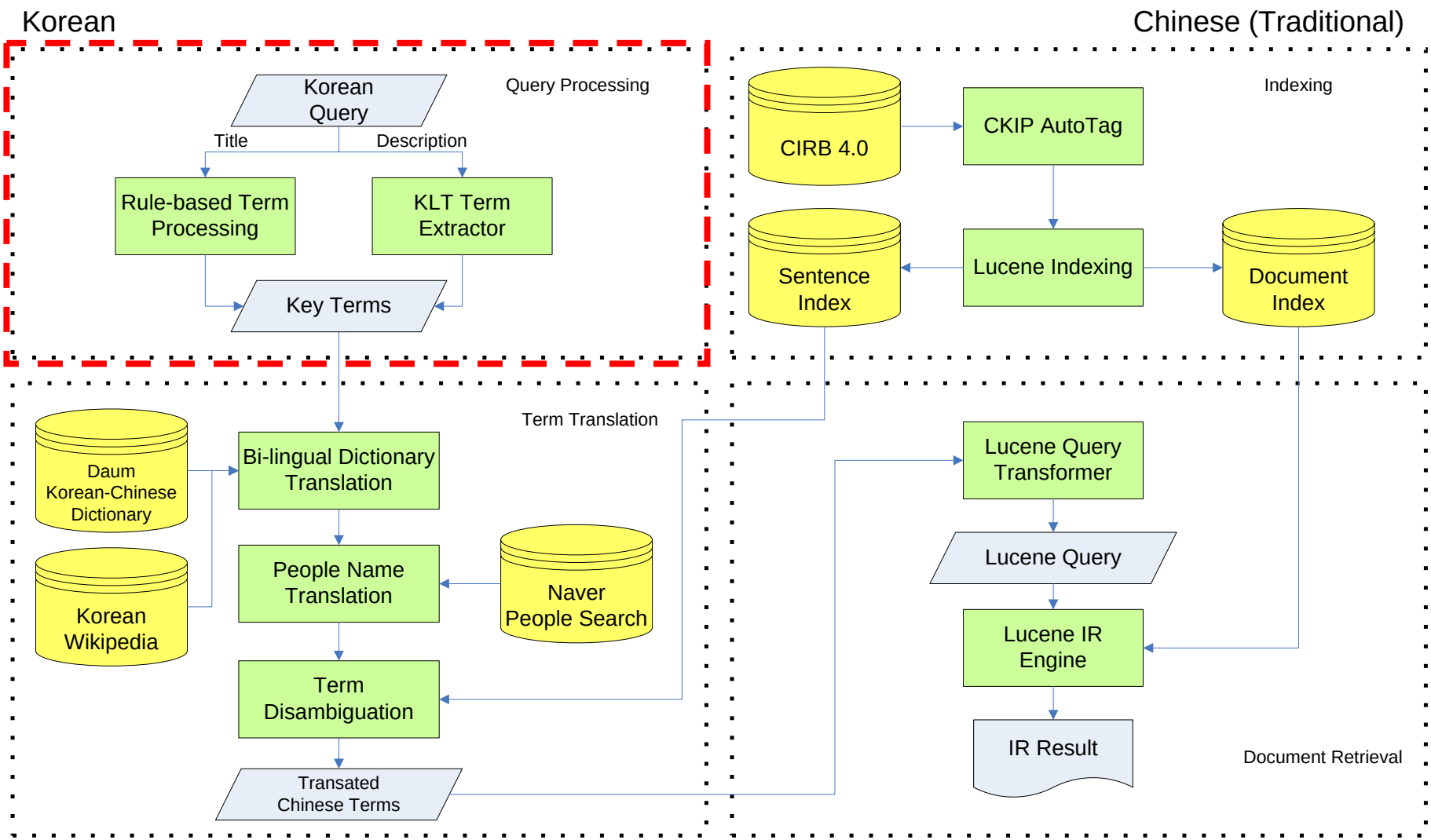
Lucene Query Transformer

Lucene Query

Lucene IR Engine

IR Result

# CLIR System Architecture

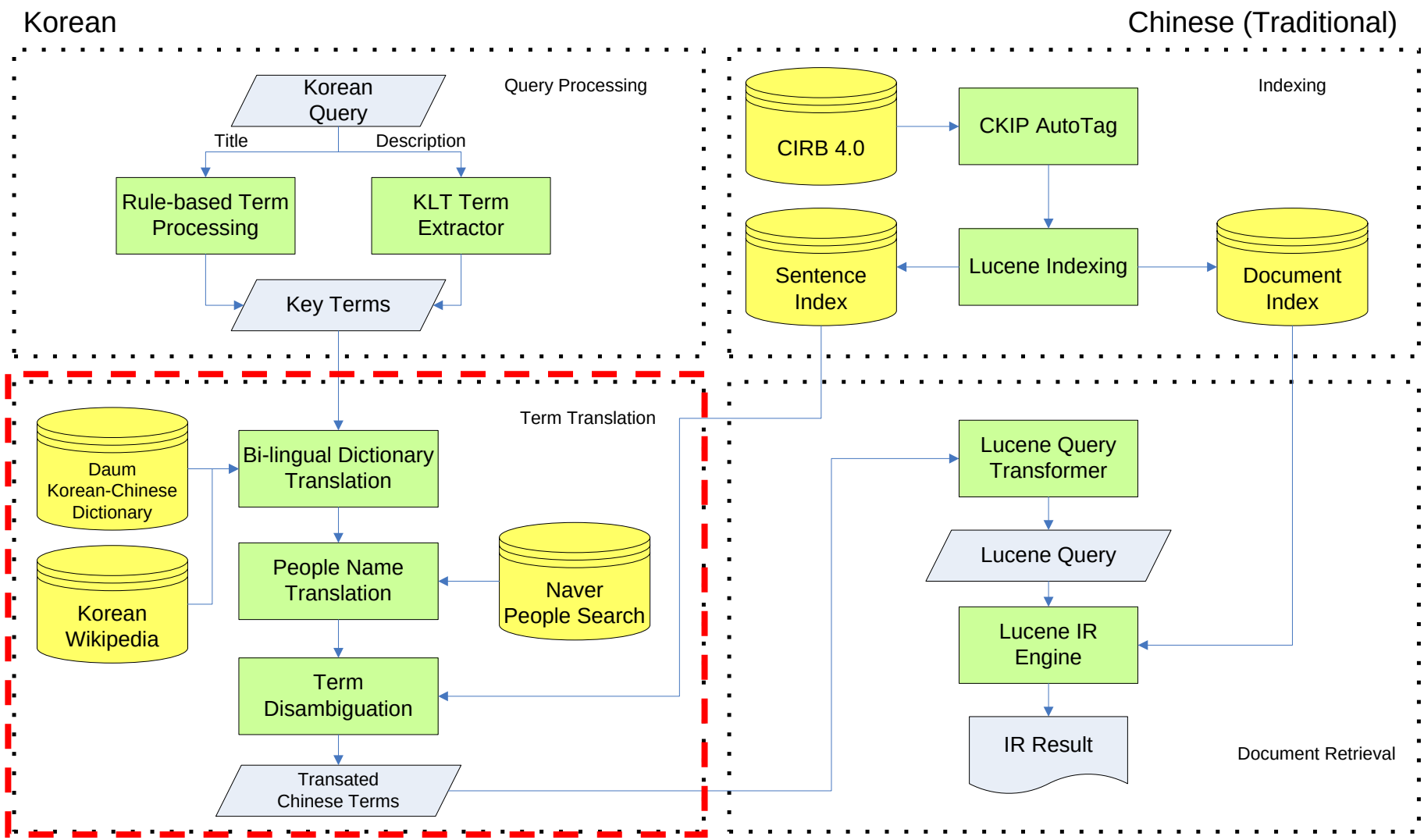


# Query Processing

- 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



# Bilingual Dictionary Translation

- Dictionary-based translation method:
  - Daum Chinese-Korean online dictionary
  - Korean Wikipedia with inter-language link to Chinese Wikipedia
- 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.

| Number of Character | Separation |
|---|---|
| 3 | ABC→A, BC<br>ABC→AB, C |
| 4 | ABCD→AB, CD<br>ABCD→A, BCD<br>ABCD→ABC, D |
| 5 | ABCDE→AB, CDE<br>ABCDE→ABC,DE |
| 6 | ABCDEF→AB, CD, EF<br>ABCDEF→ABC, DEF |
| 7 | ABCDEFG→AB, CD, EFG<br>ABCDEFG→AB, CDE, FG<br>ABCDEFG→ABC, DE, FG |
| 8 | ABCDEFGH→AB, CD, EF, GH |
| 9 | ABCDEFGHI→AB, CD, EF, GHI |
| 10 | ABCDEFGHIJ→AB, CD, EF, GH, IJ |

