Overview of WebCLEF 2006

Krisztian Balog[1] Leif Azzopardi[3] Jaap Kamps[1,2] Maarten de Rijke[1]
[1] ISLA, University of Amsterdam
[2] Archive and Information Studies, University of Amsterdam
kbalog,kamps,mdr@science.uva.nl
[3] Department of Computer and Information Sciences, University of Strathclyde
Leif.Azzopardi@cis.strath.ac.uk

Abstract

We report on the CLEF 2006 WebCLEF track devoted to crosslingual web retrieval. We provide details about the retrieval tasks, the used topic set, and the results of WebCLEF participants. WebCLEF 2006 used a stream of known-item topics consisting of: (i) manual topics (including a selection of WebCLEF 2005 topics, and a set of new topics) and (ii) automatically generated topics (generated using two techniques). Our main findings are the following. First, the results over all topics show that current CLIR systems are quite effective, retrieving on average the target page in the top few ranks. Second, when we break down the scores over the manually constructed and the generated topics, we see that the manually constructed topics result in higher performance. Third, the resulting scores on automatic topics give, at least, a solid indication of performance, and can hence be an attractive alternative in situations where manual topics are not readily available.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Web retrieval, Known-item retrieval, Multilingual retrieval

1 Introduction

The world wide web presents one of the greatest challenges for cross-language information retrieval [5]. Content on the world wide web is essentially multilingual, and web users are often polyglots. The European web space is a case in point: the majority of European speak at least one language other than their mother-tongue, and the Internet is a frequent reason to use a foreign language [4]. The challenge of crosslingual web retrieval is addressed, head-on, by WebCLEF [9].

The crosslingual web retrieval track uses an extensive collection of spidered web sites of European governments, baptized EuroGOV [7]. The retrieval task at WebCLEF 2006 is based on a stream of known-item topics in a range of languages. This task, which is labeled *mixed-monolingual retrieval*, was pioneered at the WebCLEF 2005 [8]. Participants of WebCLEF 2005 expressed the wish to be able to iron out issues with the systems they built during last year's campaign, since for many it was their first attempt at web IR with lots of languages, encoding issues, different formats, and noisy data. The continuation of this known-item retrieval task at WebCLEF 2006 allows veteran participants to take stock and make meaningful comparisons of their results over years. To facilitate this, we decided to include a selection of WebCLEF 2005 topics in the topic set (also available for training purposes), as well as a set of new known-item topics. Furthermore, we decided to experiment with the automatic generation of known-item topics [2]. By contrasting the human topics with the automatically generated topics, we hope to gain insight in the validity of the automatically generated topics, especially in a multilingual environment. Our main findings are the following. First, the results over all topics show that current CLIR systems are quite effective, retrieving on average the target page in the top few ranks. Second, when we break down the scores over the manually constructed and the generated topics, we see that the manually constructed topics result in higher performance. Third, the resulting scores on automatic topics give, at least, a solid indication of performance, and can hence be an attractive alternative in situations where manual topics are not readily available.

The remainder of this paper is structured as follows. Section 2 gives the details of the method for automatically generating known-item topics. Next, in Section 3, we discuss the details of the track set-up: the retrieval task, document collection, and topics of request. Section 4 reports the runs submitted by participants, and Section 5 discusses the results of the official submissions. Finallly, in Section 6 we discuss our findings and draw some initial conclusions.

2 Automatic Topic Construction

This year we experimented with the automatic generation of known-item topics. The main advantage of automatically generating queries is that for any given test collection numerous queries can be produced at minimal cost [2]. In the WebCLEF setting this could be especially rewarding, since manual development of topics on all the different languages would require human resources we do not dispose of.

To create simulated queries, we model the following behavior of a known-item searcher. We assume that the user wants to retrieve a particular document that they have seen before in the collection, because some need has arisen calling for this document. The user then tries to reconstruct or recall terms, phrases and features that would help identify this document, which they pose as a query.

The basic algorithm we use for generating queries was introduced by Azzopardi and de Rijke [2], and is based on an abstraction of the actual querying process, as follows:

- Initialize an empty query $q = \{\}$
- Select the document d to be the known-item with probability $p(d)$
- Select the query length k with probability $p(k)$
- Repeat k times:
  - Select a term t from the document model of d with probability $p(t|\theta_d)$
  - Add t to the query q.
- Record d and q to define the known-item/query pair.

