Thomson Legal and Regulatory at NTCIR-4:
Primarily monolingual experiments

Isabelle Moulinier
TLR, Research and Development
isabelle.moulinier@thomson.com

May 27, 2004
Overview

- System overview

- Monolingual experiments in Japanese, Chinese and Korean
  - Creating stopword lists
  - Handling compound terms in Korean
  - My first steps with Pseudo-Relevance Feedback

- Pivot-Language experiments

- Conclusion
System overview

• Research version of a production system
  – Asian languages not in production

• Handled documents in XML
  – Language identified at the collection level

• Indexing is word-based
  – Tokenization and stemming using LinguistX toolkit

• Retrieval model: a cousin of INQUERY
  – Uses structured queries
  – Uses tf-idf for concept scoring
Creating stopword lists

- Using collection information
  - with manual editing (Japanese and Chinese)
  - without manual editing (all languages)
    * 100 or 200 most frequent terms in the collection

- Using query log information
  - without manual editing (all languages)
    * terms appearing in more than 20% of the queries
Main results with stopword experiments

- Using stopword lists usually improves average precision significantly
  - Title only queries contain few stopwords

- Average Precision is not significantly different with various stopword lists
  - Typical stopwords appear in all lists
  - There is a query per query difference

- Key is balance between stopwords and concepts
  - Full queries contain strong concepts thanks to concept fields
Handling Korean compounds

- We use a stemmer to identify compound parts
  - Example: 흠런경쟁에서 stems to 흠런 #경쟁

- We think of compounds as equivalent to phrases

- Our approach
  - Index compounds and their parts
  - Use different proximity structures

<table>
<thead>
<tr>
<th></th>
<th>No Partial Credit</th>
<th>Partial Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strict (ND)</td>
<td>흠런 #경쟁</td>
<td>흠런 #경쟁 w 흠런 w₁ 경쟁 w₁</td>
</tr>
<tr>
<td>Loose</td>
<td>NPHR(흠런 경쟁)</td>
<td>NPHR(흠런 경쟁) w 흠런 w₁ 경쟁 w₁</td>
</tr>
</tbody>
</table>
Results with Korean compounds

- Partial credit is helpful
- Key is on good compound recognition
First steps with pseudo-relevance feedback

• Query expansion using PRF
  – Terms are selected using Rocchio’s formula and added to the original query
    \[ sw = \frac{\beta}{|R|} \sum_{d \in R} (ntf \times nidf) - \frac{\gamma}{|R|} \sum_{d \in \overline{R}} (ntf \times nidf) \]

• Parameter tuning using NTCIR-3 data
  – select 20 terms
  – select the 5 first documents as relevant
  – select the last 20 documents as irrelevant
  – \( \beta = \gamma = 1 \)
Some improvement over base runs but no statistical difference

Large query variations

<table>
<thead>
<tr>
<th></th>
<th>$\Delta &gt; 10% (\pm%)$</th>
<th>$\Delta &gt; 20% (\pm%)$</th>
<th>$\Delta &gt; 40% (\pm%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>tlrrd-tdnc-01</td>
<td>24 (14/10)</td>
<td>12 (7/5)</td>
<td>2 (0/2)</td>
</tr>
<tr>
<td>tlrrd-t-02</td>
<td>45 (18/27)</td>
<td>35 (13/22)</td>
<td>21 (6/15)</td>
</tr>
<tr>
<td>tlrrd-t-03</td>
<td>39 (19/20)</td>
<td>27 (14/13)</td>
<td>16 (7/9)</td>
</tr>
<tr>
<td>tlrrd-dn-04</td>
<td>27 (17/10)</td>
<td>18 (13/5)</td>
<td>4 (3/1)</td>
</tr>
</tbody>
</table>

Noticeable improvement in precision at 5 documents

Key is finding good documents in the original search
Pivot-language IR using Web resources

• Goal: Assess how well (poorly) pivot-language translation using Web resources would work

• Korean-English-Japanese
  – sentence translation using Babelfish

• Chinese-English-Japanese
  – word-based translation using Chinese-English dictionary and Babelfish from English to Japanese

• Outcome: Pivot-language IR uses Web resources works POORLY.
Conclusion

- Below average performance for our official runs
  - word-based indexing, especially tokenization
- Monolingual results
  - Stopwords are useful, independent of how they are created
  - Partial credit useful for searching compounds
  - Below expectation results for PRF
- Pivot-language results
  - The “dumb” approach does not work
  - Good news for more research