# *POSTECH at NTCIR-4: CJKE Monolingual and Korean-related Cross-Language Retrieval Experiments*

**Jun. 2, 2004**

In-Su Kang*, Seung-Hoon Na, Jong-Hyeok Lee

***Knowledge and Language Engineering Laboratory***

***Dept. of Computer Science & Engineering***

***Pohang University of Science and Technology, KOREA***

# Contents


- CJK Single Language IR
  - Motivation
  - Coupling words and n-grams
  - Coupling at a ranked list level
  - Term Extraction
  - NTCIR-4 results
  - Observations
- Korean-related Cross-Language IR
- Conclusion and Future Work

# Motivation

- CJK monolingual IR
  - Word segmentation is nontrivial
  - Words vs. n-grams

| | Words | N-grams |
|---|---|---|
| Lexical Term Space | Incomplete | Complete |
| Concept Specificity | Concentrated | Distributed |
| Weak point | Under-generation | Over-generation |

  - Combination of words and n-grams is advocated
    - We investigate ***a coupling method of words and n-grams***
- English monolingual IR (not described in this presentation)
  - Develop a new phrasal indexing unit

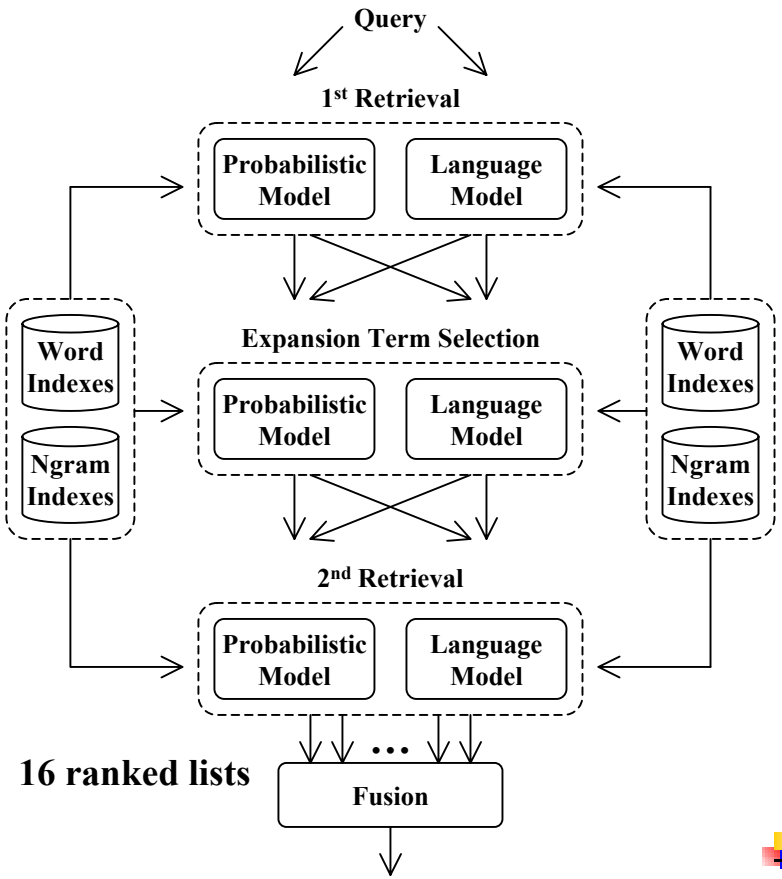
# Coupling of Words and N-grams

- Coupling methods

| Coupling Stage | Coupling Unit | | # of Indexes |
|---|---|---|---|
| Index creation | Index term | | One |
| Term weighting | TF | Sum | |
| | DF | Sum, or Union | Two |
| | Term weight | Interpolation | |
| Ranked list | Document score | Sum | Two |

  - Experiments using NTCIR-3 Korean test set
    - All but *coupling at a ranked list level* were not remarkable
- Coupling at a ranked list level
  - Basic idea → Generate & merge several ranked lists with different retrieval characteristics on words and n-grams

# Coupling at a Ranked List Level (1/2)




- Generation of ranked lists
  - Indexing units
    - Words
    - N-grams
  - 1st and 2nd retrieval models
    - Okapi probabilistic model
    - Jelinek-Mercer language model
  - Expansion term selection
    - Robertson selection value
    - Ponte's ratio formula
- Fusion by simple summation

