The University of Amsterdam at WebCLEF 2005

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

Abstract

We describe the University of Amsterdam's participation in the WebCLEF track at CLEF 2005. We submitted runs for both the *mixed monolingual* task and the *multilingual* task.

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

In the CLEF 2005 WebCLEF track, we took part in two of the retrieval tasks. We took part in the WebCLEF *mixed monolingual* task. Our participation here was aimed at evaluating the effectiveness of standard ad hoc retrieval settings for a stream of topics in various languages. Our assumption was that this would shed new light on the robustness of modern information retrieval techniques.

We also took part in the WebCLEF *multilingual* task. Our participation here was aimed at evaluating the effectiveness of straightforwardly combining runs using a number of translations of the original English queries. Such methods have previously been used successfully at the CLEF multilingual ad hoc retrieval task [5, 6].

This paper is structured as follows. In Section 2 we describe our retrieval system as well as the approaches used for the two WebCLEF tasks in which we participate. Section 3 describes our official retrieval runs for WebCLEF 2005, and Section 4 discusses the results we have obtained. Finally, in Section 5, we offer some conclusions regarding our multilingual web retrieval efforts.

2 System Description

Our retrieval system is based on the Lucene engine with a number of home-grown extensions [3, 8].

2.1 Retrieval Approach

For our ranking, we used the default similarity measure in Lucene [8], i.e., for a collection D, document d and query q containing terms t_i:

$$sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where

$$\begin{aligned} tf_{t,X} &= \sqrt{\text{freq}(t,X)} \\ idf_t &= 1 + \log \frac{|D|}{\text{freq}(t,D)} \\ norm_d &= \sqrt{|d|} \\ coord_{q,d} &= \frac{|q \cap d|}{|q|} \\ norm_q &= \sqrt{\sum_{t \in q} tf_{t,q} \cdot {idf_t}^2} \end{aligned}$$

2.2 Tokenization

We indexed the whole collection by simply extracting the full text from the documents. We did not apply any stemming nor did we use a stopword list. We applied case-folding and normalized marked characters to their unmarked counterparts, i.e., mapping ö to o, æ to ae, î to i, etc. The only language specific processing we did was a transformation of the multiple Russian encodings into an ASCII transliteration.

2.3 Translation

We used the WorldLingo machine translation [9] for translating the English topic statements into eight languages: Dutch, French, German, Greek, Italian, Portuguese, Russian, and Spanish. Combined with the English source topic statements, this gave us short topic statements in nine European languages.

2.4 Combination

We combined various 'base' runs using the unweighted CombSUM function of Fox and Shaw [2]. The runs were combined after normalizing the retrieval status values (RSVs) to the interval [0,1] as suggested in [7].

3 Runs

3.1 Mixed-monolingual task

We submitted one run to the mixed-monolingual task. The run uses the short topic statement in the ⟨title⟩ field of the WebCLEF 2005 topics. Our run uses Lucene's standard ranking formula applied on our full-text index (as discussed in Section 2 above).

3.2 Multilingual task

We submitted four runs to the multilingual task. All runs use the English version of the short topic statement in the ⟨translation language="EN"⟩ field of the WebCLEF 2005 topics, and the translations mentioned in Section 2.3.

We experimented along two dimensions. The first dimension is the number of topic languages:

	MRR	S@1	S@5	S@10
All topics	0.3497	0.2523	0.4589	0.5576
Home pages	0.2263	0.1322	0.3347	0.4380
Named pages	0.4476	0.3475	0.5574	0.6525

Table 1: Mixed Monolingual Task results by mean reciprocal rank (MRR) and success at rank 1, rank 5 and rank 10 (S@1, S@5, and S@10 respectively). We provide the scores over all topics, as well as a breakdown in home page finding and named page finding topics.

All translations Assuming that we have no knowledge of the language of the desired pages for each of the topics, it makes sense to use all available translations. That is, we use the topics in all nine languages available.

Five languages Based on knowledge of the languages in the WebCLEF topic set, we restrict the set of languages to those that occur frequently and for which we have reasonable translation methods. That is, we use the topics in the five languages: Dutch, English, German, Portuguese, and Spanish.

