Combination Methods for Crosslingual Web Retrieval

Jaap Kamps[1,2], Maarten de Rijke[2], and Börkur Sigurbjörnsson[2]

[1] Archives and Information Science, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam
{kamps,mdr,borkur}@science.uva.nl

Abstract. We investigate a range of crosslingual web retrieval tasks using the test suite of the CLEF 2005 WebCLEF track, which features a stream of known-item topics in various languages. Our main findings are: (i) straightforward indexing and retrieval is effective for mixed monolingual web retrieval; (ii) standard machine translation methods are effective for bilingual web retrieval; but (iii) standard combination methods are ineffective for multilingual web retrieval; we analyze the failure and suggest an alternative Z-score normalization that leads to effective multilingual retrieval results.

1 Introduction

The web presents one of the greatest challenges for crosslingual information retrieval. Web data is much noisier than traditional collections of newswire and newspaper data originated from a single source. Also, the linguistic variety in the collection makes it harder to apply language-dependent processing methods such as stemming algorithms. Moreover, the size of the web only allows for methods that scale well, casting doubt on the practicality of language-independent methods such as character n-gramming.

We investigate a range of approaches to crosslingual web retrieval using the test suite of the CLEF 2005 WebCLEF track, featuring a stream of known-item topics in various languages. First, we focus on the *mixed monolingual* task. We aim to evaluate the robustness of modern information retrieval techniques, by applying standard ad hoc retrieval settings for a stream of monolingual topics in various languages. Second, we focus on a range of bilingual retrieval tasks using the English translations of the WebCLEF 2005 topics. Here, our aim is to evaluate the effectiveness of machine translation for known-item search in various languages. Third, we focus on the *multilingual* task, where there is a stream of English topics targeting pages in a range of languages. Here, we investigate whether the effectiveness of straightforward run combinations carries over to crosslingual web retrieval. Such methods have previously been used successfully at earlier CLEF monolingual and multilingual ad hoc retrieval tasks [6, 7].

This paper is structured as follows. In Section 2 we describe our retrieval system as well as the specific approaches to crosslingual web retrieval. Section 3 describes our mixed monolingual experiments. The next two sections discuss our multilingual experiments, focusing on translations to individual languages in Section 4, and on combinations for all languages in Section 5. Finally, in Section 6, we offer some conclusions regarding our crosslingual web retrieval efforts.

2 System Description

Our retrieval system is based on the Lucene engine with a number of home-grown extensions [3, 9].

For our ranking, we used the default similarity measure in Lucene [9], 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

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

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.

We used the WorldLingo machine translation [12] 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.

We combine the results for runs with different translations of the topics using both rank-based methods, in particular a straightforward round robin approach, as well as score-based methods, such as the unweighted CombSUM function of Fox and Shaw [2]. The score-based methods were applied after normalizing the similarity scores. First, we use min-max normalization, $s' = \frac{s-min}{max-min}$, with min (max) the minimal (maximal) score over all topics in the run. The min-max normalization was also used in [8]. Second, we use the Z-score normalization,

$s' = \frac{s-\mu_i}{\delta_i}$, with μ_i the mean retrieval status value and δ_i the standard deviation for topic i. A variant of Z-score normalization was used earlier in [11].

3 Mixed Monolingual Experiments

For the mixed monolingual task, we investigate the effectiveness of standard ad hoc retrieval settings for a stream of topics in various languages. We create a single run using 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). The resulting run was submitted to the WebCLEF 2005 mixed monolingual task.

Table 1 reports the results 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 task. Somewhat worrying is the success rate at rank 10, with no relevant page found for over 40% of the topics. Second, when breaking down the score over the two topic types, 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]. Third, when zooming in on the different topic languages, we see that we score reasonably well over all languages. The exception are the Greek topics; because of a technical problem, the Greek topics were processed as Russian and characters outside the expected character range where treated as noise and removed.