# Coupling at a Ranked List Level (2/2)

- Selection of top 3 ranked lists out of 16
  - Selection measure
    - MAP on NTCIR-3 Korean test set
  - Selection constraint
    - Include at least one for each of words and n-grams

| | Index Unit | | |
|---|---|---|---|
| | **Word** | **N-gram** | |
| 1st Retrieval | P | P | L |
| Expansion term selection | L (Ponte's) | P (Rebertson's) | L (Ponte's) |
| 2nd Retrieval | P | P | L |
| Abbreviated notation | wPLP | nPPP | nLLL |

# Term Extraction

- Index terms

| | Terms | Stoplist |
|---|---|---|
| Chinese | Bi-gram, word | None |
| Japanese | Bi-gram, word | None |
| Korean | Bi-gram, word | 374 stopwords |

- CJK word extraction
  - By CJK taggers developed at our laboratory
- Bi-grams
  - For Japanese, bi-grams were generated for a sequence of the same character class (Hiragana, Katagana, Kanji)

# NTCIR-4 Results (Chinese)

- Chinese single language IR

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| 1st Retrieval | nP-- | 0.2297 | 0.2069 | 0.2562 | 0.2855 | 0.2911 |
| | nL-- | 0.2050 | 0.1823 | 0.2365 | 0.2708 | 0.2809 |
| | wP-- | 0.1603 | 0.1533 | 0.1789 | 0.2281 | 0.2358 |
| 2nd Retrieval | nPPP | 0.2532 | 0.2398 | 0.2681 | 0.2983 | 0.3060 |
| | nLLL | 0.2699* | 0.2686* | 0.2856* | 0.3019* | 0.3046 |
| | wPLP | 0.1853 | 0.2016 | 0.2049 | 0.2503 | 0.2693 |
| **Fusion** | **wPLP+nPPP+nLLL** | **0.2584 (-4.3%)** | **0.2535 (-5.6%)** | **0.2703 (-5.4%)** | **0.2968 (-1.7%)** | **<u>0.3103</u>* (+1.4%)** |
| **NTCIR-4 MAX** | | 0.3799 | 0.3880 | 0.3103 | | |

*** : the best performance for the query type**

**_ : NTCIR-4 best performance**

# NTCIR-4 Results (Japanese)

- Japanese single language IR

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| 1st Retrieval | nP-- | 0.3650 | 0.3424 | 0.3496 | 0.4346 | 0.4570 |
| | nL-- | 0.3260 | 0.3101 | 0.3141 | 0.4274 | 0.4435 |
| | wP-- | 0.3647 | 0.3715 | 0.3426 | 0.4439 | 0.4561 |
| 2nd Retrieval | nPPP | 0.3844 | 0.3842 | 0.3926 | 0.4539 | 0.4856 |
| | nLLL | 0.4056 | 0.4282* | 0.4207* | 0.4924* | 0.5024* |
| | wPLP | 0.4226* | 0.4103 | 0.3806 | 0.4715 | 0.4875 |
| **Fusion** | **wPLP+nPPP+nLLL** | **0.4211 (-0.4%)** | **0.4119 (-3.8%)** | **0.4105 (-2.4%)** | **0.4741 (-3.7%)** | **<u>0.4963</u> (-1.2%)** |
| **NTCIR-4 MAX** | | 0.4864 | 0.4838 | 0.4963 | | |

*** : the best performance for the query type**

**_ : NTCIR-4 best performance**

# NTCIR-4 Results (Korean)

- Korean single language IR

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| 1st Retrieval | nP-- | 0.4515 | 0.4198 | 0.4450 | 0.5249 | 0.5598 |
| | nL-- | 0.4091 | 0.3674 | 0.4081 | 0.4896 | 0.5318 |
| | wP-- | 0.4285 | 0.4184 | 0.4370 | 0.5111 | 0.5383 |
| 2nd Retrieval | nPPP | 0.4660 | 0.4347 | 0.4499 | 0.5610 | 0.6040 |
| | nLLL | 0.4967 | 0.4623 | 0.4496 | 0.5592 | 0.5873 |
| | wPLP | 0.4900 | 0.4771 | 0.4611 | 0.5806 | 0.5859 |
| **Fusion** | **wPLP+nPPP+nLLL** | **0.5226*** **(+5.2%)** | **0.4885*** **(+2.4%)** | **0.4846*** **(+5.1%)** | **0.5932*** **(+2.2%)** | **<u>0.6212</u>*** **(+2.8%)** |
| **NTCIR-4 MAX** | | 0.5361 | 0.5097 | 0.6212 | | |

