Overview of CLIR Task at the Fifth NTCIR Workshop

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

- Design of CLIR Task
- Test Collection
- Submission of results
- Relevance judgments
- Techniques
- Evaluation
- Plan for the next workshop
Design of CLIR Task

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

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

• Subtasks
  – *Multilingual CLIR (MLIR)*: e.g., C - CJKE
  – *Bilingual CLIR (BLIR)*: e.g., C - J
  – *Single Language IR (SLIR)*: e.g., C - C
Test Collection

• **Document sets** – News articles (2000-01)
  – Chinese: 901,446 docs
  – Japanese: 858,400 docs
  – Korean: 220,374 docs
  – English: 259,050 docs

• **Queries** – 50 topics
  – <TITLE>-only run (T-run), <DESC>-only run (D-run), other runs
Submission of results

• 24 groups submitted results
  – From Australia, Canada, China (including Hong Kong), Finland, Japan, Korea, Netherlands, Singapore, Spain, Switzerland, Taiwan, USA (13 countries and regions)

• No. of runs
  – SLIR: 201 runs from 18 groups
  – BLIR (or PLIR): 153 runs from 12 groups
  – MLIR: 25 runs from 2 groups
  – TOTAL: 379 runs
Relevance Judgments

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

• Multi-grade judgments

• Reducing to binary judgments (trec_eval)
  – *Rigid relevance*: S+A
  – *Relaxed relevance*: S+A+B

Unfortunately, we could not compute multi-grade relevance based indicators (DCG, Q-measure)
Techniques (1)

- Indexing methods for CJK text
  - Overlapping bi-gram
  - Word-based indexing
    - Matching with MRD
    - Morphological analyzer
  - Hybrid
Techniques (2)

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

• **Query vs. Document translation**
  – Most of groups used query translation approach
  – One group tried document translation (by MT)

• **Translation method**
  – MT systems
  – Bilingual dictionaries
Techniques (3)

• Translation disambiguation
  – Using co-occurrence statistics in the target documents collection (PIRCS, RMIT)
  – Using Web search engine (ISCAS)
  – Partial disambiguation (TSB)
Techniques (4)

• **Out-of-vocabulary problem**
  – Some groups (ISCAS, RMIT) used Web resources for specifying translations for unknown terms
Techniques (5)

• Retrieval models
  – Okapi BM25, vector space model (VSM), logistic regression model, PIRCS, language model (LM), etc.

• Query expansion techniques
  – Most of groups used pseudo-relevance feedback
  – Expansion using statistical thesaurus
  – Expansion using Web resources
  – Expansion using external document collection
  – Selective PRF (PIRCS, tlrrd)
    • Heuristic rule for selecting topics for which PRF is applied.
Techniques (6)

• Other techniques
  – Re-rank (HKPU, I2R)
  – Transliteration (tlrrd)
  – Pre-translation expansion (PIRCS)
  – Pronunciation-based indexing for Japanese text (NIIHI)
  – Identifying named entity (PIRCS)
  – Converting character codes with no translation (BRKLY)
Evaluation (1)

• Measures
  – Officially using standard output from trec_eval software
  – Mean average precision (MAP), R-precision, Recall-Precision graph, etc.
Evaluation (2)

- SLIR: C-C-D (Rigid) – top 8 groups

![Graph showing precision versus recall for different groups.](image-url)
Evaluation (3)

- SLIR: J-J-D (Rigid) – top 8 groups
Evaluation (4)

- SLIR: K-K-D (Rigid) – top 8 groups
Evaluation (5)

- SLIR: E-E-D (Rigid) – top 7 groups
Evaluation (6)

- BLIR – Comparison of MAP values between best SLIR and best BLIR runs (D-run, Rigid)

<table>
<thead>
<tr>
<th></th>
<th>C-C</th>
<th>J-J</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-C</td>
<td>.1568</td>
<td>32.5%</td>
</tr>
<tr>
<td>K-C</td>
<td>.0377</td>
<td>7.8%</td>
</tr>
<tr>
<td>E-C</td>
<td>.2682</td>
<td>55.6%</td>
</tr>
<tr>
<td>K-K</td>
<td>.5079</td>
<td></td>
</tr>
<tr>
<td>C-K</td>
<td>.3263</td>
<td>64.2%</td>
</tr>
<tr>
<td>J-K</td>
<td>.4511</td>
<td>88.8%</td>
</tr>
<tr>
<td>E-K</td>
<td>.4092</td>
<td>80.6%</td>
</tr>
</tbody>
</table>
Evaluation (7)

- MLIR – Best runs

<table>
<thead>
<tr>
<th>Run-type</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-CJKE</td>
<td>.2052</td>
</tr>
<tr>
<td>J-CJKE</td>
<td>.1890</td>
</tr>
<tr>
<td>K-CJKE</td>
<td>.1347</td>
</tr>
<tr>
<td>E-CJKE</td>
<td>.2695</td>
</tr>
</tbody>
</table>
Plan for the next workshop

• Encouraging more to try BLIR
  – In particular, J-C, C-J, K-C, C-K, K-E, E-K
• Enlarging doc collection
  – 4-years collection (1998-2001) will be used
• Collaborating with CLQA???
• Special subtask?????
  – Topics which it is hard to obtain good performance
  – Precision-oriented search
• Indian languages???????????????
Round Table Meeting for discussing NTCIR-6 CLIR Task

TODAY, after banquet
Start: 20:00
Place: Seminar room 1 on 12th floor
Up to the floor by elevator

Please give us suggestions!
Thank you for participating in NTCIR-5 CLIR Task!

And, I am sorry that some errors in the overview paper.