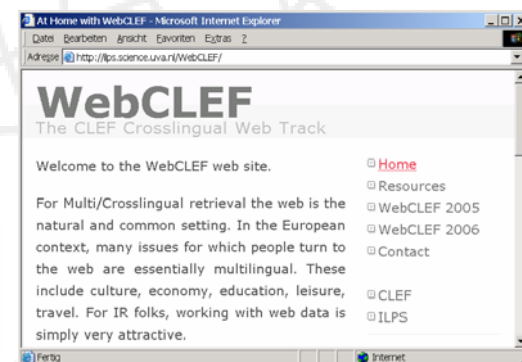


Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for CLEF 2006

Cross-Language
Evaluation Forum (CLEF)



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

- Multilingual stopword list
- One index for all languages
 - Words: no stemming
- -> no fusion problem, no language identification problem
- Search Engine: Lucene

- 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

- 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** 50 tokens from the
- **Emphazised text** "middle" of a page
 - H1 – H6, strong, em, bold, I, b

Startseite - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.consilium.europa.eu/cms3_fo/showPa... Google G Settings

Adobe Y! Suche Mein Web

Rat der Europäischen Union

Presse Rat Politik Verfassung Kontakt Dokumente Javier Solana

EU-Republic of Korea Summit - Helsinki 9 September 2006 - Joint Statement

11/9/2006 (English)- Press:250 Nr: 12643/06

Other languages

Ninth EU-China Summit - Helsinki 9 September 2006 - Joint Statement

11/9/2006 (English)- Press:249 Nr: 12642/06

ages

EU HR Javier Solana at his meeting LARIJANI in English)Nr:

Other languages

EU HR Javier SOLANA se rend à Kinshasa (République Démocratique du Congo) le 11 et 12 septembre

8/9/2006 (Français)Nr: S246/06

"Europe's answers to the global challenges" - speech by EU HR JAVIER SOLANA at the University of Copenhagen on 8 September 2006

8/9/2006 (English)Nr: S245/06

Other languages

Le HR Javier SOLANA salue la conclusion d'un accord de cessez-le-feu entre le gouvernement burundais et les Forces Nationales de Libération

7/9/2006 (English)Nr: S244/06

Time to revive Middle East peace process, say EU Ministers

EU Foreign Ministers agreed to step up efforts to revive the Middle East peace process at their informal meeting held on 1 September in Lappeenranta, Finland.

They welcomed the EU's unity and its role in helping to end the conflict in southern Lebanon. Ministers pledged to build on this contribution with renewed efforts to stabilise the region.

Ministers back meeting with Iran nuclear negotiator

EU Foreign Ministers meeting at Lappeenranta on 2 September gave support to High Representative Javier Solana to clarify remaining nuclear issues with Iran.