*** : the best performance for the query type**

**_ : NTCIR-4 best performance**

# Observations

- Words vs. n-grams
  - Coupling at a ranked list level maybe language-dependent
    - At NTCIR-4, only Korean SLIR was successful
      - Chinese : -5.6% ~ 1.4% over 2nd retrieval best
      - Japanese : -3.8% ~ -0.4% over 2nd retrieval best
      - Korean : 2.2%~ 5.2% over 2nd retrieval best
    - Our top 3 ranked lists were selected based on NTCIR-3 Korean test set
- Okapi vs. LM (language model)
  - At 1st retrieval, Okapi was better than LM
  - At 2nd retrieval, LM parallels or outperforms Okapi

# Contents


- CJK Single Language IR
- Korean-related Cross-Language IR
  - Motivation
  - QT vs. DT
  - Hybrid approach of QT and DT
  - Transliteration-based DT
  - Dictionary statistics
  - NTCIR-4 results
  - Observations
- Conclusion and Future Work

# Motivation

- Cross-language IR
  - Query translation
    - Widespread, and much explored
  - Document translation
    - Computationally expensive, and barely attempted
      - MT system or statistical translation model
    - At NTCIR-4, we tried ***a simple dictionary-based translation***
- Our interests
  - ***Combining query translation and document translation***
  - ***Coupling words and n-grams in CLIR***

# Language Translation

- Default query translation (QT)
  - Dictionary-based
    - Source-to-target bilingual dictionary
  - Target language query
    - Unstructured sequence of all translations of source language query terms
- Default document translation (DT)
  - Dictionary-based
    - Target-to-source bilingual dictionary
  - Source language document
    - Unstructured sequence of all translations of target language document terms

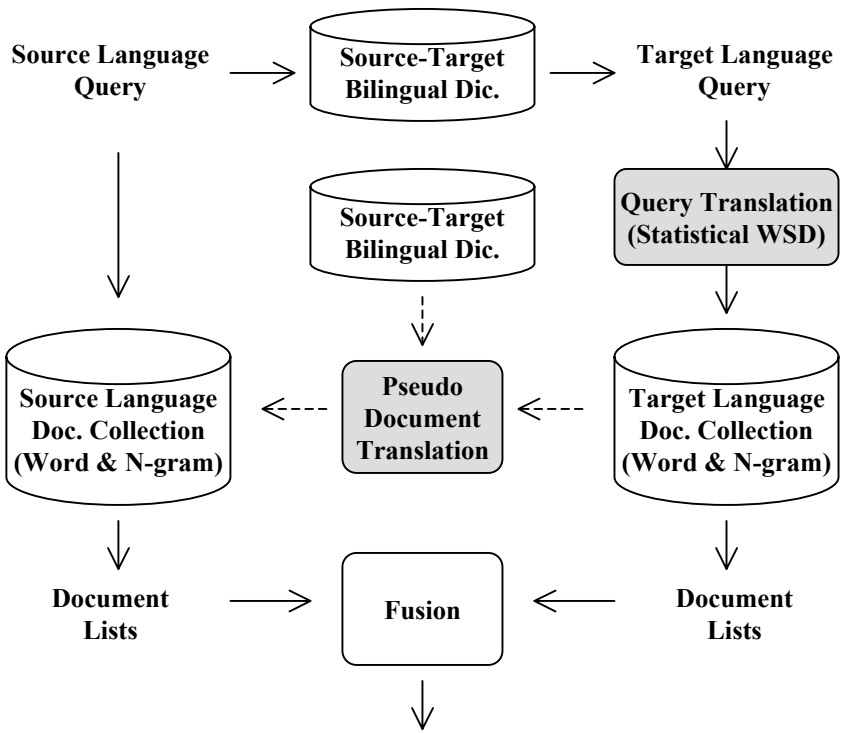
# Default QT vs. DT

- Disambiguation effect of QT and DT

| | Disambiguation context | | Disambiguation Effect |
|---|---|---|---|
| | Query | Document | |
| Default QT | Noisy | Clean | Resolves source language translation ambiguity |
| Default DT | Clean | Noisy | Resolves target language translation ambiguity |