Table 1. Mixed Monolingual Task results by mean reciprocal rank and success at rank 1, 5 and 10.

	# Topics	MRR	S@1	S@5	S@10
All topics	547	.3497	.2523	.4589	.5576
Home pages	242	.2263	.1322	.3347	.4380
Named pages	305	.4476	.3475	.5574	.6525
Danish	30	.2288	.1667	.3000	.4333
Dutch	59	.5245	.4068	.6610	.7966
English	121	.3345	.2231	.4628	.5785
French	1	1.000	1.000	1.000	1.000
German	57	.3736	.2456	.5263	.6316
Greek	16	.0000	.0000	.0000	.0000
Hungarian	35	.3731	.2571	.5143	.5714
Icelandic	5	.4654	.4000	.6000	.6000
Portuguese	59	.1934	.1017	.3051	.3898
Russian	30	.3033	.2667	.3333	.4000
Spanish	134	.4091	.3134	.5000	.5970

Table 2. Bilingual results by mean reciprocal rank and success at rank 1, 5 and 10.

	Restricted to language					All 547 topics			
	# Topics	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Dutch	59	.2709	.2203	.3051	.3729	.0540	.0420	.0640	.0823
English	121	.3289	.2149	.4628	.5702	.0882	.0585	.1207	.1499
French	1	1.000	1.000	1.000	1.000	.0303	.0201	.0366	.0494
German	57	.2008	.1754	.1930	.2807	.0447	.0329	.0530	.0695
Greek	16	.0000	.0000	.0000	.0000	.0204	.0146	.0256	.0329
Italian	0	–	–	–	–	.0284	.0201	.0366	.0475
Portuguese	59	.1047	.0508	.1525	.1695	.0412	.0256	.0567	.0713
Russian	30	.0127	.0000	.0333	.0333	.0446	.0293	.0567	.0750
Spanish	134	.2272	.1791	.2687	.3582	.0809	.0603	.0969	.1316

4 Bilingual Experiments

Although there was no separate bilingual task at WebCLEF 2005, the multilingual topics can be used to evaluate the effectiveness of the individual translations, resulting in a whole range of bilingual retrieval experiments. All runs use the English version of the short topic statement in the ⟨translation language="EN"⟩ field of the WebCLEF 2005 topics. We generate the eight translations mentioned in Section 2 above. Note that the nine languages that we cover in total differ somewhat from the languages in the WebCLEF 2005 topic set. The topic set also has topics in Danish, Hungarian, and Icelandic. Furthermore, we have a translation of the English topics into Italian, whereas the topic set contains no topics in Italian.

Table 2 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 evaluated over 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.

5 Multilingual Experiments

We move on to the multilingual task, and investigate the effectiveness of combinations of the individual bilingual runs. We experimented along two dimensions. The first dimension is the number of topic languages:

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, and similar for German, we restrict pages to come from either `.de` or `.eu.int`.[3]

Combining the two dimensions naturally suggests four different sets of bilingual runs. These are combined using unweighted CombSUM using the min-max normalization. The resulting four runs were submitted to the WebCLEF 2005 multilingual task.

5.1 CombSUM with Min-Max Normalization

Table 3 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 to the smaller set of five language used frequently in the topic set. It

[3] Note that we mainly aim for precision here, we ignore domains such as `.be` (Belgium) and `.at` (Austria) where Dutch or German pages, respectively, are abundant.

Table 3. Multilingual Task results by mean reciprocal rank and success at rank 1, 5 and 10.

Number of	All topics				Home pages				Named pages			
Languages	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Nine	.0092	.0055	.0073	.0165	.0072	.0041	.0083	.0124	.0109	.0066	.0066	.0197
Nine, restr.	.0157	.0091	.0201	.0219	.0157	.0124	.0165	.0165	.0158	.0066	.0230	.0262
Five	.0109	.0055	.0091	.0165	.0084	.0041	.0083	.0124	.0129	.0066	.0098	.0197
Five, restr.	.0166	.0091	.0201	.0238	.0163	.0124	.0165	.0207	.0168	.0066	.0230	.0262

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 successful. Fourth, the single topic language runs in Table 2 are much more effective than the combined multilingual runs, even when evaluated against the total topic set. 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.