By repeatedly performing this algorithm we can create many queries. Before doing so, the probability distributions $p(d)$, $p(k)$ and $p(t|\theta_d)$ need to be defined. By using different probability distributions we can characterize different types and styles of queries that a user may submit.

Azzopardi and de Rijke [2] conducted experiments using various term sampling methods in order to simulate different styles of queries. In one case, they set the probability of selecting a term from the document model to a uniform distribution, where $p(t|\theta_d)$ was set to zero for all terms that did not occur in the document, whilst all other terms were assigned an equal probability.

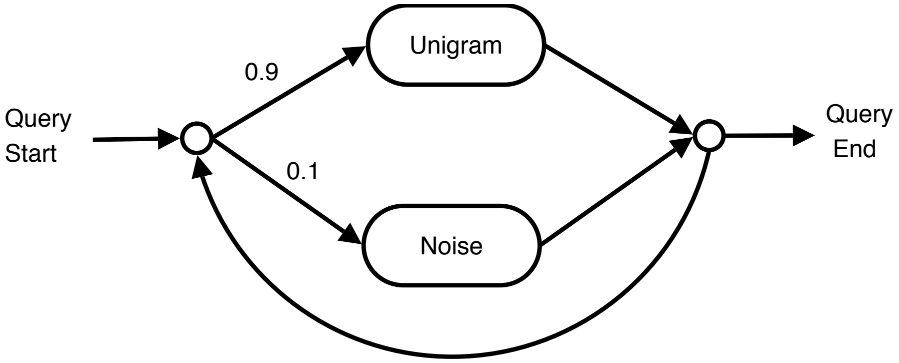

Figure 1: The process of *auto-uni* query generation

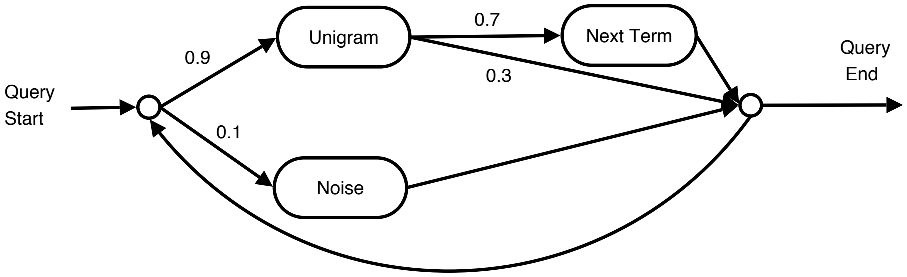

Figure 2: The process of *auto-bi* query generation

Compared to other types of queries, they found that using a uniform selection produced queries which were the most similar to real queries.

In the construction of a set of queries for the EuroGOV collection, we also use uniform sampling, but include query noise and then phrase extraction into the process to create more realistic queries. To include some noise to the process of generating a query, our model for sampling query terms is broken into two parts: sampling from the document (in our case uniformly) and sampling terms at random (i.e., noise). Figure 1 shows the sampling process; where a term is drawn from the unigram document model with some probability λ, or it is drawn from the noise model with probability $1 - \lambda$. Consequently, as λ tends to zero, we assume that the user has almost perfect recollection of the original document. Conversely, as λ tends to one, we assume that the user's memory of the document degrades to the point that they know the document exists but they have no idea as to the terms other than randomly selecting terms (from the collection). We used $\lambda = 0.1$ for topic generation. This model was used for our first setting, called *auto-uni*.

We further extended the process of sampling terms from a document. Once a term has been sampled from the document, we assume that there is some probability that the subsequent term will be drawn. For instance given the sentence, "... Information Retrieval Agent ...", if the first term sampled is "Retrieval", then the subsequent term selected will be "Agent". This was included to provide some notion of phrase extraction to the process of selecting query terms. The process is depicted in Figure 2. This model was used for our second setting, called *auto-bi*, where we either add the subsequent term with $p = 0.7$, or sample a new term independently from the document with $p = 0.3$.

We indexed each domain within the EuroGOV collection separately, using the Lemur language modeling toolkit [6]. We experimented with two different styles of queries, and for each of them we generated 30 queries per top level domain. For both settings, the query length k was selected using a Poisson distribution where the mean was set to 3. Two restrictions were placed on sampled query terms: (i) the size of a term needed to be greater than 3, and (ii) the terms should not contain

Table 1: Number of topics in the *original topic set*, and in the *new topic set* where we only retain topics for which at least one of the participants retrieved the relevant page.

	all	auto	auto-uni	auto-bi	manual	manual-o	manual-n
original	1,940	1,620	810	810	320	195	125
new	1,120	817	415	402	303	183	120
deleted	820	803	395	408	17	12	5

any numeric characters. Finally, the document prior $p(d)$ was also set to a uniform distribution.