- Hybrid of QT and DT
  - Different translation directions of the same language pair may differently influence translation disambiguation of queries

# Hybrid Approach of QT and DT




- Coupling at a ranked list level

| | QT | DT |
|---|---|---|
| KC | nPLP | nPLP |
| KJ | wPLP | nPLP |
| CK, JK | wPLP + nPLP | None |

- nPLP, wPLP
  - Selected from our experiments on NTCIR-3 Korean-to-Japanese CLIR test set

# Transliteration-based DT (1/2)

- CJK languages
  - Share ideographic Chinese characters
    - Chinese : Hanzi
    - Japanese : Kanji
    - Korean : Hanja
- In Korean text
  - Chinese characters are written in *Hangul*
    - *Hangul* : a Korean alphabet, not ideographic, but phonetic
  - M-to-1 mapping b/w Chinese characters and *Hangul*
    - 漢代(Han dynasty) → 한대
    - 寒帶(the frigid zone) → 한대

# Transliteration-based DT (2/2)

- Transliteration-based DT (in KC or KJ CLIR)
  - Chinese characters are transliterated into Hangul
  - The resulting Hangul sequence is indexed
- Advantages
  - Alleviates vocabulary mismatch problem
    - 고궁 → 古宮 (an old palace), in a KJ dictionary
    - 故宮 (an old palace), in Japanese documents
    - Their Hangul transliterations can be matched with a query term 고궁
      - 古宮 → 고궁, and 故宮 → 고궁
  - Mitigate unknown word problem
    - Unknown query term 김대중 (a former Korean president)
    - Can be matched with a document term 金大中 by Hangul transliteration

# Statistics of Bilingual Dictionaries

- Bilingual dictionaries
  - Extracted from transfer dictionaries of our lab's MT systems
    - COBALT-JK/KJ (**Co**llocation-**Ba**sed **L**anguage **T**ranslator b/w **K**orean and **J**apanese)
    - TOTAL (**T**ranslator **O**f **T**hree **A**sian **L**anguages)

| | # of Translation Pairs | # of Source Language Entries | Dictionary Ambiguity |
|---|---|---|---|
| KC | 113,312 | 81,750 | 1.39 |
| CK | 127,560 | 109,614 | 1.16 |
| KJ | 420,650 | 303,199 | 1.39 |
| JK | 434,672 | 399,220 | 1.09 |

# NTCIR-4 Results (KC and KJ)

- CLIR using Korean as a query language

(%): improvement

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| KC | QT(wP–) | 0.1436 | 0.1456 | 0.1584 | 0.1665 | 0.1778 |
| | DT(nP–) | 0.1551 (8.0%) | 0.1448 (-0.5%) | 0.1567 (-1.1%) | 0.1937 (16.3%) | 0.2057 (15.7%) |
| | QT(wP–)+DT(nP–) | 0.1687 (8.8%) | 0.1731 (18.9%) | 0.1763 (11.4%) | 0.1992 (2.8%) | 0.2089 (1.6%) |
| | QT(wPLP) + DT(nPLP) | **0.1892** (12.2%) | **0.1869** (7.9%) | **0.2028** (15.0%) | **0.2378** (19.4%) | **0.2469** (18.2%) |
| KJ | QT(wP–) | 0.2861 | 0.3039 | 0.3000 | 0.3763 | 0.3905 |
| | DT(nP–) | 0.3165 (10.6%) | 0.3207 (5.5%) | 0.3140 (4.7%) | 0.3909 (3.9%) | 0.4039 (3.4%) |
| | QT(wP–)+DT(nP–) | 0.3234 (2.2%) | 0.3362 (4.8%) | 0.3241 (3.2%) | 0.4098 (4.8%) | 0.4229 (4.7%) |
| | QT(wPLP) + DT(nPLP) | **0.3602** (11.4%) | **0.3601** (7.1%) | **0.3713** (14.6%) | **0.4471** (9.1%) | **0.4473** (5.8%) |

