Overview of WebCLEF 2005

Börkur Sigurbjörnsson[1] Jaap Kamps[1,2] Maarten de Rijke[1]
[1] Informatics Institute, University of Amsterdam
[2] Archives and Information Studies, University of Amsterdam
{borkur,kamps,mdr}@science.uva.nl

Abstract

We describe WebCLEF, the multilingual web track, that was introduced at CLEF 2005. We provide details of the tasks, the topics, and the results of WebCLEF participants. The mixed monolingual task proved an interesting addition to the range of tasks in cross-language information retrieval. Although it may be too early to talk about a solved problem, effective web retrieval techniques seem to carry over to the multilingual setting. The multilingual task, in contrast, is still very far from being a solved problem. Remarkably, using non-translated English queries proved more successful than to use translations of the English queries.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Web retrieval, Known-item retrieval, Multilingual retrieval

1 Introduction

The world wide web is a natural setting for cross-lingual information retrieval; web content is essentially multilingual, and web searchers are often polyglots. Even though English has emerged as the lingua franca of the web, planning for a business trip or holiday usually involves digesting pages in a foreign language. The same holds for searching information about European culture, education, sports, economy, or politics. To evaluate systems that address multilingual information needs on the web, a new multilingual web track, called WebCLEF, has been set up as part of CLEF 2005.

Three tasks were organized within this year's WebCLEF track: mixed monolingual, multilingual, and bilingual English to Spanish, with 242 homepage and 305 named page finding queries for the first two tasks, and 67 homepage and 67 named page finding tasks for the third task. All topics, and the accompanying assessments, were created by the participants in the WebCLEF track. In total, 11 teams submitted 61 runs for the three tasks.

The main findings of the WebCLEF track in 2005 are the following. The mixed monolingual task proved an interesting addition to the range of tasks in cross-language information retrieval. Although it may be too early to talk about a solved problem, effective web retrieval techniques

```
<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>
```

Figure 1: Example of a WebCLEF 2005 topic.

seem to carry over to the multilingual setting. The multilingual task, in contrast, is still very far from being a solved problem. Remarkably, using non-translated English queries proved more successful than to use translations of the English queries.

The remainder of the paper is organized as follows. In Section 2 we describe the WebCLEF 2005 track in more detail. Section 3 is devoted to a description of the runs submitted by the participants, while the results are presented in Section 4. We conclude in Section 5.

2 The Retrieval Tasks

2.1 Collection

For the purposes of the track a new corpus, called EuroGOV, was developed. EuroGOV is a crawl of European government-related sites, where collection building is less restricted by intellectual property rights. It is a multilingual web corpus, which contains over 3.5 million pages from 27 primary domains, covering over twenty languages. There is no single language that dominates the corpus, and its linguistic diversity provides a natural setting for multilingual web search. We refer to [2] for further details on EuroGOV.

2.2 Topics

Topic development was in the hands of the participating groups. Each group was expected to create at least 30 monolingual known-item topics, 15 homepages and 15 named page topics. Homepage topics are names of a site that the user wants to reach, and named page topics concern non-homepages that the user wants to reach. The track organizers assigned languages to groups based on their location and the language expertise available within the group. For each topic, topic creators were instructed to detect identical or similar pages in the collection, both in the

Table 1: Summary of participating teams, the number of topics they developed and the number of runs they submitted.

Group id	Group name	Subm. topics	Runs Mixed-Mono	Multilingual	BiEnEs
buap	BUAP (C.S. Faculty)	39			5
hummingbird	Hummingbird	30	5		
ilps	U. Amsterdam (ILPS)	162	1	4	
melange	Melange (U. Amsterdam)	30	5	5	
miracle	DAEDALUS S.A.	30	5	5	
ualicante	U. Alicante	30	2		1
uglasgow	U. Glasgow (IR group)	30	5		
uhildesheim	U. Hildesheim	30	3	5	
uindonesia	U. Indonesia	36	3		
uned	NLP Group - UNED	30			2
unimelb	U. Melbourne (NICTA i2d2)	47			
usal	U. Salamanca (REINA)	30	5		
sintef	Linguateca	30			
xldb	U. Lisboa (XLDB Group)	30			
metacarta	MetaCarta Inc	3			
Total		547	34	19	8

language of the target page and in other languages. Many European governmental sites provide translations of (some of) their web pages in a small number of languages, e.g., in additional official languages (if applicable), in languages of some neighboring countries, and/or in English. In addition, participants provided English translations of their topics.