5.2 Rank-Based Combination: Round Robin

We saw above that the quality of our multilingual run combinations poorly reflects the quality of the individual bilingual runs. If we look, again, at the individual language results in Table 2, we see that already the success rate at rank 1 is higher than the mean reciprocal rank for the combination runs in Table 3. Hence a combination method that preserves the order of the individual runs will be more effective. We apply a straightforward rank-based combination method, round robin, in which the individual bilingual runs are interleaved. Specifically, we only return the same document once per topic, ordering languages alphabetically by their two character iso-codes, resulting in German, Greek, English, Spanish, French, Italian, Dutch, Portuguese, and Russian. Hence, the success rate at rank 1 will be identical to that of the bilingual English to German run evaluated over all topics.

Table 4 shows the results of applying the round-robin combination. Indeed, the rank-based round robin is much more effective than the results for CombSUM in Table 3. In fact, the combination of the five frequent languages in the topic

Table 4. Round robin combination results by mean reciprocal rank and success at rank 1, 5 and 10.

Number of	All topics				Home pages				Named pages			
Languages	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Nine	.0763	.0329	.1335	.1883	.0551	.0248	.0992	.1322	.0930	.0393	.1607	.2328
Nine, restr.	.0535	.0238	.0951	.1298	.0458	.0248	.0785	.0992	.0597	.0230	.1082	.1541
Five	.0944	.0329	.1700	.2194	.0704	.0248	.1198	.1570	.1135	.0393	.2098	.2689
Five, restr.	.0687	.0238	.1243	.1645	.0575	.0248	.0950	.1157	.0776	.0230	.1475	.2033

set outperforms the best individual language run. The restriction to domains corresponding to the languages of the translations is now detrimental to the performance.

The effectiveness of rank-based round-robin combinations can be attributed to the fact that highly ranked documents in the combination are also highly ranked by some of the bilingual runs. The earlier applied combination method, CombSUM, tends to favor documents receiving scores in several of the bilingual runs. The results show that this is an undesirable behavior for the task at hand. This may be explained by the fact that the task is known-item retrieval, and this single, relevant page is generally retrieved by at most one of the bilingual runs.

5.3 CombSUM with Z-score Normalization

As we saw in Section 4, each of the bilingual runs is also capable of retrieving relevant documents in another language. That is, we may expect there to be some middle ground in which the combination does largely respect the rankings of pages in the individual bilingual runs, but at the same time does reward pages returned by several runs.

We focus on score normalization. Earlier we used the Min-Max normalization, which results in a simple linear transformation of the original scores into values between 0 and 1. We want to come up with a score normalization that gives a relatively higher weight to top ranking documents. A standard method for score normalization is the Z-score: values are normalized to the number of standard deviations that they are higher (or lower) than the mean score. At first sight, this only makes sense for normally distributed values, for example because documents not retrieved will have the mean score of the retrieved documents. On closer inspection, this will yield exactly the properties we desire. Since the similarity scores will be very skewed, with a long tail approaching zero, the mean and standard deviation will be very small. Hence, the top scoring documents will receive relatively high scores, but the score is steeply declining.

Table 5 shows the results of applying CombSUM combination to relevance scores being normalized with the Z-score value. We see that the Z-score normalization is far more effective than the Min-Max normalization in Table 3. It also improves over the rank-based round robin combination in Table 4.

Table 5. Combination results based on Z-score normalization by mean reciprocal rank and success at rank 1, 5, and 10.

