University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming.

Craig Macdonald Vassilis Plachouras Ben He Christina Lioma
Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ
United Kingdom
{craigm,vassilis,ben,xristina,ounis}@dcs.gla.ac.uk

Abstract

We investigated the use of appropriate web retrieval techniques and language specific stemming techniques in a multi-lingual environment.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Multi-lingual Web IR; European languages stemming

1 Introduction

The aim of the University of Glasgow's participation in the mono-lingual task of WebCLEF 2005 was to test the application of Web Information Retrieval (IR) techniques on a collection with many different languages. We based our CLEF 2005 participation on our IR platform, Terrier [6].

Our experiments were focused on the combination of evidence in a Web IR setting as well as the application of an appropriate stemming in a multi-lingual environment. The outline of this paper is as follows: Section 2 details the stemming techniques we applied in this work, while Section 3 details our experimental setup. We describe the official runs submitted in Section 4, and finally analyse our results in Section 5. Section 6 provides concluding remarks.

2 Stemming

Our main research hypothesis for the participation to WebCLEF 2005 was that being able to apply the correct stemmer to a document and a topic would increase the performance of the search engine. To test this hypothesis, we created three indices of the EuroGOV collection. In the first

index, the collection was indexed without applying any stemming techniques. In the second index, the collection was indexed by applying Porter's English stemmer to all documents, regardless of their domain and language. In the last index, we stemmed each document taking into account the language of each document. The language of the document was primarily determined by the language identification data provided by the WebCLEF organisers. However, as the language identification data is not precise - often giving multiple language choices for documents - we chose to supplement this with a few heuristics. For each document, we examined the suggested languages provided by the WebCLEF organisers, and look for evidence to support these languages in the URL of the document, the metadata of the document, and in a list of "normal languages" for each domain.

Having identified one language for each document, we would apply the stemmer for that language to the terms from that document. We mainly used the Snowball [2] stemmers to stem the documents. The exceptions to this were: English where we used Porter's English stemmer; Icelandic where we used the Danish Snowball Stemmer; Hungarian where we used Hunstem [1] and Greek where we did not apply any stemmer.

3 Terrier Setup

To support multi-lingual retrieval, it is essential that the IR system accurately and uniquely represent each term in the corpus. To meet this requirement, we used a version of Terrier that supports UTF-8 encoding, ensuring that we had a robust representation of the collection.

During the parsing of the collection, we used heuristics, based on the HTTP headers, the META tags and the language identifier to determine the correct content encoding for each document. Once the correct encoding for each document was determined the collection was parsed, each term being read and converted to UTF-8 representation.

As described above, we applied several stemming combinations to index the EuroGOV collection. In each index, all terms including stop words were indexed, and positional information was kept for each term in the collection. Three indices were built - one which applied no stemming during indexing, one which applied only the porter stemmer, and one which applied the stemmer deemed appropriate for each document.

We used different sources of evidence (or fields) available in a web corpus: the body of the document, the title of the document and the anchor text information. We used a new technique for combining sources of evidence during retrieval, that we call per-field normalisation, which is an alternative to [9]. The used weighting model is a per-field derivative of the following PL2 DFR model:

$$score(d,Q) = \sum_{t \in Q} \frac{qtfn}{tfn+1}\big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)\big) \qquad (1)$$

where $score(d,Q)$ is the relevance score of a document d for a query Q. t is a query term in Q. λ is the mean and variance of a Poisson distribution. $qtfn$ is the normalised query term frequency. It is given by $qtfn = \frac{qtf}{qtf_{max}}$, where qtf is the query term frequency and qtf_{max} is the maximum query term frequency among the query terms. The normalised term frequency tfn is given by the so-called *Normalisation 2*:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l}), (c > 0) \qquad (2)$$

where l is the document length and avg_l is the average document length in the whole collection. tf is the original term frequency. c is the free parameter of the normalisation method. The c parameters values were set automatically using a technique that extends our previous work [4] to take fields into account.

In our previous work [7], we found that taking the length of the URL of a document into account is particularly effective in homepage finding tasks. Following this work, we used this evidence.

4 Runs

We submitted 5 runs to the mono-lingual task of WebCLEF 2005, four of which used topic metadata of some form. For all metadata runs, we used the domain topic metadata to limit the URL domain of the returned results. For example, if the topic stated `<domain domain="eu.int"/>`, only results with URLs in the eu.int domain were returned. The official runs we submitted are detailed below:

- `uogSelStem`: This run did not use any metadata. Instead, we used the language identifier (textcat [3, 8]) to identify the language of each topic. The topic was then stemmed using the appropriate stemmer for that language. We used the index with all stemmers. This run tested the accuracy of the language identifier in determining which stemmer to apply to each topic.
- `uogNoStemNLP`: This run used only the domain metadata described above. No stemming was applied to the topics. Additionally, we used a language processing technique to deal with acronyms. No document stemming was used. This run tested the retrieval effectiveness of not applying stemming in this multi-lingual Web IR setting.
- `uogPorStem`: This run used only the domain metadata described above. Porter's English stemmer was applied to all topics and the index with Porter's stemmer was used. This run tested the retrieval effectiveness of applying the Porter's stemmer to all languages in the EuroGOV collection.
- `uogAllStem`: This run used both the domain metadata described above and the topic language metadata, which allowed the use of the correct stemmer to stem the topic. We used the index with all stemmers. This run tested the hypothesis that applying the correct stemmer to both documents and topics would improve results overall.
- `uogAllStemNP`: This run is identical to `uogAllStem` except that term order in the topics was presumed to be important. We applied a strategy where query terms are appropriately weighted to reflect the order of terms in a query. The underlying hypothesis is that in web search, the user will typically enter the most important keywords first, then apply additional terms to narrow the focus of the search.