Our initial results motivate further work with more sophisticated query generators. A natural next step would be to take structure and document priors into account.

3 The WebCLEF 2006 Tasks

3.1 Document Collection

For the purposes of the WebCLEF track the EuroGOV corpus was developed [7]. 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.

3.2 Topics

The topic set for WebCLEF 2006 consists of a stream of 1,940 known-item topics, consisting of both manual and automatically generated topics. As is shown in Table 1, 195 manual topics were reused from WebCLEF 2005, and 125 new manual topics were constructed. For the generated topics, we focused on 27 primary domains and generated 30 topics using the auto-uni query generation, and another 30 topics using the auto-bi query generation (see Section 2 for details), amounting to 810 automatic topics for each of the methods.

After the runs had been evaluated, we observed that the performance achieved on the automatic topics are frequently very poor. We found that in several cases none of the participants found any relevant page within the top 50 returned results. These are often mixed-language topics, a result of language diversity within a primary domain, or they proved to be too hard for any other reason.

In our post-submission analysis we decided to zoom in on a subset of topics and removed any topics that did not meet the following criterion: "whether any participant found the targetted page within the top 50." Table 1 presents the number of original, deleted and remaining topics. 820 out of the $1,940$ original topics were removed. Most of the removed topics are automatic (803), but there are also a few manual ones (17). The remaining topic set contains 1,120 topics, and is referred as the *new topic set*.

We decided to re-evaluate the submitted runs using this *new* topic set. Since it is a subset of the original topic collection, participants did not have to make any efforts. Submitted runs were re-evaluated using a restricted version of the (original) qrels that correspond to the new topic set.

3.3 Retrieval Task

WebCLEF 2006 saw the continuation of the *Mixed Monolingual* task of WebCLEF 2005 [8]. The mixed-monolingual task is meant to simulate a user searching for a known-item page in a European language. The mixed-monolingual task uses the title field of the topics to create a set of monolingual known-item topics.

Our emphasis this year is on the mixed monolingual task. The manual topics in the topic set contain an English translation of the query. Hence, using only the manual topics, experiments with a *Multilingual* task are possible. 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.

3.4 Submission

For each task, participating teams were allowed to submit up to 5 runs. The results had to be submitted in TREC format. For each topic a ranked list of no more than 50 results should be returned. For each topic at least 1 result must be returned. Participants were also asked to provide a list of the metadata fields they used, and a brief description of the methods and techniques employed.

3.5 Evaluation

The WebCLEF 2006 topics were known-item topics where a unique URL is targetted (unless there are page-duplicates in the collection, or near duplicates). Hence, we opted for a precision measure. The main metric used for evaluation was *mean reciprocal rank* (MRR). The reciprocal rank is, indeed, calculated as 1 divided by the rank at which the (first) relevant page is found. The mean reciprocal rank is obtained by, indeed, averaging the reciprocal ranks of a set of topics.

4 Submitted Runs

There were 8 participating teams that managed to submit official runs to WebCLEF 2006: buap; depok; hildesheim; hummingbird; isla; reina; rfia; and ucm. For details of the respective retrieval approaches to crosslingual web retrieval, we refer to the participants papers.

Table 2 lists the runs submitted to WebCLEF 2006: 35 for the mixed-monolingual task, and 1 for the bilingual task. We also indicate the use of topic metadata, either the topic's language (TL), the targetted page's language (PL), or the targetted page's domain (PD). The mean reciprocal rank (MRR) is reported over both the original and the new topic set. The official results of WebCLEF 2006 were based on the original topic set containing 1,940 topics. As detailed in Section 3.2 above, we have pruned the topic set by removing topics for which none of the participants retrieved the target page, resulting in 1,120 topics. In Appendix A, we provide scores for various breakdowns for both the original topic set and the new topic set.

The task description stated that for each topic, at least 1 result must be returned. However, several runs did not fulfill this condition. The best results for each team were achieved using 1 or more metadata fields. Knowledge of the page's primary domain (shown in the PD column in Table 2) seemed moderately effective.