Recall that WebCLEF provides a stream of topics, with topics from arbitrary languages. For the multilingual task, we use the English short topic statement. The downside of this is, of course, that finding the targeted page in the source language becomes a formidable problem. The upside is that, at least, the topic language is known, and the same holds for the translations we obtained. The second dimension we experiment with is trying to exploit this knowledge:

All results Topics in one language may likely retrieve pages in other languages as well. A case in point is WebCLEF topic WC0014, whose English topic statement ("Chancellery at the Spreebogen") could still allow us to retrieve German pages targeted by the German topic statement ("*Bundeskanzleramt am Spreebogen*"). Hence, we may simply use all pages retrieved by a topic of a particular, known language.

Language restricted Since we know the language of the topic in each of the translations, and the intention of the translated topic is to retrieve pages in that language, we may decide to restrict the pages returned by our retrieval system. We do this by restricting retrieved pages to the dominant domains. For example, for a run with the topics translated to Dutch, we restrict pages to come from either the `.nl` or the `.eu.int` domain.

Combining the two dimensions naturally suggests the four following cases:

1. using nine topic languages without restriction;
2. using nine topic languages and restricting pages to dominant domains;
3. using five topic languages without restriction; and
4. using five topic languages and restricting pages to dominant domains.

For each of the cases, we obtain five to nine different runs, which we combine using unweighted CombSUM. This results in the four runs submitted to WebCLEF 2005.

4 Results

4.1 Mixed-monolingual task

We submitted a single no-thrills run for the mixed monolingual task, using standard ad hoc document retrieval setting (as discussed in Section 3). Table 1 reports the result of the mixed monolingual run. A number of observations present themselves. First, we see that, on average, the desired page is found in the top three. That is a reassuring result for the mixed monolingual

All topics	MRR	S@1	S@5	S@10
Nine languages	0.0092	0.0055	0.0073	0.0165
Nine languages, restricted	0.0157	0.0091	0.0201	0.0219
Five languages	0.0109	0.0055	0.0091	0.0165
Five languages, restricted	0.0166	0.0091	0.0201	0.0238

Home pages	MRR	S@1	S@5	S@10
Nine languages	0.0072	0.0041	0.0083	0.0124
Nine languages, restricted	0.0157	0.0124	0.0165	0.0165
Five languages	0.0084	0.0041	0.0083	0.0124
Five languages, restricted	0.0163	0.0124	0.0165	0.0207

Named pages	MRR	S@1	S@5	S@10
Nine languages	0.0109	0.0066	0.0066	0.0197
Nine languages, restricted	0.0158	0.0066	0.0230	0.0262
Five languages	0.0129	0.0066	0.0098	0.0197
Five languages, restricted	0.0168	0.0066	0.0230	0.0262

Table 2: Multilingual Task results by mean reciprocal rank (MRR) and success at rank 1, rank 5 and rank 10 (S@1, S@5, and S@10 respectively). We provide the scores over all topics (top), as well as a breakdown in home page finding (middle) and named page finding topics (bottom).

	Restricted to language					All 547 topics		
	# Topics	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Dutch⋆	59	0.2709	0.2203	0.3051	0.3729	0.0540	0.0420	0.0640	0.0823
English⋆	121	0.3289	0.2149	0.4628	0.5702	0.0882	0.0585	0.1207	0.1499
French	1	1.0000	1.0000	1.0000	1.0000	0.0303	0.0201	0.0366	0.0494
German⋆	57	0.2008	0.1754	0.1930	0.2807	0.0447	0.0329	0.0530	0.0695
Greek	16	0.0000	0.0000	0.0000	0.0000	0.0204	0.0146	0.0256	0.0329
Italian	0	–	–	–	–	0.0284	0.0201	0.0366	0.0475
Portuguese⋆	59	0.1047	0.0508	0.1525	0.1695	0.0412	0.0256	0.0567	0.0713
Russian	30	0.0127	0.0000	0.0333	0.0333	0.0446	0.0293	0.0567	0.0750
Spanish⋆	134	0.2272	0.1791	0.2687	0.3582	0.0809	0.0603	0.0969	0.1316

Table 3: Mixed Monolingual Task results by mean reciprocal rank (MRR) and success at rank 1, rank 5 and rank 10 (S@1, S@5, and S@10 respectively). We provide the scores over all topics for each of the topic translations (not submitted as official runs). We submitted runs based on all languages, or on the five languages with a ⋆.

task. Somewhat worrying is the success rate at rank 10, with no relevant page found for over 40% of the topics. Second, named page topics score somewhat higher than home page topics, on all measures. This is well-known from other web retrieval tasks [1], which also suggests that the scores for home page finding can be substantially improved using specific web centric techniques such as various document representations and non-content priors [4].

