Buy 30, Get 547

Maarten de Rijke

Informatics Institute, University of Amsterdam

Overview of WebCLEF 2005

Maarten de Rijke

Informatics Institute, University of Amsterdam

Why?

- The web is essentially multilingual
  - Many languages
  - Two-third of users has primary language other than English
- Cross-language information retrieval
  - Original focus on monolingual user accessing documents in multiple languages
  - Recent change to polylingual users accessing documents in languages they understand
- Gey, Kando, and Peters 2001 ("Roadmap on crosslingual IR")
  - Why is there no multilingual web corpus/track/…?

Domain	Pages				Size
	Total	Duplicated	Duplicates	Unique	(compressed)
.at	10,065	457	950	9,115	24M
.be	69,011	819	2,066	66,945	115M
.cy	1,972	52	52	1,920	7.9M
.cz	324,496	10,808	25,915	298,581	519M
.de	444,794	1,682	4,658	440,136	1.1G
.dk	2,144	497	519	1,625	5.4M
.ee	16,768	486	3,960	12,808	44M
.es	35,168	3,372	9,297	25,871	298M
.eu.int	374,484	32,838	58,415	316,069	1.9G
.fi	661,559	5,815	85,289	576,270	1.3G
.fr	156,450	11,144	21,894	134,556	545M
.gr	303	10	15	288	416K
.hu	330,822	361	1,082	329,740	1.5G
.ie	12,754	1,431	1,982	10,772	32M
.it	89,836	10,056	17,011	72,825	324M
.lt	10,765	751	1,131	9,634	8.8M
.lu	8,521	52	837	7,684	33M
.lv	317,404	10,357	25,711	291,693	675M
.mt	13,991	1,300	1,372	12,619	57M
.nl	149,949	6,097	18,911	131,038	434M
.pl	66,885	3,746	4,889	61,996	330M
.pt	147,445	2,454	8,744	138,701	753M
.ru	104,659	10,676	20,049	84,610	479M
.se	102,457	2,506	15,068	87,389	155M
.si	12,434	73	224	12,210	27M
.sk	58,020	3,288	3,764	54,256	128M
.uk	66,345	1,688	2,987	63,358	331M
Total	3,589,501	122,816	336,792	3,252,709	11G

2

Tasks

- Mixed monolingual
  - Simulate a user searching for a known item in a European language
  - "Stream" of monolingual topics
- Multilingual
  - Simulate a user looking for a certain known item page in a particular European language
  - User uses English to formulate her query
- Bilingual
  - English to Spanish
- Evaluation metric: MRR (mean reciprocal rank)

```
<topic>
  <num>WC0005</num>
  <title>Minister van buitenlandse zaken</title>
  <metadata>
    <topicprofile>
      <language language="NL"/>
      <translation language="EN">dutch minister of foreign
        affairs</translation>
    </topicprofile>
    <targetprofile>
      <language language="NL"/>
      <domain domain="nl"/>
    </targetprofile>
    <userprofile>
      <native language="IS"/>
      <active language="EN"/>
      <active language="DA"/>
      <active language="NL"/>
      <passive language="NO"/>
      <passive language="SV"/>
      <passive language="DE"/>
      <passive_other>Faroese</passive_other>
      <countryofbirth country="IS"/>
      <countryofresidence country="NL"/>
    </userprofile>
  </metadata>
</topic>
```

Group name	# topics
BUAP	39
Hummingbird	30
U.Amsterdam (ILPS)	162
Melange	30
Metacarta Inc.	3
Daedalus S.A.	30
Linguateca	30
U.Alicante	30
U. Glasgow	30
U. Hildesheim	30
U. Indonesia	36
UNED (NLP group)	30
U. Melbourne (NICTA)	47
U. Salamanca (Reina)	30
U. Lisboa (XLDB)	30
Total	**547**

Runs and results

- For each task, teams could submit up to 5 runs
  - Baseline (using no metadata) required
  - Additional runs could use some or all of the metadata
- 11 teams submitted runs
  - 34 mixed monolingual runs from 9 teams
  - 19 multilingual runs from 4 teams
  - 8 bilingual runs from 3 teams
- Strategies: web-specific, linguistic, cross-lingual
  - Mixed monolingual: do-able
  - Multilingual and bilingual: hard

Mixed monolingual	MRR
U. Glasgow	0.5135
Hummingbird	0.4780
U. Amsterdam (ILPS)	0.3497
Multilingual	
U. Hildesheim	0.1479
Miracle	0.0902
U. Amsterdam (ILPS)	0.0175
Bilingual	
UNED	0.0930
BUAP	0.0844
U. Alicante	0.0385

3 Presentations

Glasgow

10:10–10:25

10:25–10:40

UNED

10:40–10:55

Hildesheim

Wrap-up

- Thank you
  - Carol Peters, Jaap Kamps, Börkur Sigurbjörnsson, Ian Soboroff
  - University of Glasgow, UNED
  - Participants and topic developers
- Next year
  - Same tasks as this year?
  - Pilot task on blog data?
  - Join the WebCLEF breakout session, 14:30–16.00 today, **HS7**
- Remember
  - If you want to do crosslingual/multilingual retrieval, do it on web data (what else?), and do it at WebCLEF!