# 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.

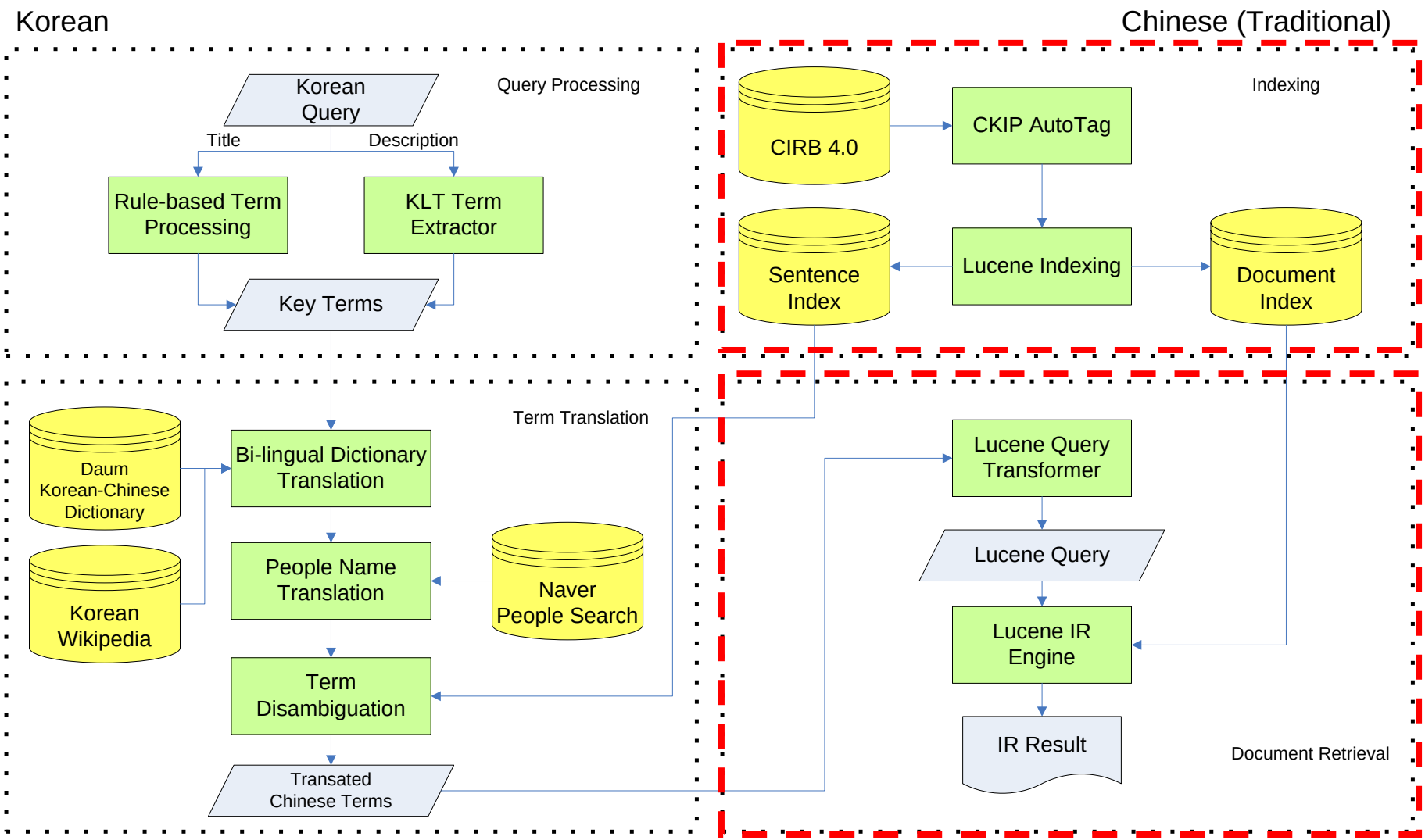
# Term Disambiguation

- 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
    - The Hangul word "이상" corresponds to four different Chinese words: "理想"(ideal), "異常"(unusual), "以上"(above), "異狀" (indisposition).

- 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



# 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

| 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 |

# 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.
  - "외국인 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 general Korean-Chinese dictionary and Wikipedia
  - 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*