Overview of CLIR Task at the Fourth NTCIR Workshop 4

Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, Sung Hyon Myaeng, Koji Eguchi

Design of CLIR Task

• Purpose
  - To promote researches of cross-lingual information retrieval (CLIR) on East-Asian languages

• Languages
  - Chinese (C), Japanese (J), Korean (K), English (E)

• Subtasks
  - Multilingual CLIR (MLIR) - e.g., C-JE
  - Bilingual CLIR (BLIR) - e.g., C-J
  - Single Language IR (SLIR) - e.g., C-C
  - Pivot Bilingual CLIR (PLIR) - e.g., C-E-J

Test Collection

• Document sets - News articles (1998-99)
  - Chinese: 381,681 docs
  - Japanese: 596,058 docs
  - Korean: 245,438 docs
  - English: 347,550 docs

• Queries - 60 topics
  - TITLE only run (T-run), DESC only run (D-run), other runs

Submission of results

• 26 groups submitted results
  - From Australia, Canada, China (including Hong Kong), Japan, Korea, Singapore, Switzerland, Taiwan, USA (9 countries)
  - No. of runs:
    - SLIR: 182 runs from 19 groups
    - BLIR (or PLIR): 149 runs from 17 groups
    - MLIR: 75 runs from 13 groups
    - TOTAL: 368 runs

Evaluation (1)

• Measures
  - Officially using standard output from tre_c eval software
    - Mean average precision (MAP), R-precision, Recall-Precision graph, etc.

Relevance Judgments

• Use of standard pooling method
  - Top-ranked documents from each run were merged, and judged

• Multi-grade judgments
  - “S: highly relevant,” “A: relevant,” “B: partially relevant,” “C: irrelevant”
  - Reducing to binary judgments (REC, eval)
    - Rigid relevance: S+A
    - Relaxed relevance: S+A+B

Techniques (1)

• Indexing methods
  - Morphological analyzers
  - Matching with MRD
  - Overlapping bigram, etc.

• Removing stopwords
  - Some groups tried to remove stopwords identified by term frequency, from documents or queries

• Decompounding
  - Korean and Japanese compound words were decomposed by special techniques

• Query vs. Document translation
  - Most of groups used query translation approach
  - Two groups employed a combination of document translation and query translation

• Translation methods
  - MT systems
  - Bilingual dictionaries
  - Parallel corpora

• Resources for translation
  - Various translation resources were used (see task overview in the working notes)

Techniques (2)

• Translation disambiguation
  - Using parallel corpus
  - Using co-occurrence statistics in the target documents collection
  - Using the number of Web pages including a pair of translation candidates
  - Using a probabilistic method based on a language model
  - Using Web directory
  - Pre-translation expansion

• Out-of-vocabulary problem
  - Some groups used Web resources for specifying translations for unknown terms

• Transliteration
  - Two groups applied transliteration techniques for unknown terms

• Cognate matching
  - Converting character codes with no translation

• Retrieval models
  - Okapi BM11 and BM25, vector space model (VSM), logistic regression model, INQUERY, PIRCS, language model (LM), etc.

• Query expansion techniques
  - Most of groups used pseudo-relevance feedback (PRF)
  - Expanded using knowledge ontology
  - Expanded using statistical thesaurus

• Merging strategies for MLIR
  - Round robin strategy, raw-score merging, normalized-score method, z-score, normalized-by-top-k strategy

Other

• Other various techniques were tried to be used such as multi-word term translation, estimation of translation probabilities, combination of MT systems, data fusion, etc.

Evaluation (2)

• SLIR: K-K-D (Rigid) - top 8 groups
  - MAP
    - C-KJE: 0.3255, J-KJE: 0.3204
  - precision
    - J-J: 0.60%, C-KJE: 0.72%
  - recall
    - J-J: 0.60%, C-KJE: 0.72%

• BLIR - Comparison of MAP values between best SLIR and best BLIR runs (D-run, Rigid)
  - C-C: 0.3255, J-J: 0.3204

Evaluation (3)

• BLIR
  - 40 runs using pivot approach were submitted (including MLIR runs)
  - Run types: K-E, C-E-J, E-C, C-E, K-C-J
  - Performance of pivot language approach is lower than that of non-pivot runs (except C-J-E runs)

• MLIR - Best runs
  - MAP
    - C-CE-J: 0.0923, J-CE-J: 0.1566, E-C-J: 0.1604, J-C-E: 0.1296, E-J-C: 0.1766

Conclusion

• Various techniques for improving search performance were used.

• BLIR on Chinese document sets shows relatively low performance. Meanwhile, BLIR on Korean documents seems to reach at a very high level.

• Performance of pivot language approach is lower than that of non-pivot runs.

• Performance of MLIR was low.