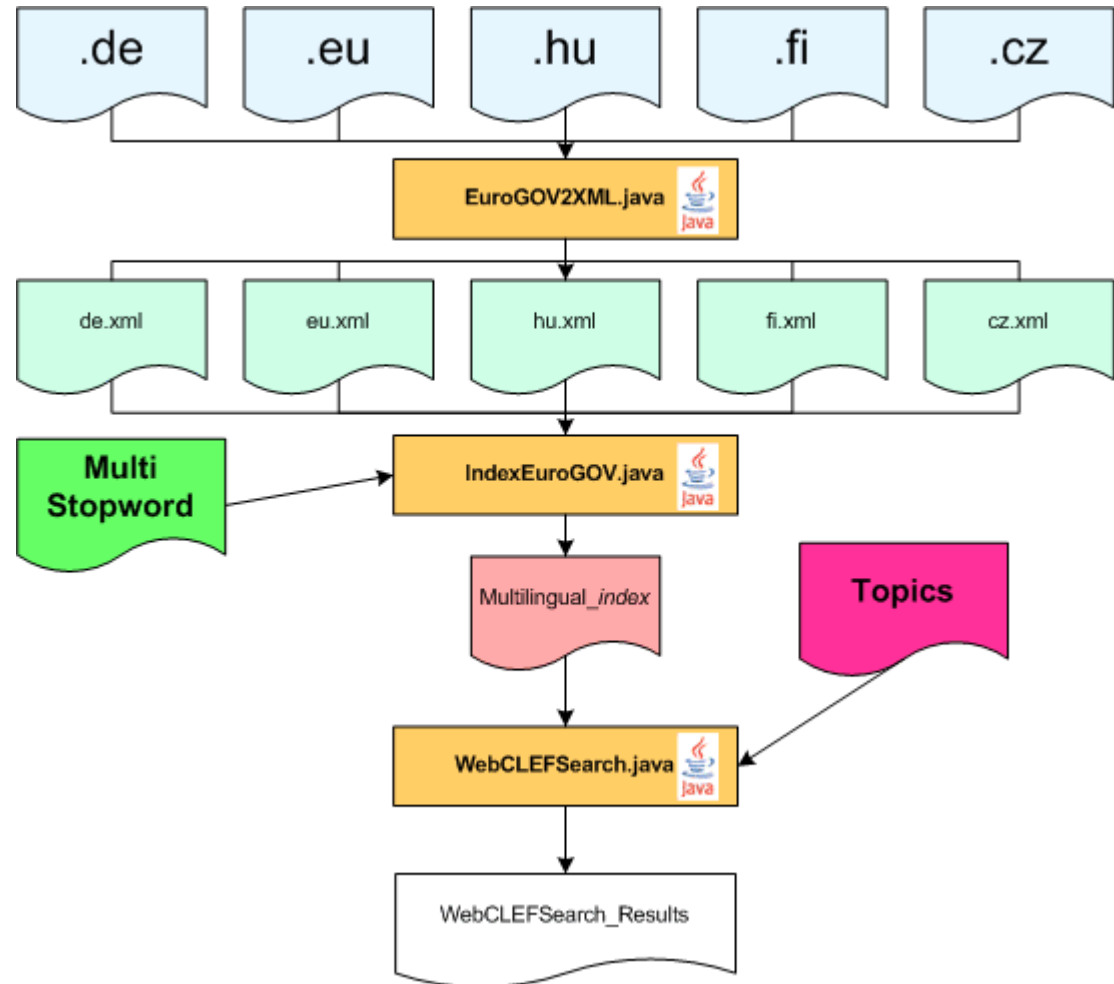
Focus on EU-Russia relations

In Lappeenranta, EU Foreign Ministers also discussed how best to develop the EU's strategic relationship with Russia.