5 Results

This year our focus is on the Mixed-Monolingual task. A large number of topics were made available, consisting old manual, new manual, and automatically generated topics. Evaluation results showed that the performance achieved on the automatic topics are frequently very poor, and we made a new topic set where we removed topics for which none of the participants found any relevant page within the top 50 returned results. All the results presented in this section correspond to the new topic set consisting of 1,120 topics.

Table 2: Summary of all runs submitted to WebCLEF 2006. The 'metadata usage' columns indicate usage of topic metadata: topic language (TL), page language (PL), page domain (PD). Mean Reciprocal Rank (MRR) scores are reported for both the original and the new topic set. For each team, its best scoring non-metadata run is in italics, and its best scoring metadata run is in boldface. Scores reported at the Multilingual section are based only on the manual topics.

		Metadata usage			topics	
Group id	Run name	TL	PL	PD	original	new
Monolingual runs:						
buap	*allpt40bi*			Y	*0.0157*	*0.0272*
depok	UI1DTA	Y			0.0404	0.0699
	UI2DTF	Y			**0.0918**	**0.1589**
	UI3DTAF	Y			0.0253	0.0439
	UI4DTW	Y			0.0116	0.0202
hildesheim	UHi1-5-10			Y	0.0718	0.1243
	UHi510			Y	0.0718	0.1243
	UHiBase			Y	**0.0795**	**0.1376**
	UHiBrf1			Y	0.0677	0. 1173
	UHiBrf2			Y	0.0676	0.1171
	UHiTitle			Y	0.0724	0.1254
hummingbird	humWC06				0.1133	0.1962
	humWC06dp				*0.1209*	*0.2092*
	humWC06dpc				0.1169	0.2023
	humWC06dpcD			Y	**0.1380**	**0.2390**
	humWC06p				0.1180	0.2044
isla	*Baseline*				*0.1694*	*0.2933*
	Comb				0.1685	0.2918
	CombMeta			Y	0.1947	0.3370
	CombNboost			Y	0.1954	0.3384
	CombPhrase			Y	**0.2001**	**0.3464**
reina	usal_base	Y			0.0100	0.0174
	usal_mix	Y			0.0137	0.0237
	USAL_mix_hp	Y	Y		**0.0139**	**0.0241**
	usal_mix_hp	Y			0.0139	0.0241
	usal_mix_hp_ok	Y			0.0139	0.0241
rfia	DPSinDiac	Y		Y	0.0982	0.1700
	ERConDiac	Y		Y	0.1006	0.1742
	ERFinal	Y		Y	**0.1021**	**0.1768**
	ERSinDiac	Y		Y	0.1021	0.1768
ucm	**webclef-run-all-2006**	Y			**0.0870**	**0.1505**
	webclef-run-all-2006-def-ok	Y			0.0870	0.1505
	webclef-run-all-2006-def-ok-2	Y			0.0870	0.1505
	webclef-run-all-2006-ok-conref	Y			0.0870	0.1505
	webclef-run-all-OK-definitivo	Y			0.0870	0.1505
Multilingual runs:						
hildesheim	*UHiMu*				*0.2553*	*0.2686*

Table 3: Best overall results using the *new topic set*. The results are reported on *all* topics, the *auto*matic and *manual* subsets of topics, and *average* is calculated from the auto and manual scores.

Group id	Run	all	auto	manual	average
isla	combPhrase	0.3464	0.3145	0.4411	0.3778
hummingbird	humWC06dpcD	0.2390	0.1396	0.5068	0.3232
depok	UI2DTF	0.1589	0.0923	0.3386	0.2154
rfia	ERFinal	0.1768	0.1556	0.2431	0.1993
hildesheim	UHiBase /5-10	0.1376	0.0685	0.3299	0.1992
ucm	webclef-run-all-2006-def-ok-2	0.1505	0.1103	0.2591	0.1847
buap	allpt40bi	0.0272	0.0080	0.0790	0.0435
reina	USAL_mix_hp	0.0241	0.0075	0.0689	0.0382

Table 4: Best runs using the automatic topics in the *new topic set*.

