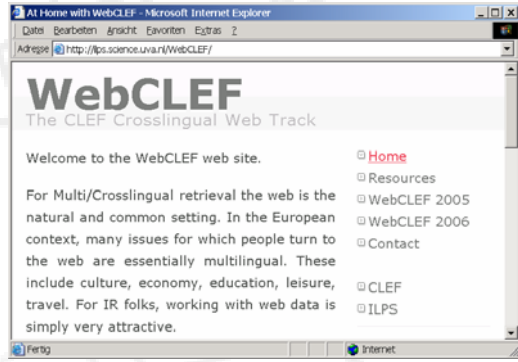
**Ben Heuwing, Thomas Mandl,
Robert Strötgen**

Information Science
Universität Hildesheim
mandl@uni-hildesheim.de

Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for CLEF 2006



Cross-Language
Evaluation Forum (CLEF)

Overview

- Challenges
- Indexing Approach
  - Fields Extracted
  - Content Indexing
  - Blind Relevance Feedback
- Results for WebCLEF 2005
- Results for WebCLEF 2006

Retrieval Approaches

- Multilingual stopword list
- One index for all languages
  - Words: no stemming

- -> no fusion problem, no language identification problem
- Search Engine: Lucene

HTML Titles

- Very effective at WebCLEF 2005
- Assumption: many titles might be of low quality
  - “no title”, „startpage“, etc. in many languages
- Goal: create a stop title list
- Finding: EuroGOV has good titles
  - valuable text
- Nevertheless, stopword list from last year was extended with the most frequent title words

Content Indexing

- Full Content
  - Used for searching
- Partial Content
  - Used to BRF
    (because of efficiency)

