

Overview of WebCLEF 2006

Krisztian Balog, Leif Azzopardi, Jaap Kamps, Maarten de Rijke





Overview

- 🔺 A look back
- 🔺 Aims for 2006
- 🔺 Topic generation
- 🔺 Tasks
- 🔺 Submissions
- 🔺 Results
- 🔺 Answers
- 🔺 Questions



A reminder: WebCLEF 2005



Tasks

-  **Mixed monolingual:** stream of known-item topics in a variety of European languages; language of topic is language of target page
-  **Multilingual:** stream of English known-item topics; desired pages may cover any language

Document collection

-  **EuroGOV:** Crawl of European governmental sites; 3M pages

Outcomes

-  **11 teams participated; use of metadata makes big difference; CLIR effective for mixed monolingual task**
-  **549 (!) known-item topics, but issues with collection, issues with task**

Main aims for 2006

❖ “In between” year

❖ Participants: don't change the collection

❖ Organizers: why bother manually creating, say, 500 more known-item topics? We already have 549!

❖ Participants: little interest in helping with ad hoc assessments

❖ Compromise

❖ Old vs new topics

❖ Manual vs automatically generated topics



Topic generation 2006

- ❖ New manual topics created using same interface and settings as last year
- ❖ Automatic
 - ❖ Modeling before. S that would
 - ❖ Repeat k times:
 - ❖ Select a term t from doc model
 - ❖ Add t to query q
 - ❖ Unigrams, bigrams, noise model
- ❖ More detail



Topics 2006

- ❖ 1940 in total
- ❖ 125 new manual topics in 5 languages
 - ❖ Dutch, English, German, Hungarian, and Spanish
 - ❖ Created by UCM (Madrid) and the track organizers
- ❖ 195 topics randomly sampled from 2005 test set
- ❖ 1620 automatic topics in 27 languages
 - ❖ 810 unigram-based, 810 bigram-based



Tasks 2006

❖ Stream of known-item topics in a range of languages

❖ *Mixed-monolingual* task

❖ Stream of monolingual known-item topics; language of topic is language of target document

❖ Manual topics contain an English translation, which allowed for a *Multilingual* task

❖ Stream of English topics; the desired pages may cover any language or domain in the collection



Submissions 2006

- ❖ For each task, submit up to 5 runs
- ❖ For each topic at least 1 and at most 50 results should be returned
 - ❖ Provide a list of metadata used
- ❖ Runs submitted by 8 teams
 - ❖ BUAP, Depok, Hildesheim, Hummingbird, Reina, RFIA, UCM, UvA
 - ❖ Mixed monolingual: 35
 - ❖ Multilingual: 1



Results 2006

- ❖ Embarrassment ...
- ❖ Performance on automatic topics frequently very poor
 - ❖ *All* vs *new* topics (1,940 vs 1,120—leaving out “0-scoring topics”)
- ❖ Best run for each team achieved using metadata fields
 - ❖ Knowledge of page’s primary domain moderately effective
- ❖ Old manual vs new manual
 - ❖ Returning participants better on old manual topics (compared to 2005)
 - ❖ All participants better on new manual than old manual
- ❖ Automatic topics harder than manual ones
 - ❖ Max/avg MRR automatic vs manual: 0.3134/0.1015 vs 0.5068/0.2780



Comparing rankings of runs

		all	auto	auto-uni	auto-bi	manual	manual-new	manual-old
all	τ		0.8182	0.7726	0.8125	0.5935	0.6292	0.5707
	p		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
auto	τ			0.9412	0.9688	0.4108	0.4575	0.3945
	p			0.0000	0.0000	0.0006	0.0001	0.0010
auto-uni	τ				0.9097	0.3717	0.4183	0.3619
	p				0.0000	0.0019	0.0005	0.0025
auto-bi	τ					0.4029	0.4762	0.3800
	p					0.0008	0.0000	0.0016
manual	τ						0.9123	0.9642
	p						0.0000	0.0000
manual-new	τ							0.8769
	p							0.0000



Answers 2006

- ❖ **Current CLIR systems quite effective**
 - ❖ Impressive as 27 primary domains were involved
 - ❖ Manually created topics result in higher performance than automatically created ones
 - ❖ Progress on the old topics, new manual topics seem to confirm this
- ❖ **Mixed conclusion on usage of automatic topics**
 - ❖ Substantial differences between automatic topics
 - ❖ But scores on automatic topics give a solid indication of performance



Questions 2006

Lack of interest?

 Task, document collection, ...

How to move forward with WebCLEF?

 Task, collection, assessments, ...

 Wanted: reasonable abstraction of a realistic task, using web data, where multi-linguality fits naturally

 More at the WebCLEF // session, Friday, 9.00–10.30