Group id	Run	auto	auto-uni	auto-bi
isla	combNboost	0.3145	0.3114	0.3176
rfia	ERFinal	0.1556	0.1568	0.1544
hummingbird	humWC06dpcD	0.1396	0.1408	0.1384
ucm	webclef-run-all-2006	0.1103	0.1128	0.1077
depok	UI2DTF	0.0923	0.1024	0.0819
hildesheim	UHiBase	0.0685	0.0640	0.0731
buap	allpt40bi	0.0080	0.0061	0.0099
reina	USAL_mix_hp	0.0075	0.0126	0.0022

5.1 Mixed-Monolingual

We look at each team's best scoring run, independent of whether it was a baseline run or used some of the topic metadata. Table 3 presents the scores of the participating teams. We report the results over the whole new qrel set (*all*), and over the *auto*matic and *manual* subsets of topics. What is striking is that the automatic topics proved to be more difficult than manual ones. This may be due in part to the fact that the manual topics cover 11 languages, but the generated topics cover all 27 domains in EuroGOV including the more difficult domains and languages. Another important factor may be the imperfections in the generated topics. Apart from the lower scores, the auto topics also dominate the manual topics in number. Therefore we also used the average of the auto and manual scores for ranking participants. Defining an overall ranking of teams is not straightforward, since one team may outperform another on the automatic topics, but perform worse on the manual ones. Still, we observe that participants can be unambiguously assigned into one out of three bins based on either the *all* or the *average* scores: the first bin consisting of hummingbird and isla; the second bin of depok, hildesheim, rfia, and ucm; and the third bin of buap and reina.

5.2 Evaluation on Automatic Topics

Automatic topics were generated using two different methods, as described in Section 2 above. The participating teams' scores did not show significant variance between the difficulty of topics, using the the two generators. Table 4 provides details of the best runs when evaluation is restricted to automatically generated topics only.

Note that the scores included in Table 4 are measured on the new topic set. Notice, by the way, that there is very little difference between the number of topics within the *new topic set* for the two automatic topic subsets (*auto-uni* and *auto-bi* in Table 1).

In general, the two query generation methods perform very similarly, and it is system specific whether one type of automatic topics is preferred over the other. Our initial results with auto-

Table 5: Best manual runs using the *new topic set*.

Group id	Run	manual	old	new
hummingbird	humWC06dpcD	**0.5068**	**0.4936**	0.5269
isla	combPhrase	0.4411	0.3822	**0.5310**
depok	UI2DTF	0.3386	0.2783	0.4307
hildesheim	UHi1-5-10	0.3299	0.2717	0.4187
ucm	webclef-run-all-2006-def-ok-2	0.2591	0.2133	0.3289
rfia	DPSinDiac	0.2431	0.1926	0.3201
buap	allpt40bi	0.0790	0.0863	0.0679
reina	USAL_mix_hp	0.0689	0.0822	0.0488

matically generated queries are promising, but still a large portion of these topics are not realistic. This motivates us to work further on more advanced query generation methods.

5.3 Evaluation on Manual Topics

The manual topics include 183 old and 120 new queries. Old topics were randomly sampled from last year's topics, while new topics were developed by *Universidad Complutense de Madrid (UCM)* and the track organizers. The new topics cover only languages for which expertise was available: Dutch, English, German, Hungarian, and Spanish.

In case of the old manual topics we witnessed improvements for all teams that took part in WebCLEF 2005, compared to their last year's scores. Moreover, we found that most participating systems performed better on the new manual topics, compared to the old ones. A possible explanation is the nature of the topics, namely the new topics may be more appropriate for know-item search. Also, language coverage of the new manual topics could play a role.

5.4 Comparing Rankings

We use Kendall's tau to determine correlations between the rankings of runs resulting from different topic sets. First, we find weak (0.2–0.4) to moderate (0.4–0.6) positive correlations between ranking of runs resulting from automatic topics, and rankings of runs resulting from manual topics, only new manual topics, and only old manual topics; see Table 6. The rankings resulting from the topics generated with the "auto-bi" method are somewhat more correlated with the manual rankings than the ranking resulting from the topics generated with the "auto-uni" method. A very strong

Table 6: Kendall tau rank correlation, two-sided p-value.

		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
manual-old	τ							
	p							

positive correlation (0.8–1.0) is found between the ranking of runs obtained using new manual topics and the ranking of runs resulting from using old manual topics. Note that the new topic set we introduced does not affect the relative ranking of systems, thus the correlation scores we reported here are exactly the same for the *original* and for the *new topic sets*.

5.5 Multilingual Runs

Our main focus this year was on the monolingual task, but we allowed submissions for multilingual experiments within the mixed-monolingual setup. The manual topics (both old and new ones) are provided with English titles. The automatically generated topics do not have English translations.

