Zürcher Hochschule
für Angewandte Wissenschaften

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

Martin Braschler
Zurich University of Applied Sciences
Winterthur, Switzerland

Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

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

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

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

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Research approach, Methodology

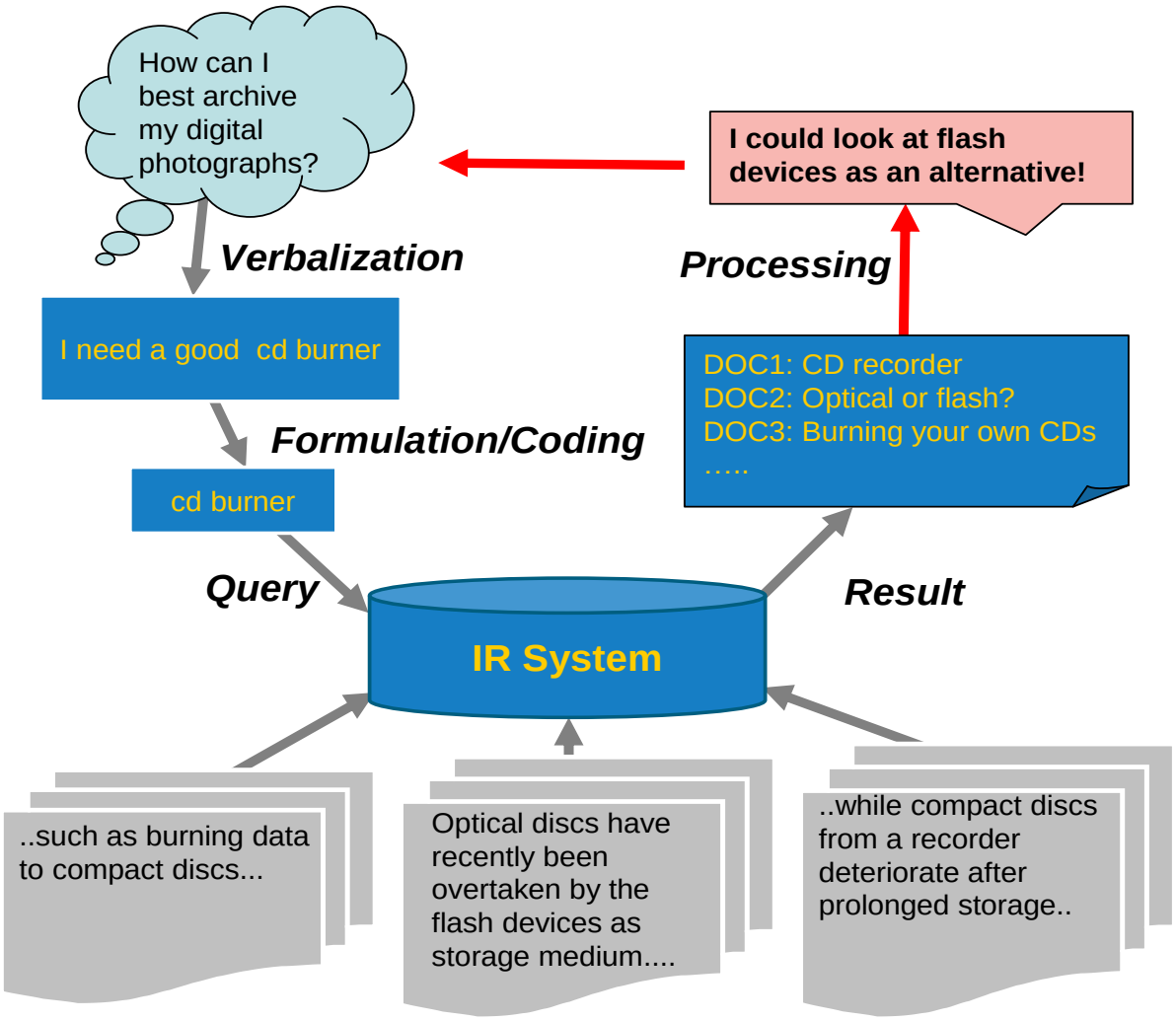
- Statistical analysis of the text of experiment descriptions

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

Example

Term	cf	df
average.precis	1541	294
cross.languag	1314	338
relev.document	1279	296
queri.expans	1267	238
question.answer	1255	175
document.collect	1144	171

