Best Practices in System-Oriented Aspects for Multilingual Information Access Applications

Martin Braschler
Zurich University of Applied Sciences
Winterthur, Switzerland
Motivation, problem area

- Growing interest in multilingual information access (MLIA)/cross-language information retrieval (CLIR):
  - access information in a language different in which querier has little or no skill
- Ever growing digital universe
- Internet becoming more multilingual
- Increasing academic output in the MLIA field (CLEF campaign)

- HOWEVER:
- Lack of commercial uptake of MLIA/CLIR technology!
Research Objectives

• Goal: to compile best practice guidelines in system-oriented MLIA

• "Digest" corpus of academic work in the field

• "Unify" the conclusions from a vast range of different experiments

• "Present" the results in the form of a best practice report and on a best practices portal
Research approach, Methodology

• We have used a number of sources to compile the recommendations
  – Overview papers from CLEF working notes/proceedings
  – Statistical analysis of the text of experiment descriptions
  – Feedback from workshop on operational, system-oriented MLIA
    (October 2008 at Winterthur, Switzerland)
  – Earlier analysis from 2003
  – Studies on searching web portals and company intranets
Research approach, Methodology

• Statistical analysis of the text of experiment descriptions

• Word/phrase frequency analysis
• Lists are scanned for terms that indicate the use of specific techniques and algorithms
• These terms are used as "seed queries" for exploration
• Boosts coverage of our analysis, low overhead
## Example

<table>
<thead>
<tr>
<th>Term</th>
<th>cf</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>average.precis</td>
<td>1541</td>
<td>294</td>
</tr>
<tr>
<td>cross.languag</td>
<td>1314</td>
<td>338</td>
</tr>
<tr>
<td>relev.document</td>
<td>1279</td>
<td>296</td>
</tr>
<tr>
<td>queri.expans</td>
<td>1267</td>
<td>238</td>
</tr>
<tr>
<td>question.answer</td>
<td>1255</td>
<td>175</td>
</tr>
<tr>
<td>document.collect</td>
<td>1144</td>
<td>171</td>
</tr>
</tbody>
</table>
Information Acquisition Cycle

Verbalization

I need a good cd burner

Formulation/Coding

cd burner

I could look at flash devices as an alternative!

Processing

DOC1: CD recorder
DOC2: Optical or flash?
DOC3: Burning your own CDs

Query

IR System

Result

...such as burning data to compact discs...

Optical discs have recently been overtaken by the flash devices as storage medium....

...while compact discs from a recorder deteriorate after prolonged storage..
Major Outcomes/Results

Flow is divided into

- Indexing
- Translation
- Matching
## Major Outcomes/Results

<table>
<thead>
<tr>
<th>Use weighted retrieval</th>
<th>copes with translation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Unicode/XML</td>
<td>covers different scripts</td>
</tr>
<tr>
<td>Use minimal stopword elimination</td>
<td>keep maximum information</td>
</tr>
<tr>
<td>Remove diacritics, special characters</td>
<td>tolerant towards inconsistent spelling</td>
</tr>
<tr>
<td>Use stemming</td>
<td>covers different word forms</td>
</tr>
<tr>
<td>Use decompounding</td>
<td>tolerant towards different phrasings</td>
</tr>
<tr>
<td>Use character n-grams</td>
<td>helps with languages with scarce language resources</td>
</tr>
</tbody>
</table>
# Major Outcomes/Results

<table>
<thead>
<tr>
<th>Major Outcomes/Results</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximize coverage of translation resources</td>
<td>reduces retrieval failure due to missing translations</td>
</tr>
<tr>
<td>Use document translation to solve merging problems</td>
<td>if combined results in multiple languages are needed</td>
</tr>
<tr>
<td>Combine different types of translation resources</td>
<td>minimizes mistranslations inherent to the individual resources</td>
</tr>
<tr>
<td>Use an interlingua</td>
<td>covers language pairs with no direct translation resources</td>
</tr>
<tr>
<td>Use high-performing weighting schemes</td>
<td>weighting schemes with robust performance over different types of text</td>
</tr>
<tr>
<td>Use pseudo-relevance feedback</td>
<td>boosts recall (coverage of results)</td>
</tr>
</tbody>
</table>
Major Outcomes/Results

• Blueprint

• Effective, well-tuned monolingual retrieval for as many languages as possible

• Combination of different sources of translation information

• Merging of multiple, well-tuned bilingual results
Major Outcomes/Results

• Lessons from studies of website portals and intranets

• Analyzed nearly 100 websites portals (DE/CH)

• Careful monitoring of the index coverage

• Good maintenance of metadata

• Good result presentation, follow existing leads
Conclusion and outlook

• Potential for increased use of MLIA/CLIR
• Still limited commercial uptake

• There are clear recommendations of what works in a large range of different settings (stemming from evaluation campaigns and studies)
• New initiatives aim to "formalize" some of these aspects (GridCLEF)
• Academic initiatives (CLEF) are interested in moving closer to industrial stakeholders
Thank you for your time!

Questions?