We received only one multilingual submission, from the *University of Hildesheim*. The evaluation of the multilingual run is restricted to the manual topics in the topic set, Table 2 summarizes the results of that run. A detailed breakdown over the different topic types is provided in Appendix A (Tables 7 and 8)

6 Conclusion

The world-wide-web is a natural reflection of the language diversity in the world, both in terms of web content as well as in terms of web users. Effective cross-language information retrieval (CLIR) techniques have clear potential for improving the search experience of such users. The WebCLEF track at CLEF 2006 attempts to realize some of this potential, by investigating known-item retrieval in a multilingual setting. Known-item retrieval is a typical search task on the web [3]. This year's track focused on mixed monolingual search, in which the topic set is a stream of known-item topics in various languages. This task was pioneered at WebCLEF 2005 [8]. The collection is based on the spidered content of web sites of European governments. This year's topic set covered all 27 primary domains in the collection, and contained both manually constructed search topics and automatically generated topics. Our main findings for the mixed-monolingual task are the following. First, the results over all topics show that current CLIR systems are quite effective. These systems retrieve, on average, the target page in the top few ranks. This is particularly impressive when considering that the topics of WebCLEF 2006 covered no less than 27 European primary domains. Second, when we break down the scores over the manually constructed and the generated topics, we see that the manually constructed topics result in higher performance. The manual topics consisted of both a set of newly constructed topics, and a selection of WebCLEF 2005 topics. For veteran participants, we can compare the scores over years, and we see progress for the old manual topics. The new manual topics (which were not available for training) seem to confirm this progress.

Building a cross-lingual test collection is a complex endeavor. Information retrieval evaluation requires substantial manual effort by topic authors and relevance assessors. In a cross-lingual setting this is particularly difficult, since the language capabilities of topic authors should sufficiently reflect the linguistic diversity of the used document collection. Alternative proposals to traditional topics and relevance assessments, such as term relevance sets, still require human effort (albeit only a fraction) and linguistic capacities by the topic author.[1] This prompted us to experiment with techniques for automatically generating known-item search requests. The automatic construction of known-item topics has been applied earlier in a monolingual setting [2]. At WebCLEF 2006, two refined versions of the techniques were applied in a mixed-language setting. The general set-up of the the WebCLEF 2006 track can be viewed as an experiment with automatically constructing topics. Recall that the topic set contained both manual and automatic topics. This allows us to critically evaluate the performance on the automatic topics with the manual topics, although the comparison is not necessarily fair given that the manual and automatic subsets of topics differ both in number and in the domains they cover. Our general conclusion on the automatic topics is a mixed one: On the one hand, our results show that there are still some substantial differences

[1] Recall that term relevance sets (T-rels) consisting of a set of terms likely to occur in relevant documents, and a set of irrelevant terms (especially disambiguation terms avoiding false-positives) [1].

between the automatic topics and manual topics, and it is clear that automatic topics cannot simply substitute manual topics. Yet on the other hand, the resulting scores on automatic topics give, at least, a solid indication of performance, and can hence be an attractive alternative in situations where manual topics are not readily available.

Acknowledgments Thanks to *Universidad Complutense de Madrid* (UCM) for providing additional Spanish topics.

Krisztian Balog was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.065.120 and 612.000.106. Jaap Kamps was supported by NWO under project numbers 612.066.302, 612.066.513, 639.072.601, and 640.001.501; and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

[1] E. Amitay, D. Carmel, R. Lempel, and A. Soffer. Scaling ir-system evaluation using term relevance sets. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 10–17. ACM Press, New York USA, 2004.

[2] L. Azzopardi and M. de Rijke. Automatic construction of known-item finding test beds. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 603–604, New York, NY, USA, 2006. ACM Press.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[4] Eurobarometer. Europeans and their languages. Special Eurobarometer 243, European Commision, 2006. URL: http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.

[5] F. C. Gey, N. Kando, and C. Peters. Cross language information retrieval: a research roadmap. *SIGIR Forum*, 36(2):72–80, 2002.

[6] Lemur. The Lemur toolkit for language modeling and information retrieval, 2005. URL: http://www.lemurproject.org/.

[7] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. EuroGOV: Engineering a multilingual Web corpus. In C. Peters, F. C. Gey, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke, editors, *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science*. Springer Verlag, Heidelberg, 2006.