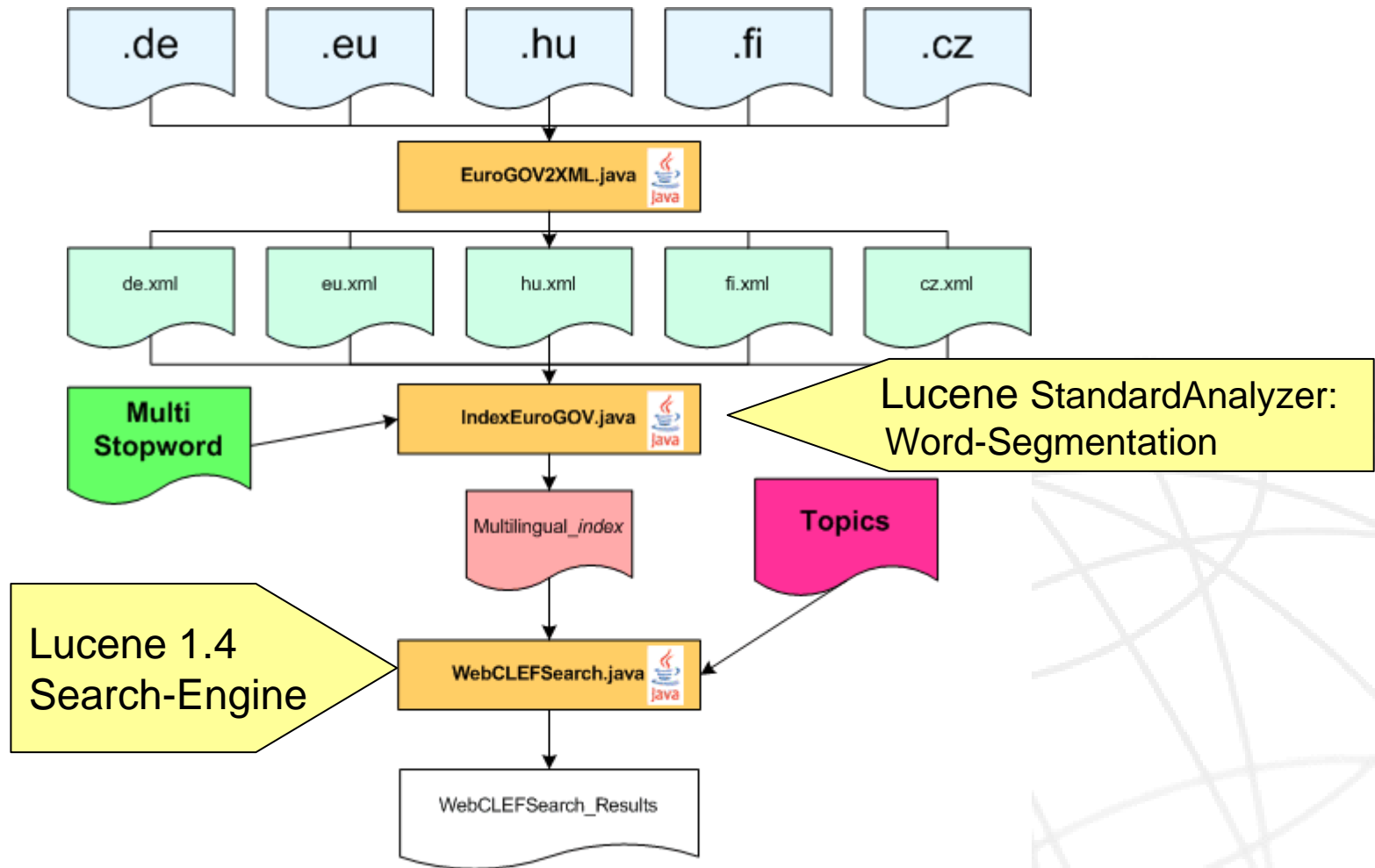
Internet

WebCLEFSearch Prozess

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



WebCLEFSearch Prozess



- Full content significantly better than partial content
- Title should to be weighted high



Multilingual Run	MRR
<i>Best submission 2005</i>	<i>0.137</i>
<i>Best post experiment Hildesheim</i>	<i>0.212</i>
<i>Best (Hildesheim) run this year</i>	<i>0.224</i>

Additional fields (H1), metadata
and weighting helped

Parameters for Submitted Runs

Name of Run	Weights
<i>UHiBase</i>	<i>content¹ emphasised^{0.1} title²⁰</i>
<i>UHiTitle</i>	<i>content¹ emphasised¹ title²⁰</i>
<i>UHi1-5-10</i>	<i>content¹ emphasised⁵ title¹⁰</i>
<i>UHiBrf1</i>	<i>content¹ emphasised¹ title²⁰ blind relevance feedback (weight of expanded query: 1)</i>
<i>UHiBrf2</i>	<i>blind relevance feedback (weight of expanded query: 0.5)</i>
<i>UHiMu</i>	<i>(multilingual) content¹ emphasised¹ title²⁰ - translation¹⁰</i>

High title weights, brf weighted low

Results for WebCLEF 2005 Topics

	<i>all topics</i>		<i>manually generated topics</i>	
	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

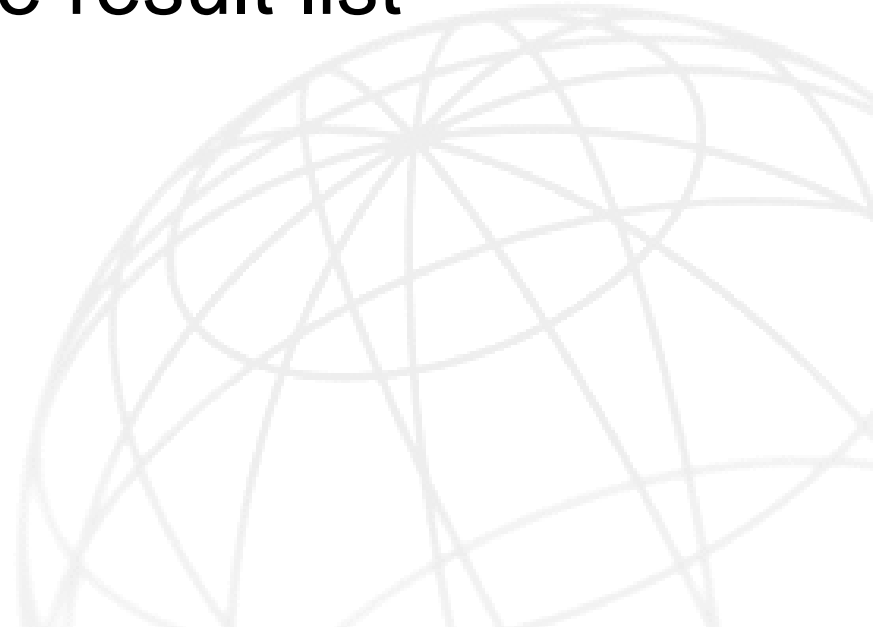
Results for Submitted Runs

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

- 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

- 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?