# Observations (KC and KJ)

- Overall, a default DT was better than a default QT
  - QT (KC or KJ) is more ambiguous than DT (CK or JK)
  - Transliteration of DT may improve recall
- A hybrid of QT and DT outperforms QT or DT alone
  - QT and DT has different disambiguation effects on queries
- Post-translation feedback works well

| | KC | | | KJ | | |
|---|---|---|---|---|---|---|
| **QT** | 0.1584 | | | 0.3314 | | |
| **DT** | 0.1712 | 8.09% | 8.09% | 0.3492 | 5.38% | 5.38% |
| **QT + DT (no feedback)** | 0.1852 | 8.20% | 16.96% | 0.3633 | 4.03% | 9.63% |
| **QT + DT (feedback)** | 0.2127 | 14.83% | 34.31% | 0.3972 | 9.34% | 19.87% |

# NTCIR-4 Results (CK and JK)

- CLIR using Korean as a document language
  - Coupling effect of words and n-grams (%): improvement

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| CK | QT(wP–) | 0.3466 | 0.3193 | 0.3364 | 0.4004 | 0.4299 |
| | QT(nP–) | 0.3572 (3.1%) | 0.3342 (4.7%) | 0.3466 (3.0%) | 0.4099 (2.4%) | 0.4355 (1.3%) |
| | QT(wP–)+QT(nP–) | 0.3663 (2.5%) | 0.3463 (3.6%) | 0.3557 (2.6%) | 0.4259 (3.9%) | 0.4538 (4.2%) |
| | QT(wPLP) + QT(nPLP) | **0.4343** (18.6%) | **0.4314** (24.6%) | **0.4083** (14.8%) | **0.5060** (18.8%) | **0.5138** (13.2%) |
| JK | QT(wP–) | 0.3559 | 0.3431 | 0.3451 | 0.4243 | 0.4450 |
| | QT(nP–) | 0.3490 (-1.9%) | 0.3501 (2.0%) | 0.3587 (3.9%) | 0.4536 (6.9%) | 0.4607 (3.5%) |
| | QT(wP–)+QT(nP–) | 0.3634 (2.1%) | 0.3666 (4.7%) | 0.3833 (6.9%) | 0.4632 (2.1%) | 0.4773 (3.6%) |
| | QT(wPLP) + QT(nPLP) | **0.4559** (25.5%) | **0.4306** (17.5%) | **0.4593** (19.8%) | **0.5383** (16.2%) | **0.5446** (14.1%) |

# Observations (CK and JK)

- N-grams (nP--) are better than words (wP--)
  - N-grams are robust to segmentation errors
    - So, alleviates missing word problem in CLIR
- A hybrid of words and n-grams (wP-- + nP--)
  - Words and n-grams collaboratively help in CLIR
- Post-translation feedback works well

| | CK | | | JK | | |
|---|---|---|---|---|---|---|
| **QT(wP–)** | 0.3665 | | | 0.3827 | | |
| **QT(nP–)** | 0.3767 | 2.77% | 2.77% | 0.3944 | 3.07% | 3.07% |
| **QT(wP–)+QT(nP–)** | 0.3896 | 3.43% | 6.30% | 0.4108 | 4.14% | 7.34% |
| **QT(wPLP) + QT(nPLP)** | 0.4588 | 17.75% | 25.17% | 0.4857 | 18.25% | 26.93% |

# NTCIR-4 Results (SLIR vs. CLIR)

- SLIR vs. CLIR
  - CLIR is compared with SLIR best performance
    - Note that most literatures compare CLIR with SLIR baseline

| | SLIR | CLIR | % of SLIR |
|---|---|---|---|
| KC | 0.2779 (CC) | 0.2127 | 0.76 |
| KJ | 0.4428 (JJ) | 0.3972 | 0.90 |
| CK | 0.5420 (KK) | 0.4588 | 0.85 |
| JK | 0.5420 (KK) | 0.4857 | 0.90 |

Each figure : Average of AvgPre over T,D,C,DN, and TDNC

# Conclusion and Future Work

- CJK monolingual IR
  - Coupling of words and n-grams at a ranked list level
- Korean-related CLIR
  - A simple dictionary-based DT, and transliteration-based DT
  - A hybrid approach of QT and DT even at its default mode
    - Performs collaboratively
- In future
  - More analysis of NTCIR-4 results such as
    - Query-by-query analysis
    - Language-dependent coupling of words and n-grams
    - Net effect of transliteration-based DT