[8] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Overview of WebCLEF 2005. In C. Peters, F. C. Gey, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke, editors, *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science*. Springer Verlag, Heidelberg, 2006.

[9] WebCLEF. Cross-lingual web retrieval, 2006. URL: http://ilps.science.uva.nl/WebCLEF/.

A Breakdown of Scores over Topic Types

We provide a breakdown of scores over the different topic types, both for the original topic set in Table 7 and for the new topic set in Table 8.

Table 7: Original topic set: breakdown of submission results over topic types (MRR) for all runs submitted to WebCLEF 2006. For each team, its best scoring run is in boldface.

RUN	ALL	AUTO			MANUAL		
	topics	all	uni	bi	all	old	new
buap							
allpt40bi	**0.0157**	**0.0040**	**0.0031**	**0.0049**	**0.0750**	**0.0810**	**0.0657**
depok							
UI1DTA	0.0404	0.0234	0.0296	0.0173	0.1263	0.1099	0.1522
UI2DTF	**0.0918**	**0.0466**	**0.0525**	**0.0406**	**0.3216**	**0.2611**	**0.4168**
UI3DTAF	0.0253	0.0142	0.0116	0.0168	0.0819	0.0644	0.1094
UI4DTW	0.0116	0.0025	0.0020	0.0030	0.0583	0.0284	0.1053
hildesheim							
UHi1-5-10	0.0718	0.0242	0.0231	0.0253	**0.3134**	0.2550	**0.4051**
UHi510	0.0718	0.0242	0.0231	0.0253	**0.3134**	0.2550	**0.4051**
UHiBase	**0.0795**	**0.0346**	**0.0328**	**0.0363**	0.3076	**0.2556**	0.3893
UHiBrf1	0.0677	0.0220	0.0189	0.0251	0.3000	0.2485	0.3812
UHiBrf2	0.0676	0.0221	0.0188	0.0253	0.2989	0.2464	0.3816
UHiTitle	0.0724	0.0264	0.0245	0.0283	0.3061	0.2542	0.3876
UHiMu *(multilingual)*	0.0489	0.0083	0.0063	0.0102	0.2553	0.2146	0.3192
hummingbird							
humWC06	0.1133	0.0530	0.0572	0.0488	0.4194	0.3901	0.4657
humWC06dp	0.1209	0.0528	0.0555	0.0501	0.4664	0.4471	0.4967
humWC06dpc	0.1169	0.0472	0.0481	0.0464	0.4703	0.4553	0.4939
humWC06dpcD	**0.1380**	**0.0704**	**0.0721**	**0.0687**	**0.4814**	**0.4633**	**0.5099**
humWC06p	0.1180	0.0519	0.0556	0.0482	0.4538	0.4252	0.4988
isla							
baseline	0.1694	0.1253	0.1397	0.1110	0.3934	0.3391	0.4787
comb	0.1685	0.1208	0.1394	0.1021	0.4112	0.3578	0.4952
combmeta	0.1947	0.1505	**0.1670**	0.1341	0.4188	**0.3603**	0.5108
combNboost	0.1954	**0.1586**	0.1595	**0.1576**	0.3826	0.3148	0.4891
combPhrase	**0.2001**	0.1570	0.1639	0.1500	**0.4190**	0.3587	**0.5138**
reina							
usal_base	0.0100	0.0028	0.0044	0.0011	0.0468	0.0550	0.0340
usal_mix	0.0137	0.0038	0.0065	0.0011	0.0640	0.0747	0.0472
USAL_mix_hp	**0.0139**	**0.0038**	**0.0065**	**0.0011**	**0.0655**	**0.0771**	**0.0472**
usal_mix_hp	0.0139	0.0038	0.0065	0.0011	0.0655	0.0771	0.0472
usal_mix_hp_ok	0.0139	0.0038	0.0065	0.0011	0.0655	0.0771	0.0472
rfia							
DPSinDiac	0.0982	0.0721	0.0736	0.0706	**0.2309**	**0.1808**	0.3098
ERConDiac	0.1006	0.0771	0.0795	0.0746	0.2203	0.1693	0.3006
ERFinal	**0.1021**	**0.0785**	**0.0803**	**0.0766**	0.2220	0.1635	**0.3140**
ERSinDiac	**0.1021**	**0.0785**	**0.0803**	**0.0766**	0.2220	0.1635	**0.3140**
ucm							
webclef-run-all-2006-def-ok-2	**0.0870**	**0.0556**	**0.0578**	**0.0534**	**0.2461**	**0.2002**	**0.3183**
webclef-run-all-2006-def-ok	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-2006-ok-conref	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-2006	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-OK-definitivo	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183

Table 8: New topic set: breakdown of submission results over topic types (MRR) for all runs submitted to WebCLEF 2006. For each team, its best scoring run is in boldface.

