The University of Amsterdam at WebCLEF 2006

Krisztian Balog   Maarten de Rijke
ISLA, University of Amsterdam
kbalog,mdr@science.uva.nl

Abstract

Our aim for our participation in WebCLEF 2006 was to investigate the robustness of information retrieval techniques to crosslingual retrieval, such as compact document representations, and query reformulation techniques. Our focus was on the mixed monolingual task. Apart from the proper preprocessing and transformation of various encodings, we did not apply any language-specific techniques. Instead, the target domain meta field was used in some of our runs. A standard combSUM combination using Min-Max normalization was used to combine runs, based on a separate content and title indexes of documents. We found that the combination is effective only for the human generated topics. Query reformulation techniques can be used to improve retrieval performance, as witnessed by our best scoring configuration, however these techniques are not yet beneficial to all different kinds of topics.

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 Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Web retrieval, Multilingual retrieval

1 Introduction

The world wide web is a natural setting for cross-lingual information retrieval, since the web content is essentially multilingual. On the other hand, web data is much noisier than traditional collections, eg. newswire or newspaper data, which originated from a single source. Also, the linguistic variety in the collection makes it harder to apply language-dependent processing methods, eg. stemming algorithms. Moreover, the size of the web only allows for methods that scale well.

We investigate a range of approaches to crosslingual web retrieval using the test suite of the CLEF 2006 WebCLEF track, featuring a stream of known-item topics in various languages. The topics are a mixture of human generated (manually) and automatically generated topics. Our focus is on the mixed monolingual task. Our aim for our participation in WebCLEF 2006 was to investigate the robustness of information retrieval techniques, such as compact document representations (titles or incoming anchor-texts), and query reformulation techniques.

The remainder of the paper is organized as follows. In Section 2 we describe our retrieval system as well as the specific approaches we applied. In Section 3 we describe the runs that we submitted, while the results of those runs are detailed in Section 4. We conclude in Section 5.

2 System Description

Our retrieval system is based on the Lucene engine [4].

For our ranking, we used the default similarity measure of Lucene, 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 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 á to a, ö to o, æ to ae, î to i, etc. The only language specific processing we did was a transformation of the multiple Russian and Greek encodings into an ASCII transliteration.

We extracted the full text from the documents, together with the title and anchor fields, and created three separate indexes:

- **Content**: Index of the full document text.
- **Title**: Index of all <title> fields.
- **Anchors**: Index of all incoming anchor-texts.

We performed three base runs using the separate indexes. We evaluated the base runs using the WebCLEF 2005 topics, and decided to use only the *content* and *title* indexes.

2.1 Run combinations

We experimented with the combination of *content* and *title* runs, using standard combination methods as introduced by Fox and Shaw [1]: `combMAX` (take the maximal similarity score of the individual runs); `combMIN` (take the minimal similarity score of the individual runs); `combSUM` (take the sum of the similarity scores of the individual runs); `combANZ` (take the sum of the similarity scores of the individual runs, and divide by the number of non-zero entries); `combMNZ` (take the sum of the similarity scores of the individual runs, and multiply by the number of non-zero entries); and `combMED` (take the median similarity score of the individual runs).

Fox and Shaw [1] found `combSUM` to be the best performing combination method. Kamps and de Rijke [2] conducted extensive experiments with the Fox and Shaw combination rules for nine european languages, and demonstrated that combination can lead into significant improvements.

Moreover, they proved that the effectiveness of combining retrieval strategies differs between English and other European languages. In Kamps and de Rijke [2], `combSUM` emerges as the best combination rule, confirming Fox and Shaw's findings.

Similarity score distributions may differ radically across runs. We apply the combination methods to normalized similarity scores. That is, we follow Lee [3] and normalize retrieval scores into a $[0, 1]$ range, using the minimal and maximal similarity scores (min-max normalization):

$$s' = \frac{s - min}{max - min}, \tag{1}$$

where s denotes the original retrieval score, and min (max) is the minimal (maximal) score over all topics in the run.

For the WebCLEF 2005 topics the best performance was achieved using the `combSUM` combination rule, which is in conjunction with the findings in [1, 2], therefore we used that method for our WebCLEF 2006 submissions.

2.2 Query reformulation

In addition to our run combination experiments, we conducted experiments to measure the effect of phrase and query operations. We tested query-rewrite heuristics using phrases and word n-grams.

Phrases In this straightforward mechanism we simply add the topic terms as a phrase to the topic. For example, for the topic WC0601907, with title "diana memorial fountain", the query becomes: *diana memorial fountain "diana memorial fountain"*. Our intuition is that rewarding documents that contain the whole topic as a phrase, not only the individual terms, would be beneficial to retrieval performance.