Number of	All topics				Home pages				Named pages			
Languages	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Nine	.0914	.0494	.1298	.1846	.0659	.0372	.0992	.1364	.1186	.0689	.1705	.2197
Nine, restr.	.0638	.0347	.0859	.1371	.0467	.0289	.0579	.0868	.0674	.0295	.1016	.1705
Five	.1096	.0640	.1609	.2029	.0770	.0413	.1157	.1529	.1352	.0820	.1967	.2590
Five, restr.	.0841	.0475	.1261	.1572	.0649	.0413	.0909	.1074	.0947	.0492	.1475	.2000

Table 6. Combination results based on domain information by mean reciprocal rank and success at rank 1, 5 and 10. Top half: using nine languages. Bottom half: using five languages.

Combination	All topics				Home pages				Named pages			
Method	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10	MRR	S@1	S@5	S@10
Min-Max	.0530	.0165	.0731	.1353	.0474	.0165	.0744	.1074	.0574	.0164	.0721	.1574
Round robin	.1382	.0676	.2267	.2834	.1038	.0579	.1612	.2107	.1654	.0754	.2787	.3410
Z-score	.1605	.1042	.2303	.2797	.1104	.0661	.1653	.1983	.2001	.1344	.2820	.3443
Min-Max	.0612	.0219	.0823	.1609	.0523	.0207	.0785	.1281	.0683	.0230	.0852	.1869
Round robin	.1472	.0695	.2358	.2907	.1098	.0620	.1653	.2107	.1769	.0754	.2918	.3541
Z-score	.1676	.1079	.2468	.2852	.1193	.0785	.1777	.2025	.2060	.1311	.3016	.3508

5.4 Exploiting Additional Information: Target Domain

What if our user provides us with further information, such as the language or the domain of the desired page? We investigate this scenario by using some of the additional metadata fields. In particular, we use the additional information about the domain of the target page in the ⟨domain domain=“*top-level domain*” /⟩ field. Table 6 shows the results of applying (i) CombSUM combination of the min-max normalization; (ii) round robin; and (iii) CombSUM combination of the Z-score normalization. We see that information about the domain of the desired page can effectively be exploited by all three combination methods. The relative effectiveness of the combination methods mimic the earlier combination scores closely, with the CombSUM method with Z-score normalization the most effective.

6 Conclusions

The EuroGOV collection used at the CLEF 2005 WebCLEF Track 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 our crosslingual web retrieval retrieval experiments we use a stream of known-item topics in various languages. For the *mixed monolingual* task, our main finding is that such a straightforward approach is relatively effective, even without specific web 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 bilingual retrieval, we experimented with machine translations of the English queries. The individual query translations are relatively successful in targeting their share of relevant pages. For the *multilingual* task, we experimented with various combination methods. A standard CombSUM combination using Min-Max normalization is ineffective.

This result deviates from earlier experiences with combination methods for corpora in various languages [5], or with known-item retrieval on an English web corpus [10]. We show that rank-based combination methods fare much better, and propose an alternative Z-score normalization method that turns out to be effective for crosslingual web retrieval.

Acknowledgments We want to thank Valentin Jijkoun for help with the Russian collection. This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 365-20-005, 612.000.106, 612.000.207, 612.013.001, 612.-066.302, 612.069.006, 640.001.501, and 640.002.501

Bibliography

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

[2] E. Fox and J. 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 *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 307–308, 2005.

[5] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for European languages. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, pages 1073–1077, 2004.

[6] J. Kamps, S. Fissaha Adafre, and M. de Rijke. Effective translation, tokenization and combination for cross-lingual retrieval. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, pages 123–134, 2005.

[7] J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjörnsson. Language-dependent and language-independent approaches to cross-lingual text retrieval. In *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, pages 152–165, 2004.

[8] J. 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, 1995.

[9] Lucene. The Lucene search engine, 2005. http://lucene.apache.org/.

[10] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 143–150, 2003.

[11] J. Savoy. Report on CLEF-2003 multilingual tracks. In *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, pages 64–73, 2004.

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