The topic authors were also asked to fill out a form where they provided various types of metadata, including their language knowledge, birth place and residence. This information was used to augment the topics with additional metadata. Figure 1 provides an example of the topic format used at WebCLEF 2005. The track organizers reviewed the topics, suggested improvements, and finally selected the final set of topics.

As few participants had facilities to search the EuroGOV collection during the topic development phase, the organizers provided a Lucene-based search engine for the collection, and the University of Glasgow provided access to the collection through Terrier, for which we are very grateful. Both search engines were at a proof-of-concept level only and were not specially adapted for the task.

Table 1, column 3, shows a summary of the number of topics submitted by each participating team. The WebCLEF 2005 topic set contained 547 topics, 242 homepage topics and 305 named page topics. The target pages were in 11 different languages: Spanish (ES), English (EN), Dutch (NL), Portuguese (PT), German (DE), Hungarian (HU), Danish (DA), Russian (RU), Greek (EL), Icelandic (IS), and French (FR). Since topic development depended on language knowledge within participating groups the distribution between languages in the test set varies considerably. Table 2 provides more detailed statistics of the WebCLEF 2005 topic set.

During topic development, topic authors were asked to try to identify duplicates and translations of the target page. Table 2 shows the number of duplicates/translations available. We list both the number of topics having a duplicate/translation and also the total count of duplicates/translations. The category *Readable trans.* refers to the number of translations whose language matches the language knowledge identified by the user. The number of translations naturally varies from one domain to another. As an example, 78 topics target pages were located in the `eu.int` domain (14% of the topics), and those pages have 232 translations (60% of identified translations). The identification of translations is a difficult and labor intensive process. Due to a lack of resources we have not been able to verify the completeness of duplicate/translation identification. This must be taken into account when interpreting results using the duplicate/translation

Table 2: Number of topics per language for both homepages (HP) and named pages (NP). The languages are sorted by the number of available topics. The bottom part of the table shows how many duplicates/translations were identified. We list both the number of topics having a duplicate/translation and also the total count of duplicates/translations.

	Total	ES	EN	NL	PT	DE	HU	DA	RU	EL	IS	FR
Total	547	134	121	59	59	57	35	30	30	16	5	1
HP	242	67	50	25	29	23	16	11	15	5	1	–
NP	305	67	71	34	30	34	19	19	15	11	4	1
Duplicates (topics)	191	37	47	21	15	38	11	12	8	1	1	–
Duplicates (total)	473	82	109	40	95	90	18	26	11	1	1	–
Translations (topics)	114	25	24	9	4	13	6	15	6	7	5	–
Translations (total)	387	100	47	18	7	39	17	101	11	19	28	–
Readable trans. (topics)	72	17	6	9	2	10	6	9	5	7	1	–
Readable trans. (total)	143	29	8	16	3	26	6	30	6	13	6	–

information.

2.3 Tasks

Due to limited resources for evaluation all tasks at WebCLEF 2005 were restricted to known-item searches. The following tasks were organized for WebCLEF 2005.

- *Mixed-Monolingual* The mixed-monolingual task is meant to simulate a user searching for a known-item page in an European language. The mixed-monolingual task used the title field of the topics to create a set of monolingual known-item topics.
- *Multilingual* The multilingual task is meant to simulate a user looking for a certain known-item page in a particular European language. The user, however, uses English to formulate her query. The multilingual task used the English translations of the original topic statements.
- *Bilingual English to Spanish* For this task a special topic set was used. It contained a reviewed translation of the Spanish topics. The reviewed and revised translations were provided by the NLP group at UNED, for which we are very grateful.

2.4 Submission

For each of the tasks, teams were allowed to submit up to 5 runs. Each run could contain 50 results for each topic.

2.5 Evaluation

Since each NP and HP topic is developed with a URL in mind, the only judging task is to identify URLs of equivalent (near-duplicate or translated) pages. As described previously, this task was carried out during the topic development phase.

From the assessments obtained during the topic development stage we are able define a number of qrel sets, including the following.

- *Monolingual* This set of qrels contains for each topic, the target page and all its duplicates.
- *Multilingual* This set of qrels contains for each topic, the target page, its duplicates and all its translations.

Table 3: Summary of the runs submitted for the Mixed-Monolingual task. The 'metadata usage' columns indicate usage of topic metadata: topic language (TL), page language (PL), page domain (PD), and user's native or active languages (UN, UA, respectively). For each team, its best scoring non-metadata run is in italics, and its best scoring metadata run is in boldface.