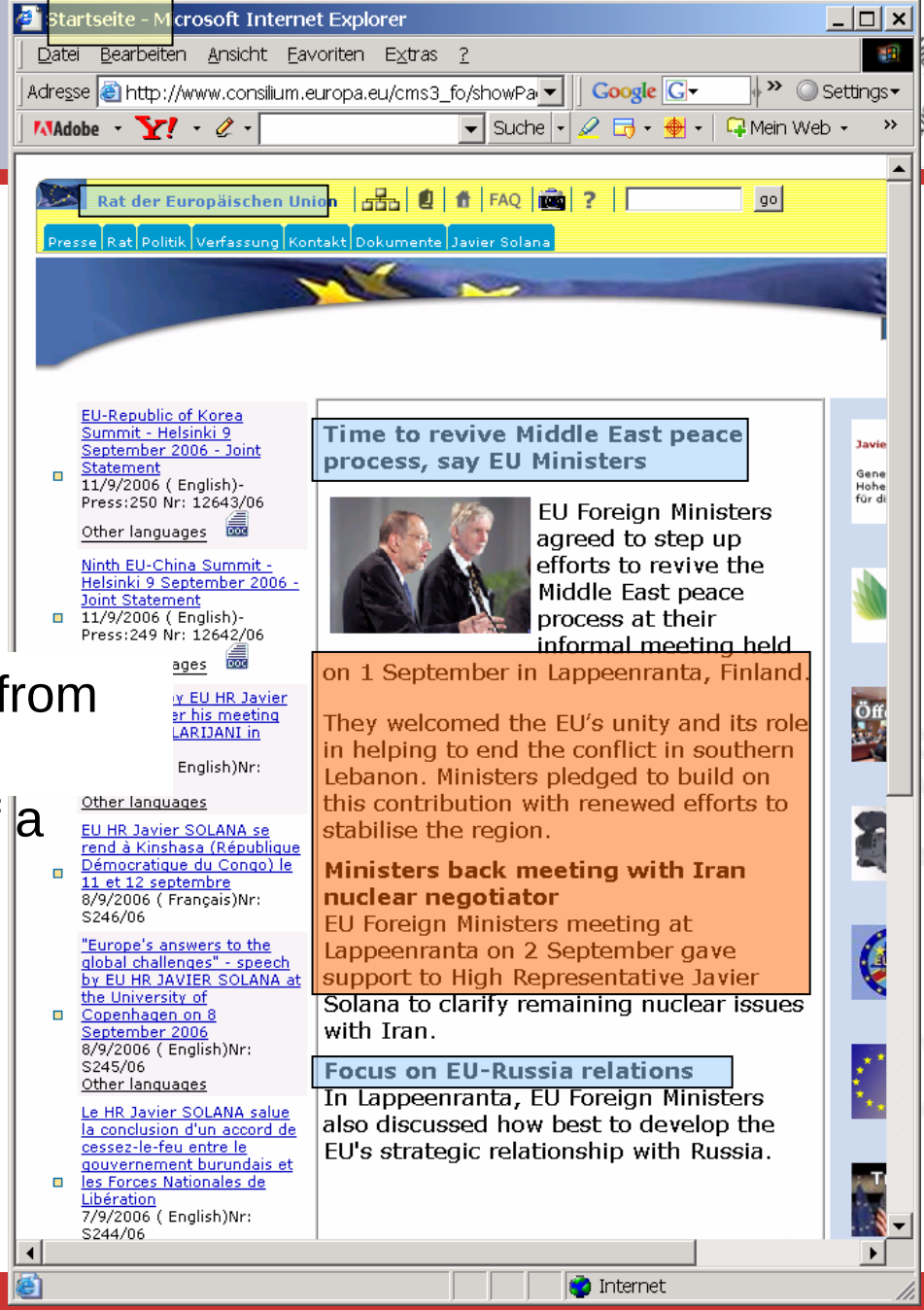
Partial Content

- Assumption
  - Partial content might be better
  - Eliminate menus, footers, headers
  - Several approaches try to identify the „important“ content