N-grams In our approach every word n-gram from the query is added to the query as a phrase with weight n. This means that longer phrases get bigger boost. Using the previous topic as an example, the query becomes: *diana memorial fountain "diana memorial"*^{2} *"memorial fountain"*^{2} *"diana memorial fountain"*^{3}, where the number in the upper index denotes the weight attached to the phrase (the weight of the individual terms is 1).

3 Runs

We submitted five runs to the mixed monolingual task:

Baseline Base run using the *content* only index.

Comb Combination of the *content* and *title* runs, using the `CombSUM` method.

CombMeta The `Comb` run, using the *target domain* meta field. We restrict our search to the corresponding domain.

CombPhrase The `CombMeta` run, using the *Phrase* query reformulation technique.

CombNboost The `CombMeta` run, using the *N-grams* query reformulation technique.

4 Results

Table 1 lists our results in terms of mean reciprocal rank. Runs are listed along the left-hand side, while the labels indicate either all topics (*all*) or various subsets: automatically generated (*auto*)—further subdivided into automatically generated using the unigram generator (*auto-u*) and automatically generated using the bigram generator (*auto-b*)—and manual (*manual*), which is further subdivided into *new* manual topics and *old* manual topics.

Significance testing was done using the two-tailed Wilcoxon Matched-pair Signed-Ranks Test, where we look for improvements at a significance level of 0.05 (^{1}), 0.001 (^{2}), and 0.0001 (^{3}). Signficant differences noted in Table 1 are with respect to the `Baseline` run.

Table 1: Submission results (Mean Reciprocal Rank)

runID	all	auto	auto-u	auto-b	manual	man-n	man-o
`Baseline`	0.1694	0.1253	0.1397	0.1110	0.3934	0.4787	0.3391
`Comb`	0.1685^1	0.1208^3	0.1394^3	0.1021	0.4112	0.4952	0.3578
`CombMeta`	0.1947^3	0.1505^3	$\mathbf{0.1670}^3$	0.1341^3	0.4188^3	0.5108^1	$\mathbf{0.3603}^1$
`CombPhrase`	$\mathbf{0.2001}^3$	0.1570^3	0.1639^3	0.1500^3	**0.4190**	**0.5138**	0.3587
`CombNboost`	0.1954^3	$\mathbf{0.1586}^3$	0.1595^3	$\mathbf{0.1576}^3$	0.3826	0.4891	0.3148

Combination of the *content* and *title* runs (`Comb`) increased performance only for the manual topics. The use of the title tag does not help when the topics are automatically generated. Instead of employing a language detection method, we simply used the target domain meta field. The `CombMeta` run improved the retrieval performance significantly for all subsets of topics. Our query reformulation techniques show mixed, but promising results. The best overall score was achieved when the topic, as a phrase, was added to the query (`CombPhrase`). The comparison of `CombPhrase` vs `CombMeta` reveals that they achieved similar scores for all subsets of topics, except for the automatic topics using the bigram generator, where the query reformulation technique was clearly beneficial. The n-gram query reformulation technique (`CombNboost`) further improved the results of the *auto-b* topics, but hurt accuracy on all other subsets, especially on the manual topics. The `CombPhrase` run demonstrates that even a very simple query reformulation technique can be used to improve retrieval scores. However, we need to further investigate how to automatically detect whether it is beneficial to use such techniques (and if yes, which technique to apply) for a given a topic.

Comparing the various subsets of topics, we see that the automatic topics have proven to be more difficult than the manual ones. Also, the new manual topics seem to be more appropriate for known-item search than the old manual topics. There is a clear ranking among the various subsets of topics, and this ranking is independent from the applied methods: $man - n \gg man - o \gg auto - u \gg auto - b$.

5 Conclusions

Our aim for our participation in WebCLEF 2006 was to investigate the robustness of information retrieval techniques to crosslingual web retrieval. The only language-specific processing we applied was the transformation of various encodings into an ASCII transliteration. We did not apply any stemming nor did we use a stopword list. We indexed the collection by extracting the full text and the title fields from the documents. A standard combSUM combination using Min-Max normalization was used to combine the runs based on the content and title indexes. We found that the combination is effective only for the human generated topics, using the title field did not improve performance when the topics are automatically generated. Significant improvements (+15% MRR) were achieved by using the target domain meta field. We also investigated the effect of query reformulation techniques. We found, that even very simple methods can improve retrieval performance, however these techniques are not yet beneficial to retrieval for all subsets of topics. Although it may be too early to talk about a solved problem, effective and robust web retrieval techniques seem to carry over to the mixed monolingual setting.

6 Acknowledgments

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. 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, and 640.002.501.

References

[1] 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.

[2] 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.

[3] J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 180–188, New York, NY, USA, 1995. ACM Press. ISBN 0-89791-714-6. doi: http://doi.acm.org/10.1145/215206.215358.

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