		Metadata usage					
Group id	Run name	TL	PL	PD	UN	UA	MRR
hummingbird	humWC05dp						0.4334
	humWC05dpD	Y	Y	Y			0.4707
	humWC05dplD	Y	Y	Y			**0.4780**
	humWC05p						0.4154
	humWC05rdp						*0.4412*
ilps	*UAmsMMBaseline*						*0.3497*
melange	BaselineMixed						0.0226
	AnchorMixed						*0.0260*
	DomLabelMixed			Y			**0.0366**
	LangCueMixed						0.0226
	LangLabelMixed	Y					0.0275
miracle	*MonoBase*						*0.0472*
	MonoExt			Y			0.1030
	MonoExtAH1PN			Y			0.1420
	MonoExtH1PN			Y			**0.1750**
	MonoExtUrlKy			Y			0.0462
ualicante	**final**	Y					**0.1191**
	final.lang						*0.0000*[1]
uglasgow	*uogSelStem*						*0.4683*
	uogNoStemNLP			Y			**0.5135**
	uogPorStem			Y			0.5107
	uogAllStem	Y		Y			0.4827
	uogAllStemNP	Y		Y			0.4828
uhildesheim	UHi3TiMo						0.0373
	UHiScoMo						0.1301
	UHiSMo						*0.1603*
uindonesia	*UI-001*						*0.2165*
	UI-002			Y			**0.2860**
	UI-003			Y			0.2714
usal	usal0	Y		Y			0.0537
	usal1	Y	Y				0.0685
	usal2	Y		Y			0.0626
	usal3	Y		Y			**0.0787**
	usal4	Y		Y			0.0668

[1] This run had an error in topic-result mapping. Corrected run has MRR of 0.0923.

- *User readable* This set of qrels contains for each topic, the target, all its duplicates, and all translations which are in a language that the topic author marked as her native/active/passive language.

Each of these qrel sets can be further divided into subsets based on the language of the topic or the domain of the target page. In this report we will only use the language base subsets.

The main metric used for evaluation was *mean reciprocal rank* (MRR).

Table 4: Summary of the runs submitted for the Multilingual task. The 'metadata usage' columns indicate topic metadat usage: topic language (TL), page language (PL), page domain (PD), and the user's native or active languages (UN, UA, respectively). MRR is reported using the monolingual, multilingual, and the user readable assessment sets. For each team, its best scoring non-metadata run is in italics, while its best scoring metadata run is in boldface.

		Metadata usage					MRR		
Group id	Run name	TL	PL	PD	UN	UA	mono	multi	u.r.
ilps	ILPSMuAll						0.0092	0.0097	0.0097
	ILPSMuAllR						0.0157	0.0164	0.0164
	ILPSMuFive						0.0109	0.0117	0.0117
	ILPSMuFiveR						*0.0166*	*0.0175*	*0.0175*
melange	BaselineMulti						0.0082	0.0091	0.0091
	AnchorMulti						0.0074	0.0083	0.0083
	AccLangsMulti				Y	Y	0.0082	0.0092	0.0092
	LangCueMulti						*0.0086*	*0.0092*	*0.0092*
	SuperMulti			Y			**0.0086**	**0.0092**	**0.0092**
miracle	*MultiBase*						*0.0314*	*0.0401*	*0.0387*
	MultiExt			Y			0.0588	0.0684	0.0669
	MultiExtAH1PN			Y			0.0633	0.0736	0.0733
	MultiExtH1PN			Y			**0.0762**	**0.0903**	**0.0902**
	MultiExtUrlKy			Y			0.0338	0.0397	0.0383
uhildesheim	UHi3TiMu						0.0274	0.0282	0.0282
	UHiScoMu						0.1147	0.1235	0.1225
	UHiSMu						*0.1370*	*0.1488*	*0.1479*
	UHi3TiMuBo91						0.0139	0.0160	0.0159
	UHiSMuBo91						0.0815	0.0986	0.0974

3 Submitted Runs

Table 1 shows a summary of the number of runs submitted by each team. The mixed-monolingual task was the most popular task with 34 runs submitted by 9 teams; Table 3 provides details of the runs submitted. The multilingual task was the second most popular task with 19 runs submitted by 4 teams; the details are given in Table 4. For the bilingual English to Spanish task, 8 runs were submitted by 3 teams; consult Table 5 for details.