- Heuristic approach
  - Take from the „middle“

Approach

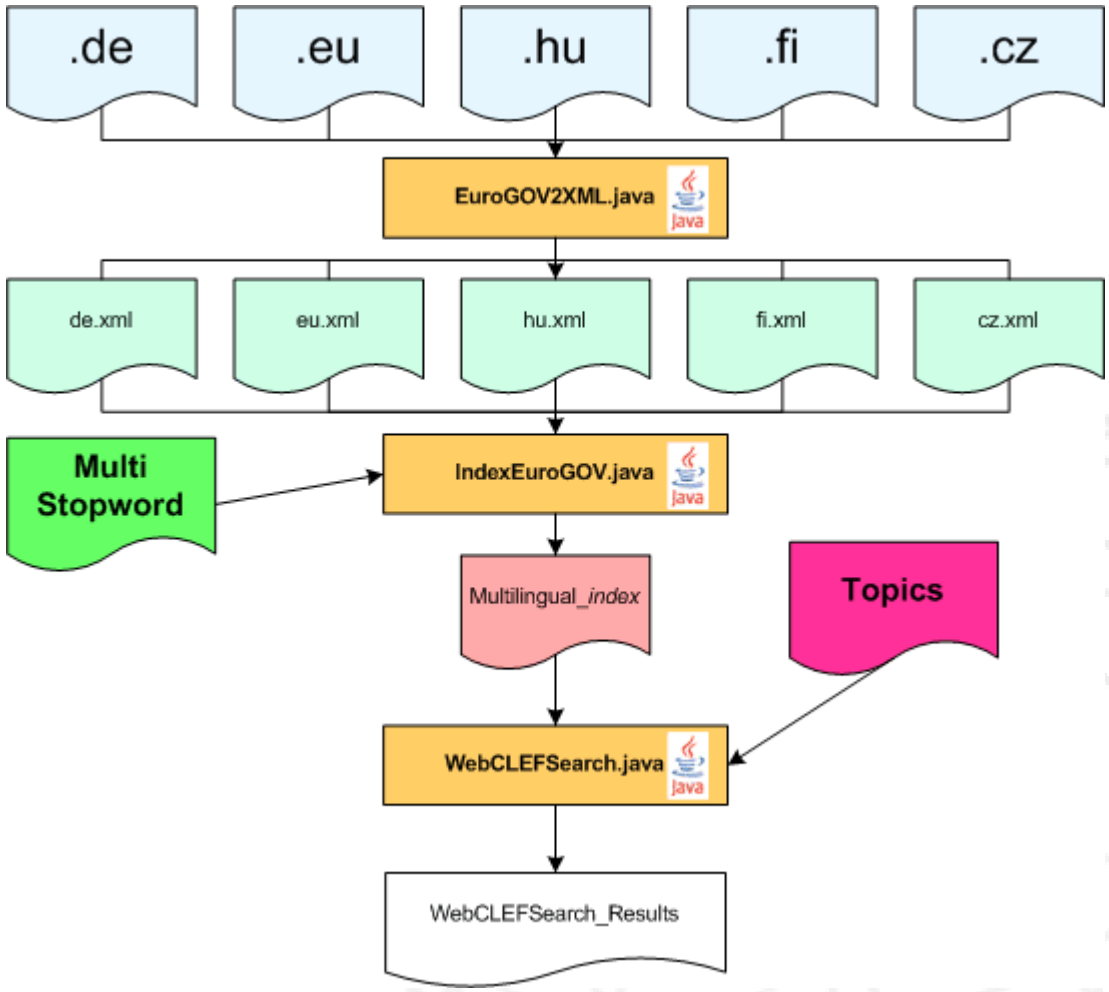
- Titles
  - + H1
- Content
  - Full & partial
- Emphazised text
  - H1 – H6, strong, em, bold, I, b