4.2 Multilingual task

We submitted four runs for the multilingual task (as discussed in Section 3). We will first look at the overall results, and then focus on the effectiveness for each of the languages in which we translated the English topics.

Table 2 reports the result of the multilingual runs. Again, we make a number of observations. First, we see that scores are substantially lower than for the mixed monolingual task. The complexity of the multilingual task can hardly be overestimated: given an English query we have to guess what page in any language has to be returned to the user. Obvious ways of limiting this

wealth of options are the use of topic meta-fields, or of sophisticated techniques to extract target language cues. Second, our experiment with the number of translations to use, points conclusively to the smaller set of five language used frequently in the topic set. It is a reassuring fact that the improvement is moderate, and the extended set of translations is far from detrimental to the performance. Note that the extended set includes, for example, Italian, which is not used in any of the topics. Third, our experiment with restricting our intention to pages in the language of the topic translation is clearly successfull. It leads to substantial improvement of the score.

We now zoom in on the effectiveness of the individual translations. Table 3 lists the results of the translated queries, both evaluated against the whole topic set, as well as against all topics targeting a page in the language at hand. We see the following. First, when looking at the restricted topic sets, effectiveness varies from total failure (Greek) to perfection (French). The score for the five frequent languages is reasonable compared to those of the mixed monolingual task. Hence, one may conclude that the automatic topic translations are effective. Second, when looking at all topics, the scores are generally unimpressive and mirroring the frequency with which a topic of the given language appears in the topic set. This comes as no surprise, given that the topic set covers eleven languages, and each of the topic translations will dominantly target only one of them. Third, the translated topics pick up relevant pages in languages other than the target language. In particular, the Italian topics do pick up a relevant page for 35 of the topics. Fourth, the single topic language runs are still much more effective than the combined multilingual runs in Table 2. This is a disappointing result, and clearly indicates that the straightforward run combination is ineffective. On a more positive note, however, the results for the individual translations strongly suggest that more sensible methods are possible.

5 Conclusions

This paper documents the University of Amsterdam's participation in the CLEF 2005 WebCLEF track. The EuroGOV collection used at WebCLEF is based on a crawl of governmental information from a range of sites. Such a collection of web data is much noisier than traditional collections of newswire and newspaper data originating from a single source. Moreover, the linguistic variety in the collection makes it harder to apply language-specific processing methods such as stemming algorithms. Hence, we simply indexed the collection by extracting the full text from the documents.

For the *mixed monolingual* task, we submitted a single, standard ad hoc retrieval run. Our main finding is that such a straightforward approach is relatively effective, that uses no web specific settings. Considering the fact that we are dealing with a stream of topics in eleven languages, and with an even greater number of languages in the collection, this sheds new light on the robustness of modern information retrieval techniques.

For the *multilingual* task, we experimented with different numbers of translations of the English queries, and with restricting the returned pages to the now-known language of the query at hand. Our experiments show beneficial effects for restricting the number of translation to those occurring frequently in the topic set, as well as for limiting query translations to return only pages in the language of the query. In general, however, the combined results for the multilingual task are unimpressive. The individual query translations, however, seem relatively successful in targeting their share of relevant pages. This casts considerable doubt on the effectiveness of standard combination methods for this particular task.

6 Acknowledgments

We want to thank Valentin Jijkoun for help with the Russian collection. 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] E.A. Fox and J.A. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215, 1994.

[3] ILPS. The ILPS extension of the Lucene search engine, 2005. `http://ilps.science.uva.nl/Resources/`.

[4] J. Kamps. Web-centric language models. In *Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005)*. ACM Press, New York NY, USA, 2005.

[5] J. Kamps, S. Fissaha Adafre, and M. de Rijke. Effective translation, tokenization and combination for cross-lingual retrieval. In C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, volume 3491 of *Lecture Notes in Computer Science*, pages 123–134. Springer Verlag, Heidelberg, 2005.

[6] J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjörnsson. Language-dependent and language-independent approaches to cross-lingual text retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, volume 3237 of *Lecture Notes in Computer Science*, pages 152–165. Springer, 2004.

[7] J.H. Lee. Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–188. ACM Press, New York NY, USA, 1995.

[8] Lucene. The Lucene search engine, 2005. `http://jakarta.apache.org/lucene/`.

[9] Worldlingo. Online translator, 2005. `http://www.worldlingo.com/`.