We will now provide an overview of features used by the participating teams. We divide the overview in three parts: *web-specific*, *linguistic*, and *cross-lingual* features.

The teams used a wide variety of web-based features. Many teams indexed titles separately: Hummingbird, Miracle, U. Alicante, U. Glasgow, U. Indonesia, and U. Salamanca. A few teams

Table 5: Summary of the runs submitted for the BiEnEs task. For each team, the score of its best scoring run is in boldface.

Group id	Run name	MRR
buap	BUAP_Full	0.0465
	BUAP_PT10	0.0331
	BUAP_PT40	**0.0844**
	BUAP_PT60	0.0771
	BUAP_PT20	0.0446
ualicante	**BiEn2Es**	**0.0395**
uned	UNED_bilingual_baseline	0.0477
	UNED_bilingual_exp1	**0.0930**

also build special indexes for other HTML tags: Hummingbird, Miracle, and UNED. Several teams used separate index for anchor text: Melange, U. Glasgow, and U. Salamanca. Miracle also built an index for URL text. Hummingbird, U. Glasgow and U. Salamanca used URL length in their ranking. PageRank was used by Melange and U. Salamanca. Neither U. Amsterdam (ILPS) nor U. Hildesheim used any web-specific features.

The teams also used a wide variety of linguistic features. Language specific stemming was performed by a number of teams: Hummingbird, Melange, U. Alicante, and U. Glasgow. U. Amsterdam (ILPS) limited themselves to a simple accent normalization, but did do a ASCII transliteration for Russian. Miracle extracted proper nouns and keywords and indexed separately. U. Hildesheim experimented with character tri-grams. U. Indonesia did not use any language specific features. U. Salamanca applied a special stemmer for Spanish.

In the multilingual task, two different techniques were used by participating groups to bridge the gap between query language (English) and target page language. Neither U. Hildesheim nor Miracle did use any translation. I.e., both teams used simply the English version of the topics. Both ILPS and Melange used an on-line translator.

In the bilingual English to Spanish task two different approaches were used to translate the English queries to Spanish. UNED used an English to Spanish dictionary, but BUAP and U. Alicante use on-line translators.

4 Results

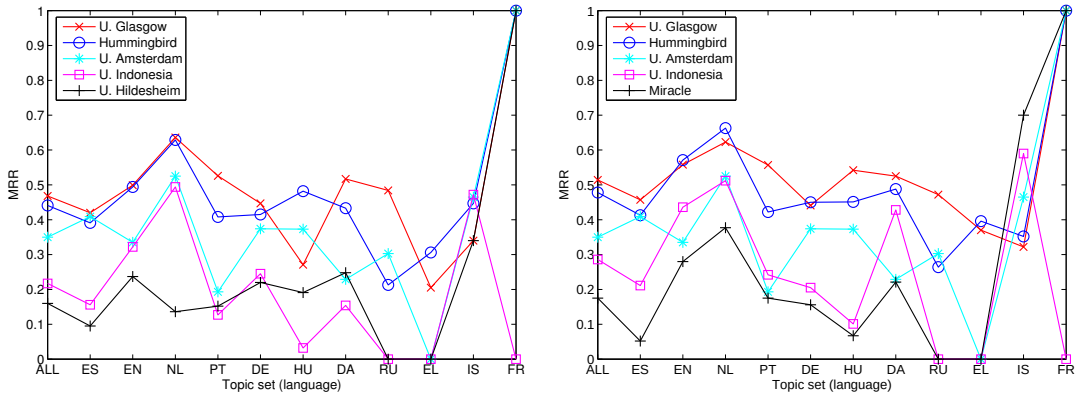
4.1 Mixed-Monolingual Task

First we look at each team's best scoring baseline run. Figure 2 (left) shows the scores of the 5 best scoring teams. The left-most point shows the MRR over all topics. The successive points show MRR scores for a subset of the topics: one for each language. The languages are sorted by the number of topics: from Spanish (ES) with the most topics (134) to French (FR) with only one topic.

Now, let's look at each team's best scoring run, independent of whether it was a baseline run or used some of the topic metadata. Figure 2 (right) shows the scores of the 5 best scoring teams. For the top scoring teams only U. Amsterdam (ILPS) uses no metadata.