50 tokens from the middle“ of a page

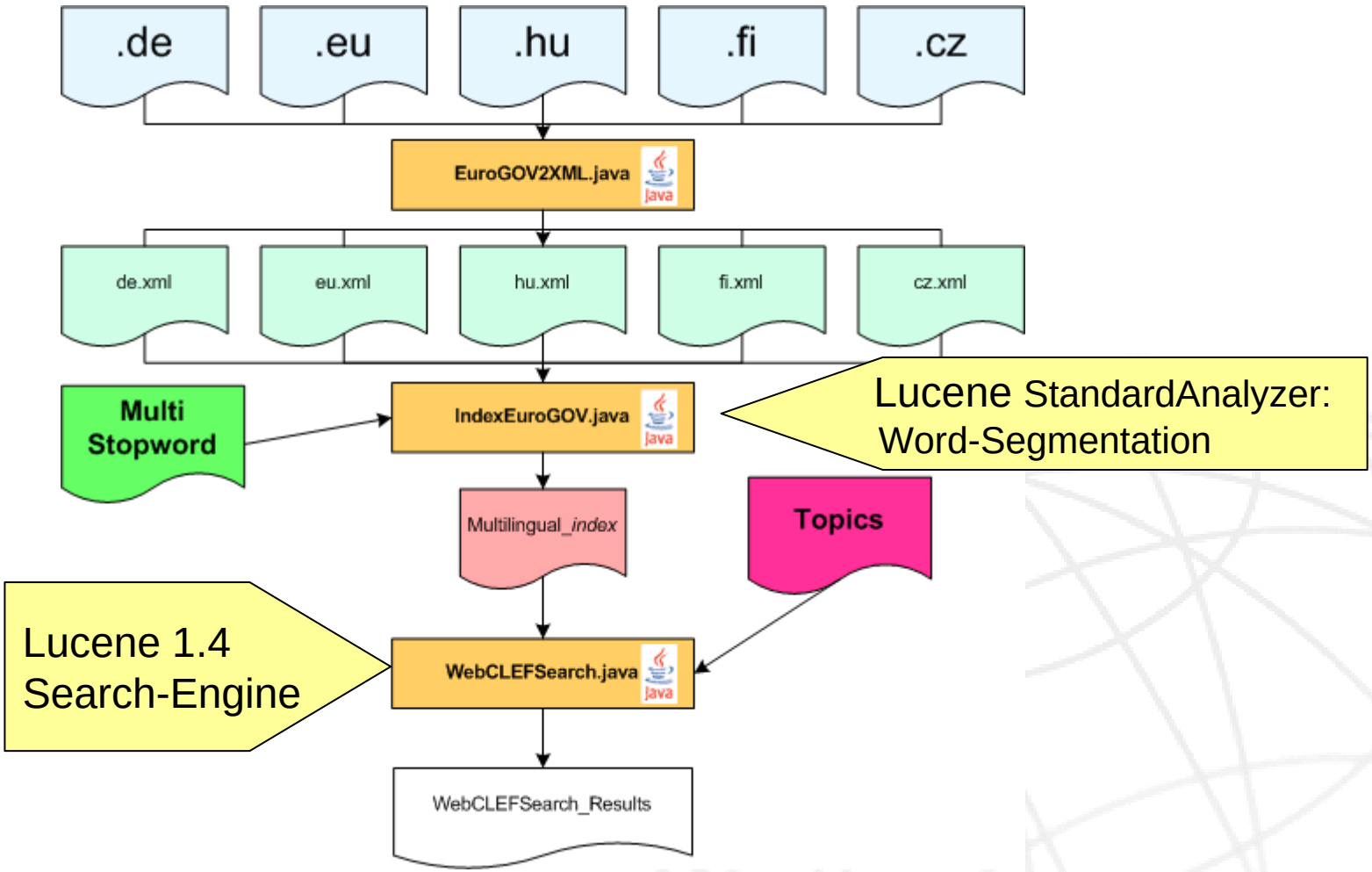


WebCLEFSearch Prozess

Lists from Neuchatel
+
Czech list assembled in Hildesheim
+
Frequent title words



WebCLEFSearch Prozess



Initial Findings

- Full content significantly better than partial content

- Title should to be weighted high

Multilingual Task

Multilingual Run	MRR
Best submission 2005	*0.137*
Best post experiment Hildesheim	*0.212*
Best (Hildesheim) run this year	*0.224*

Additional fields (H1), metadata and weighting helped

Parameters for Submitted Runs

Name of Run	Weights
UHiBase	*content^1 emphasised^0.1 title^20*
UHiTitle	*content^1 emphasised^1 title^20*
UHi1-5-10	*content^1 emphasised^5 title^10*
UHiBrf1	*content^1 emphasised^1 title^20 blind relevance feedback (weight of expanded query: 1)*
UHiBrf2	*blind relevance feedback (weight of expanded query: 0.5)*
UHiMu	*(multilingual) content^1 emphasised^1 title^20 - translation^10*

High title weights, brf weighted low

Results for WebCLEF 2005 Topics

	all topics		*manually generated topics*	
	MRR	Average success at 10	MRR	Average success at 10
UHiBase	0.0795	0.1377	0.3076	0.4451
UHiTitle	0.0724	0.1253	0.3061	0.4420
UHi1-5-10	0.0718	0.1233	0.3134	0.4577
UHiBrf1	0.0677	0.1104	0.3000	0.4295
UHiBrf2	0.0676	0.1124	0.2989	0.4295
UHiMulti	0.0489	0.0758	0.2553	0.3824

Results for Submitted Runs

Run	UHi Base	UHi Title	UHi 1-5-10	UHi Brf1	UHi Brf2
Mean reciprocal rank	0.282	0.281	0.281	0.273	0.277
Average success at 10	0.417	0.413	0.419	0.395	0.404

All runs quite similar

BRF – No positive Results (yet)

- No improvement using BRF
  - base run brings best results
  - but it does not hurt much
- Web Retrieval different?
  - BRF might be useless for page finding
  - there cannot be many similar pages in the first hits, if we look for only one page
- Maybe field specific BRF works better

Metadata

- Target domain was used
  - MRR higher
  - Success at 10 not much better
  - -> hits are higher in the result list

Conclusion

- A great corpus with many topics! Let‘s continue!
- Thanks U Amsterdam!
- Ample room for improvement still?