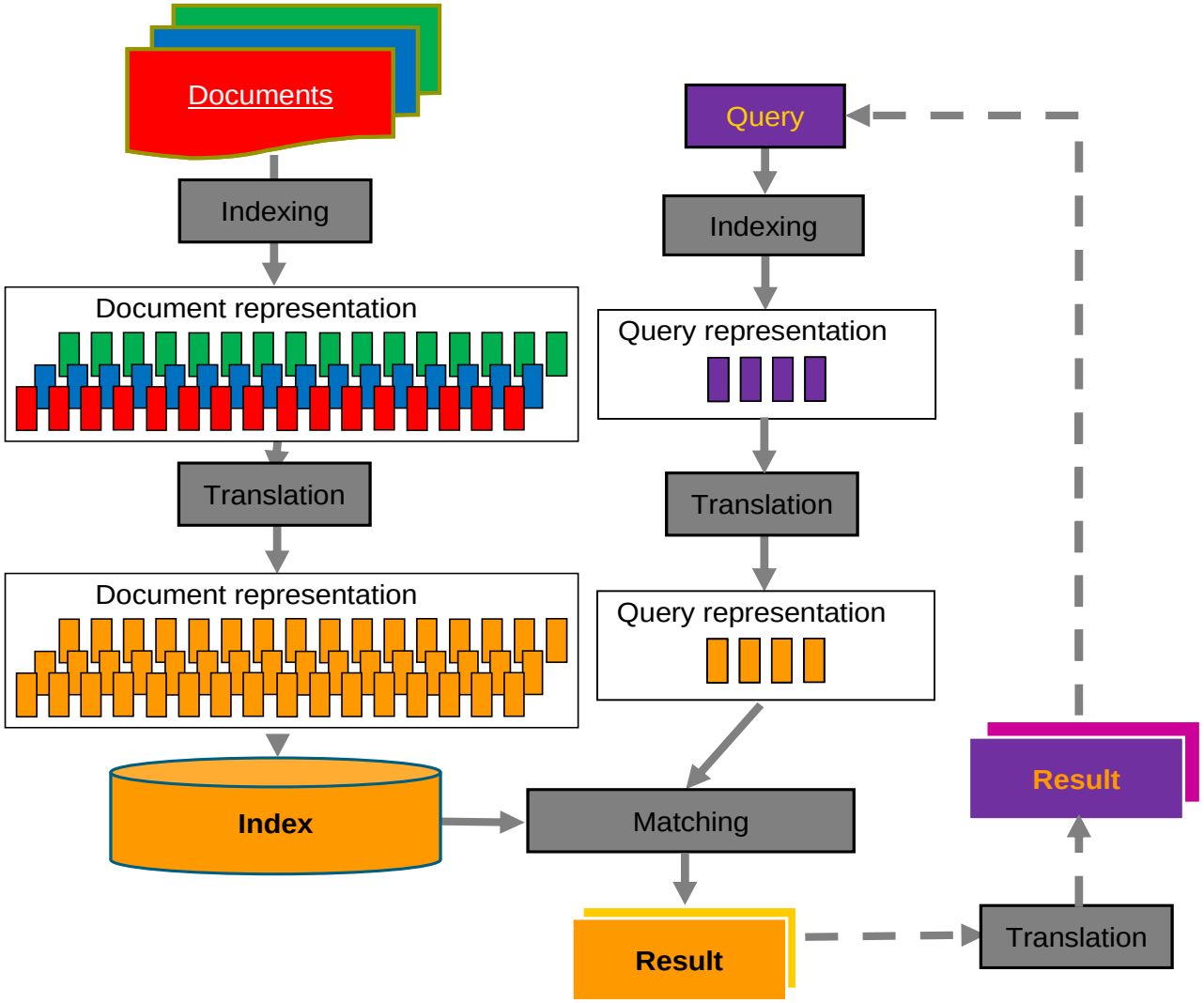
Information Acquisition Cycle



Major Outcomes/Results

Flow is divided into

- Indexing
- Translation
- Matching



Major Outcomes/Results

Use weighted retrieval	copes with translation error
Use Unicode/XML	covers different scripts
Use minimal stopword elimination	keep maximum information
Remove diacritics, special characters	tolerant towards inconsistent spelling
Use stemming	covers different word forms
Use decompounding	tolerant towards different phrasings
Use character n-grams	helps with languages with scarce language resources

Major Outcomes/Results

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

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

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

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

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Thank you for your time!

Questions?