

CLEF Web Track: A Proposal

Jaap Kamps
Maarten de Rijke
Börkur Sigurbjörnsson

Informatics Institute
University of Amsterdam



Motivation

- ▶ Multi/Crosslingual retrieval
 - web is the natural and common setting
- ▶ In the European context, many issues for which people turn to the web are essentially multilingual
 - culture, economy, education, leisure, travel
- ▶ For IR folks, working with web is simply data attractive


Multilingual Aspects

- ▶ Use cases
 - even if people are not able to generate queries in a language, they may understand web pages in that language
 - ▶ Dutch person looking for name of Danish Foreign minister
 - ▶ de Deense minister van buitenlandse zaken
 - ▶ <http://www.um.dk/da/menu/OmOs/Udenrigsministeren>

Multilingual Aspects

- ▶ Use cases
 - many European students learn several foreign languages
 - ▶ Danish student learning German and English, looking for info on German soccer, not too much trouble reading Norwegian, Swedish
 - ▶ tysk mester i fodbold
 - ▶ tysk mästare i fotboll
 - ▶ deutsche fussball meister

Task Description

- ▶ For the first year, multi-lingual navigation task
 - home page finding
 - named page finding
 - ▶ Possible tasks
 - $X \rightarrow X$
 - $EN \rightarrow X$
 - $X \rightarrow \text{Everything}$
 - more complex mixtures
 - ▶ What info to reveal?
 - language of topic, of docs, of relevant docs,...
- 

Document Collection

Country	Domain	Fetches pages
EU-land	eu.int	65,879
Austria	at	13,258
Belgium	be	5,757
Cyprus	cy	487
Czech republic	cz	32,651
Denmark	dk	25,139
Estonia	ee	26,057
Finland	fi	392,923
France	fr	42,262
Germany	de	22,311
Greece	gr	8,160
Hungary	hu	12,584
Ireland	ie	21,787
Italy	it	25,170
Latvia	lv	19,510
Lithuania	lt	9,866
Luxembourg	lu	1,294
Malta	mt	2,966
Poland	pl	13,237
Portugal	pt	2,756
Slovakia	sk	12,087
Slovenia	si	5,577
Spain	es	9,677
Sweden	se	313,958
The Netherlands	nl	24,639
The United Kingdom	uk	27,814

Document Collection

- ▶ Aim
 - 1 to 2M pages
 - at least 50K docs per language
 - have more languages than we will actually use
 - at least 10 languages/countries
 - HTML, TXT, PDF
- ▶ Build on insights gained down under while building W10G and .GOV

Topic Creation

- ▶ Follow the CLEF or QA@CLEF model:
 - participants generate topics
 - topics are translated into the source languages
- ▶ Alternative: take queries from log file and translate these into multiple languages
- ▶ 150 NPs and 150 HPs

Assessment & Evaluation

- ▶ Participants return 50 results per topic
- ▶ Topic creators assess their own topics
- ▶ Open issue: can we exclude that a relevant page also occurs in an unexpected language?
 - probably not
 - but may not be a serious problem
- ▶ Measures: MRR, S@10

What the Organizers Provide

- ▶ Crawl, cleaned-up, chopped up in compressed CD images available for download
- ▶ Web interface for topic development
- ▶ Topic development guidelines
- ▶ Web interface for assessment
- ▶ Distribution of topics
- ▶ Collecting of participants' results, etc.

▶ Schedule

- September: web site, mailing list
- December: freeze document collection, task details
- January: release document collection
- February: complete topic development
- after that: CLEF schedule

▶ Contact:

- ir@science.uva.nl

CLEF Multilingual Web Track