RUN	ALL	AUTO			MANUAL		
	topics	all	uni	bi	all	old	new
buap							
allpt40bi	**0.0272**	**0.0080**	**0.0061**	**0.0099**	**0.0790**	**0.0863**	**0.0679**
depok							
UI1DTA	0.0699	0.0465	0.0578	0.0348	0.1330	0.1171	0.1572
UI2DTF	**0.1589**	**0.0923**	**0.1024**	**0.0819**	**0.3386**	**0.2783**	**0.4307**
UI3DTAF	0.0439	0.0281	0.0226	0.0339	0.0862	0.0686	0.1130
UI4DTW	0.0202	0.0049	0.0038	0.0060	0.0613	0.0302	0.1088
hildesheim							
UHi1-5-10	0.1243	0.0480	0.0451	0.0510	**0.3299**	0.2717	**0.4187**
UHi510	0.1243	0.0480	0.0451	0.0510	**0.3299**	0.2717	**0.4187**
UHiBase	**0.1376**	**0.0685**	**0.0640**	**0.0731**	0.3238	**0.2724**	0.4023
UHiBrf1	0.1173	0.0436	0.0369	0.0505	0.3159	0.2648	0.3939
UHiBrf2	0.1171	0.0438	0.0367	0.0510	0.3147	0.2625	0.3943
UHiTitle	0.1254	0.0524	0.0479	0.0570	0.3222	0.2709	0.4005
UHiMu (*multilingual*)	0.0846	0.0164	0.0124	0.0205	0.2686	0.2286	0.3297
hummingbird							
humWC06	0.1962	0.1051	0.1116	0.0984	0.4416	0.4156	0.4812
humWC06dp	0.2092	0.1047	0.1084	0.1009	0.4910	0.4764	0.5132
humWC06dpc	0.2023	0.0937	0.0939	0.0935	0.4952	0.4852	0.5104
humWC06dpcD	**0.2390**	**0.1396**	**0.1408**	**0.1384**	**0.5068**	**0.4936**	**0.5269**
humWC06p	0.2044	0.1030	0.1086	0.0971	0.4777	0.4530	0.5154
isla							
baseline	0.2933	0.2485	0.2726	0.2237	0.4141	0.3614	0.4946
comb	0.2918	0.2394	0.2720	0.2058	0.4329	0.3812	0.5117
combmeta	0.3370	0.2985	**0.3259**	0.2701	0.4409	**0.3839**	0.5278
combNboost	0.3384	**0.3145**	0.3114	**0.3176**	0.4028	0.3355	0.5054
combPhrase	**0.3464**	0.3112	0.3199	0.3023	**0.4411**	0.3822	**0.5310**
reina							
usal_base	0.0174	0.0055	0.0087	**0.0023**	0.0493	0.0586	0.0351
usal_mix	0.0237	0.0075	0.0126	0.0022	0.0674	0.0796	0.0488
USAL_mix_hp	**0.0241**	**0.0075**	**0.0126**	0.0022	**0.0689**	**0.0822**	**0.0488**
usal_mix_hp	0.0241	0.0075	0.0126	0.0022	0.0689	0.0822	0.0488
usal_mix_hp_ok	0.0241	0.0075	0.0126	0.0022	0.0689	0.0822	0.0488
rfia							
DPSinDiac	0.1700	0.1429	0.1436	0.1422	**0.2431**	**0.1926**	0.3201
ERConDiac	0.1742	0.1528	0.1552	0.1503	0.2320	0.1804	0.3106
ERFinal	**0.1768**	**0.1556**	**0.1568**	**0.1544**	0.2337	0.1743	**0.3244**
ERSinDiac	**0.1768**	**0.1556**	**0.1568**	**0.1544**	0.2337	0.1743	**0.3244**
ucm							
webclef-run-all-2006-def-ok-2	**0.1505**	**0.1103**	**0.1128**	**0.1077**	**0.2591**	**0.2133**	**0.3289**
webclef-run-all-2006-def-ok	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-2006-ok-conref	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-2006	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-OK-definitivo	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289