5 Results & Analysis

Table 1 details the mono-lingual baseline task MRR achieved for each run. We have also broken the MRR down into the component languages of the topics, and the home-page (HP) and named-page (NP) topics. From Table 1, we can draw the following conclusions:

- From initial inspection of the overall performance of the the runs, it would appear that the run which did not apply any stemmers performed best overall. Indeed, the `uogNoStemNLP` run gives the best results over the baseline tasks, closely followed by the `uogPorStem` run.
- However, the runs with the correct stemming applied (`uogAllStem` & `uogAllStemNP`) have comparable results to the `uogNoStemNLP` & `uogPorStem` runs, even though the performance of the Hungarian queries is considerably reduced. In particular, the last line of Table 1 shows the MRR of all runs with all Hungarian topics removed. This shows that the stemming makes minor difference - the `uogAllStem` & `uogAllStemNP` runs achieve approximately the same MRR as the `uogPorStem` run, and are very comparable to the `uogNoStemNLP` run.
- It would appear that the Porter's English stemmer is more effective than either no stemming or the appropriate Snowball stemmer for Dutch and Russian. English and Portuguese topics give the best performance without any stemming applied.

Topic Set	uogSelStem	uogNoStemNLP	uogPorStem	uogAllStem	uogAllStemNP
All	0.4683	**0.5135**	0.5107	0.4827	0.4828
DA	0.5168	0.5246	0.5098	**0.5857**	0.5829
DE	0.4469	0.4414	0.4567	**0.4780**	0.4689
EL	0.2047	0.3704	0.3659	0.3586	**0.4003**
EN	0.4988	**0.5578**	0.5240	0.5188	0.5239
ES	0.4198	0.4571	0.4635	0.4602	**0.4647**
FR	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
HU	0.2713	**0.5422**	**0.5422**	0.1142	0.1003
IS	**0.3400**	0.3222	0.3222	**0.3400**	**0.3400**
NL	0.6362	0.6226	**0.6551**	0.6444	0.6447
PT	0.5262	**0.5565**	0.5336	0.5048	0.5028
RU	0.4838	0.4724	**0.4975**	0.4838	0.4625
NP only	0.4803	**0.5353**	0.5232	0.4952	0.4956
HP only	0.4531	0.4862	**0.4949**	0.4669	0.4666
All minus HU	0.4818	**0.5116**	0.5085	0.5078	0.5089

Table 1: Mean Reciprocal Rank (MRR) of the submitted runs to the mono-lingual task.

- The particularly poor performance when applying the correct stemmer to the Hungarian topics (the `uogAllStem` & `uogAllStemNP` runs) implies that the Hungarian stemmer is not effective. However, when the language identifier classifies the Hungarian topics (as in `uogSelStem`), performance is improves slightly (0.2713 vs. 0.1142). Hence we hypothesise that the classification of the Hungarian documents in the collection was often incorrect, and that the classifier made the same mistakes with the topics, leading to the improved performance.

- By comparing the `uogAllStem` and the `uogSelStem` runs, we can see identify that the language identifier is deficient at recognising the language of topics in many languages, and poor at Greek and Danish in particular. For Hungarian topics, when language identifier mis-classifies the language of the topics, performance is actually improved. In the final version of this paper, we will assess the accuracy of the language classification for the topics, to detect correlations between correct classification and increased performance.

- Query term ordering overall showed little retrieval performance difference when applied. However, its application was particularly effective to the Greek topics (0.3586 to 0.4003), but showed very little positive or negative change for most languages.

6 Conclusions

We performed experimentation around the correct application of stemmers in a multi-lingual setting. We found that generally applying the correct stemmer for the language of the document and topic worked in most cases, however if language classification of the documents was incorrect, the retrieval performance could be harmed. The bare-system approach of applying no stemming at all, is a very safe and stable option where the results will not be very different from that produced by the best approach for that language.

While it is clear that stemming with respect to a language can assist in retrieval performance, determining the language of a topic or document is still an active research problem for reliable application in a multi-lingual document collection.

For this paper, we investigated the average topic length - in particular for the German, Spanish and English topic sets, and found these to be 3.3, 6.3 terms and 5.7 terms respectively. In contrast, a recent study by Jansen & Spink [5] found the an average length of 1.9, 2.6 and 5.0 terms for

German, Spanish and English queries, suggesting that the topics used in WebCLEF 2005 were not representative of real European user queries on a multi-lingual collection. For a future WebCLEF participation, we would like to see queries gathered from commercial European search engines, as these are definitive of real user needs.

References

[1] Hunspell & Hunstem: Hungarian version of Ispell & Hungarian stemmer, http://magyarispell.sourceforge.net/.

[2] Snowball stemmers, http://snowball.tartarus.org/.

[3] W.B. Cavnar and J.M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.

[4] B. He and I. Ounis. A study of the Dirichlet Priors for term frequency normalisation. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–471, New York, NY, USA, 2005. ACM Press.

[5] B.J. Jansen and A. Spink. An analysis of Web searching by European AlltheWeb.com users. *Inf. Process. Manage.*, 41(2):361–381, 2005.

[6] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005. URL: http://ir.dcs.gla.ac.uk/terrier/

[7] V. Plachouras, B. He, and I. Ounis. University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier. In *Notebook of 13thText REtrieval Conference (TREC2004)*, NIST, MD. USA, October 2004.

[8] G. van Noord. Textcat language guesser, http://odur.let.rug.nl/ vannoord/textcat/.

[9] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Notebook of 13thText REtrieval Conference (TREC2004)*, NIST, MD. USA, October 2004.