Observe that, for each of the top five scoring runs, there is a considerable amount of variation across languages. For some languages the "hardness" seems independent of systems. Most systems score relatively high for Dutch; relatively low for Russian and Greek; but the score for German is close to their average score. The different performance between languages is only partially caused

Figure 2: Scores per-language for the 5 best scoring runs for the Mixed-Monolingual task using MRR. **(Left)**: Best scoring baseline run per team. **(Right)**: Best scoring run per team.



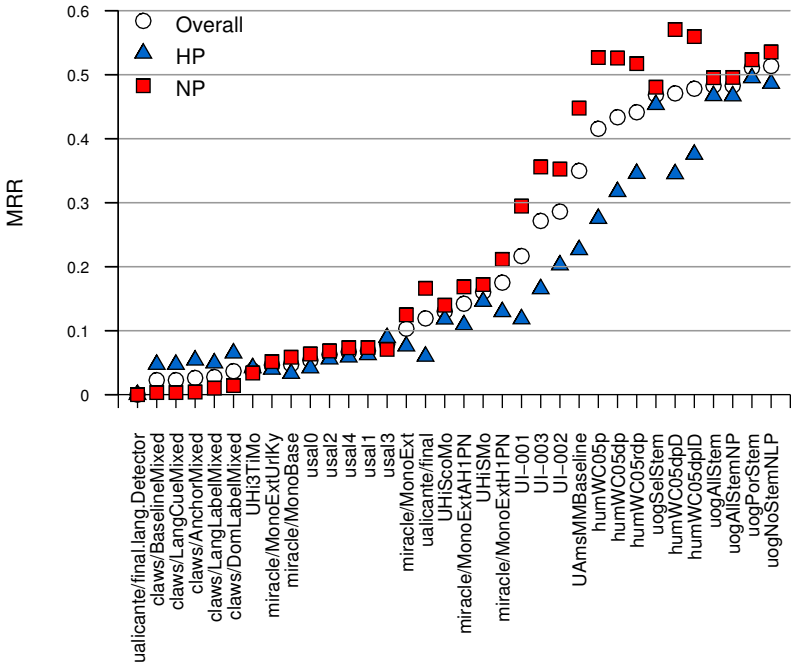


Figure 3: Homepages vs. named pages.

by the "hardness" of the particular language. Since the topics are not the same across languages, the "hardness" of the topics may also play a role.

Let's turn to the use of metadata now. The highest scoring runs are ones that use metadata. No team used user metadata; information about the domain of the target page proved to be the most popular type of metadata, and using it to restrict retrieval systems' outputs seems to be a sensible strategy, as is witnessed by the fact that it's the only type of metadata that each of the 5 top ranking runs uses.

Finally, for many runs, there is a clear gap between scores for NPs and HPs, with the named page queries scoring higher than the home page queries. For the best scoring runs, both types of known-item topics in relative balance. This phenomenon is illustrated in Figure 3, and mirrors a similar phenomenon at TREC's web track in 2003 and 2004 [1].

4.2 Multilingual Task

For the multilingual task we can actually look at 3 tasks. The tasks differ w.r.t. the translations being used in the qrels. Figure 4 (Top row) shows the results if only the target page and its duplicates are considered relevant. The second row shows the results if all translations are added to the relevant set. And the bottom row shows the results if only "user readable" translations are added to the relevant set. From Table 4 we see that the overall MRR increases when translations are added to the relevant set. This effect is, obviously, due to an increase in the amount of relevant pages. There is little difference between the two sets of translations, which may have been caused by several reasons: such as the completeness of the translation identification is not known, and there might be a bias toward identifying "readable" translations rather than "unreadable" translations. Note that, the relative ranking of the submitted runs does not change if translations are added to the relevant set.

Table 6: Non-English queries with the highest mean MRR over all runs submitted to the multilingual track

Topic	Lang.	Original query	English query
WC0528	Dutch	cv balkenende	cv balkenende
WC0185	German	Europa Newsletter	Europa Newsletter
WC0070	French	Le professeur Henri Muller nommé Ambassadeur de l'Hellénisme	Prof. Henri Muller named ambassador for Hellenism
WC0232	Danish	Regeringen Poul Hartling	The cabinet of Poul Hartling
WC0456	Icelandic	upplýsingar um europol	europol factsheet
WC0404	Dutch	CV minister-president Jan-Peter Balkenende	CV of the Dutch prime minister Jan-Peter Balkenende
WC0149	German	Ernst Breit 80. Geburtstag	80th birthday of Ernst Breit
WC0536	German	Interviews mit Staatsminister Rolf Schwanitz	Interviews with Minister of State Rolf Schwanitz
WC0025	Greek	–	Historical sources of the Hellenic parliament
WC0198	Spanish	El Palacio de la Moncloa	Moncloa Palace
WC0327	German	Autobahn Südumfahrung Leipzig	Southern Autobahn Ring Road of Leipzig
WC0202	Danish	Dansk Færøsk kulturfond	danish faroese culture fund
WC0497	Greek	–	Home page of the Hellenic parliament for kids
WC0491	German	Francesca Ferguson Architektur-Biennale 2004	Francesca Ferguson for Germany at achitecture Biennale 2004

The highest MRR for the multilingual task is substantially lower than the highest MRR for the mixed monolingual task: 0.1370 vs. 0.5135. The top score of the best scoring team on the multilingual task, U. Hildesheim, is over 14% below their top score on the mixed monolingual task. For the teams that score second and third best on the multilingual task, the corresponding differences are even more dramatic (56% for Miracle, and 95% for U. Amsterdam).

The success of approaches which did not apply translation is interesting and deserves a closer look. Let's look at the 40 topics which received the highest mean MRR over all submitted runs, using the monolingual result set. Thereof, 26 topics are in English. The remaining 14 topics are listed in Table 6. For the high scoring non-English topics we see that proper names are common, such as Jan-Peter Balkenende, Henri Muller, Paul Hartling, Europol etc. For these queries a translation is hardly needed.

It is difficult to say whether metadata helped in the multilingual task, since we have very few runs to compare. It is however tempting to say that the metadata did indeed help Miracle.

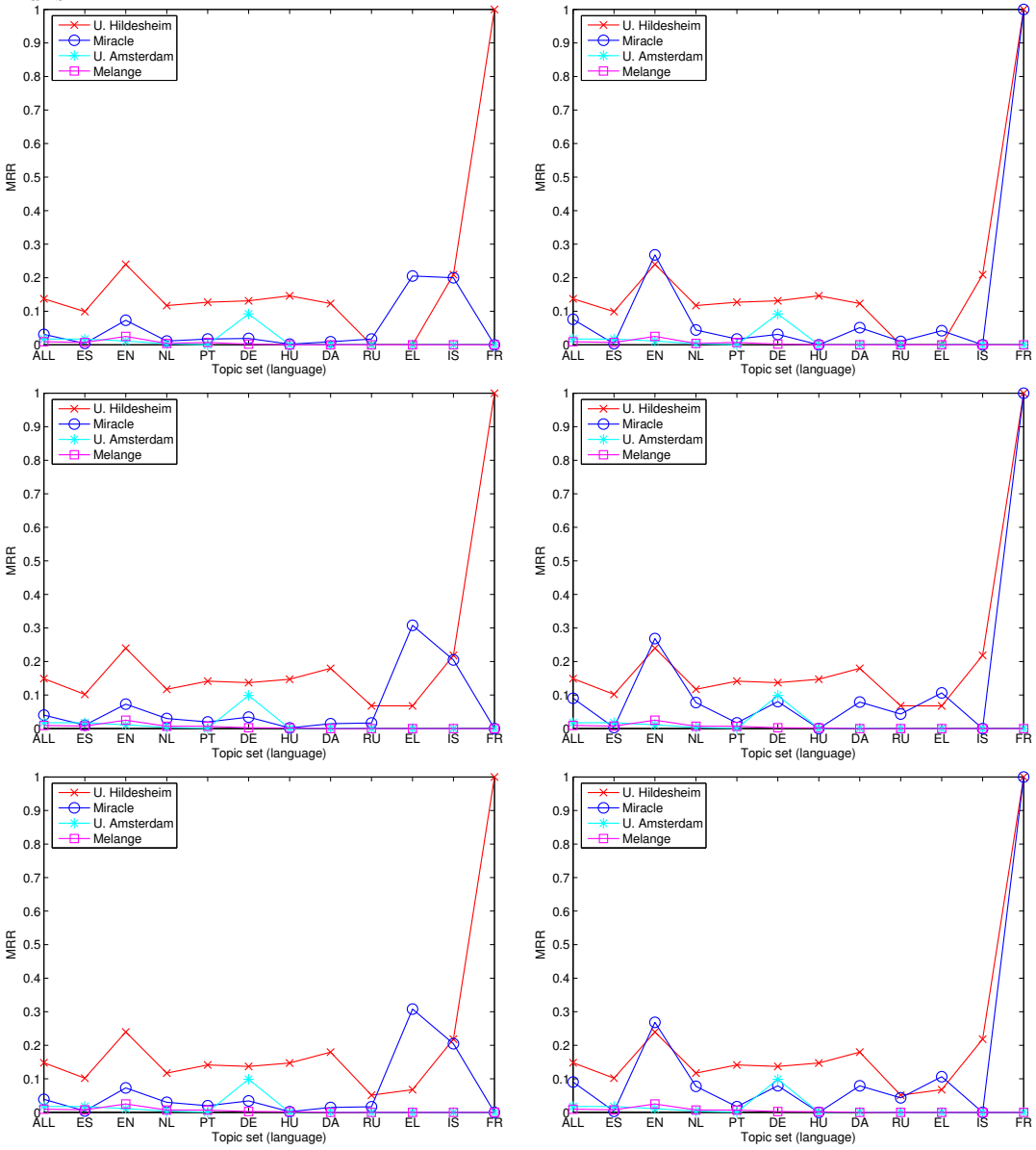
4.3 Bilingual English to Spanish Task

The results for the bilingual English to Spanish task can be seen from Table 5. We refer to the individual participants' papers for a more detailed analysis of the results.

5 Conclusions

The mixed monolingual task proved an interesting addition to the range of tasks in cross-language information retrieval. A number of participant build effective systems, that cope well with all the eleven languages in the topic set. Specific web-centric techniques or additional knowledge from the metadata fields leads to further improvement. Although it may be too early to talk about a solved problem, effective web retrieval techniques seem to carry over to the multilingual setting. The

Figure 4: **(Top row)**: Scores per-language for the best scoring runs for the Multilingual task using MRR and only target pages and duplicates. (Left): Baseline runs. (Right): All runs. **(Second row)**: Scores per-language for the 5 best scoring runs for the Multilingual task using MRR and target pages, duplicates and ALL translations. (Left): Baseline runs. (Right): All runs. **(Bottom row)**: Scores per-language for the best scoring runs for the Multilingual task using MRR and target pages, duplicates and *user readable* translations. (Left): Baseline runs. (Right): All runs.



multilingual task, in contrast, is still very far from being a solved problem. Remarkably, using non-translated English queries proved more successful than to use translations of the English queries. A closer look at the best scoring queries revealed that a large portion of them had indeed an English target. As for the best scoring queries which had non-English target, a majority contained a proper name which does not require translation.

Future of WebCLEF WebCLEF 2005 was an important first step toward a cross lingual web retrieval test collection. There are a number of steps that can be taken to further improve the quality of the current test collection. Here we list a few.

- *User data* More user data was collected during topic development phase than was used as topic metadata. This serves as an important resource to better understand the challenges of multilingual web retrieval. The data is available to all groups who participated in the topic development process.
- *Duplicates* It is not clear how complete the duplicate detection is. It remains as future work to investigate this completeness. Furthermore, we need to analyze how incomplete duplicate detection affects system ranking.
- *Translations* As with duplicates, the translations are likely to be incomplete. It is rather complicated to achieve complete list of translations. It remains as future work to investigate if the creation of the set of translation can be partly automated.

If we look a bit further ahead and speculate about future WebCLEF tasks, there are a number of new tasks we can look at.

- *X to English* Non-native English speakers are often more comfortable with posting queries in their native language even if they have no problem with reading English results.
- *Ad-hoc retrieval* If assessment resources are allocated for the WebCLEF task it would be possible to do ad-hoc retrieval.

6 Acknowledgments

We want to thank the participating teams for their valuable input that helped to make this test collection a reality. We are thankful to the University of Glasgow for providing additional search engine access to the collection during the topic development phase. We thank UNED for providing a reviewed set of translations for the bilingual English to Spanish task. We would like to thank Ian Soboroff and TREC for their help with creating the topic development guidelines.

Jaap Kamps was supported by a grant from the Netherlands Organization for Scientific Research (NWO) under project numbers 612.066.302 and 640.001.501. Maarten de Rijke was supported by grants from NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, and 612.069.006.

References

[1] N. Craswell and D. Hawking. Overview of the TREC-2004 Web Track. In *Proceedings TREC 2004*, 2005.

[2] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. EuroGOV: Engineering a Multilingual Web Corpus. In *This Volume*, 2005.