Web Retrieval Experiments with the EuroGOV Corpus at the University of Hildesheim

Niels Jensen, René Hackl, Thomas Mandl, Robert Strötgen

Information Science, University of Hildesheim,
Marienburger Platz 22
D-31141 Hildesheim, Germany

mandl@uni-hildesheim.de

Abstract

In the CLEF 2005 initiative, multlingual web retrieval was integrated as a task for the first time. This paper describes experiments based on one multilingual index carried out at the University of Hildesheim. Several indexing strategies based on a multi-lingual index have been tested with the EuroGOV corpus. Boosting topic fields with higher weight led to best results during post submission runs. The experiments also led to experiences in working with large test collections and the challenges associated with them.

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

Web Retrieval, Multilingual Information Retrieval, N-gram Indexing, Evaluation

1 Introduction

Web search engines has become a part of every day life for many people. The development of information retrieval systems for the web is faced with many challenges (Arasu et al. 2001). Systems give different answers to these challenges and it is difficult to judge the effect of decisions during the design of search enigne. As a consequence, there is a great need for evaluation in web retrieval (Hawking 2000). The web is also a natural source for multilingual documents.

Within the Cross Language Evaluation Forum (CLEF) the web track has been created (Sigurbjörnsson et al. 2005b). A large multilingual corpus has been collected and distributed (Sigurbjörnsson et al. 2005a). In our first participation, we intended to tune our system to the challanges of a large web corpus. For the experiments, language resources in all languages were not available from ad-hoc retrieval. As a consequence, we considered n-gram indexing for the web retrieval task (McNamee & Mayfield 2004).

2 Data Pre-Processing

Since the files of the EuroGOV corpus were not released in well formed XML, substantial effort for data pre-processing was necessary. A corpus in well formed XML would allow us to use the System implemented during the CLEF 2004 campaign for multilingual ad-hoc tasks (Hackl et al. 2005). The two main items in the EuroGOV files that needed replacing were predeclared entities. This step was required for ampersand with the associated entity reference in the URL fields of the individual documents and all nested CDATA tags. The first attempt to reformat the files has been carried out by a Perl-script. At the first view, it seemed that the Perl-scirpt would work perfectly for our needs. Unfortunately, we realized that during the process of indexing the corpus, the XML

parser would frequently report “parser exceptions” that we traced back to the fact that the XML files still contained a couple of not adjusted predeclared entities. Having this in mind, a Java program was developed that worked through the whole corpus perfectly. It seems that Perl is not able to process EuroGOV files bigger than 250 MB since we successfully tested the Perl-script with the small-size files (22 MB, 59 MB & 220 MB) of the corpus.

3 Submitted Retrieval Experiments with EuroGOV

As mentioned in the introduction, one multilingual index was created. In order to generate a slim index we assembled a multilingual stopwordlist. The bases for this list were the stopwordlists supplied by the University of Neuchatel[1] and a list developed specifically for the Czech language (Hofman Miquel 2005). All lists were combined and revised into one file. This multilingual stopwordlist covers twelve languages and was used for the indexing process of the corpus.

For our retrieval experiments, we created three different multilingual indexes. Two were created with the Lucene StandardAnalyzer[2], which does not implement any linguistic processing apart from word segmentation. The first index covered the whole corpus whereas the second index cut off the indexing process after a maximum of 200 characters for each individual document. Due to this approach, the sizes of the indexes varies from 5 GB to 700 MB.

The third index was created with a NGram Analyzer also applied to multilingual ad-hoc retrieval before (Hackl et al. 2005). Because of performance and time restrictions the trigram approach was only applied to the title field of the individual documents in the corpus files. As a result the size of the index is down to 300 MB which led to a very quick and stable performance at retrieval time. These three indexes are the foundation for our experiments. As a main retrieval engine, we used Lucene 1.4[3]. Some of the basic code for retrieval and n-gram analysis was adopted from previous CLEF ad-hoc experiments (Hackl et al. 2005). Six different baseline runs were submitted. We did not use any of the metadata that was supplied by the topics due to time and resource constraints. Our monolingual queries were created with the title field of the topic whereas the multilingual queries were based on the monolingual title field and the translation language English field. Both types of queries were sent to one multilingual index. Results are shown in table 1.

Table 1. WebCLEF 2005 results University of Hildesheim

	UHi3TiMo	UHi3TiMu	UHiScoMo	UHiScoMu	UHiSMo	UHiSMu
Mean reciprocal rank	0.0373	0.0274	0.1301	0.1147	0.1603	0.137
Average success at 1	0.0219	0.0146	0.1024	0.0932	0.1261	0.1097
Average success at 5	0.0512	0.0402	0.1627	0.1353	0.2011	0.1627
Average success at 10	0.064	0.0494	0.1883	0.1609	0.2194	0.1927
Average success at 20	0.075	0.064	0.2322	0.192	0.2523	0.2249
Average success at 50	0.1024	0.0878	0.2505	0.2157	0.287	0.2578

Looking at the results of the submitted runs it becomes clear that the trigram index did not confirm the expectations. On average, the monolingual runs differ from the multilingual runs by about 0.0162 MRR points. Having those results in mind the method of indexing the corpus with the Lucene StandardAnalyzer turned out to be more effective than the trigram strategy. The post experiments will illustrate this effect more clearly.

3 Post Submission Experiments with EuroGOV

For our post experiments we decided to generate another trigram index covering the whole corpus. The purpose of this experiment was to confirm the results from the official runs or to improve them by providing a better or more a complete index respectively. We also wanted to see if through boosting of the individual query fields (title & translation language English) the difference between the mono- and multilingual runs could be

[1] Stopwordlists: http://www.unine.ch/Info/clef/ verified August 11th 2005
[2] Lucene StandardAnalyzer: http://lucene.apache.org verified on August 11th 2005
[3] Lucene: http://lucene.apache.org verified August 11th 2005

compensated or even improved. As table 2 shows quite obviously, even a more complete index was not able to improve the MRR of the trigram runs. The results declined by approx. 50 %.

Table 2. Results of the trigram index run

	UHi3Mo	UHi3Mu
Mean reciprocal rank	0.0169	0.0099
Average success at 1	0.0091	0.0037
Average success at 5	0.0238	0.0183
Average success at 10	0.0366	0.0238
Average success at 20	0.042	0.0311
Average success at 50	0.0548	0.0402

In the second part of our post experiments we took the four indexes we had generated, and modified the weights of the query fields. The ratio for the two query fields were 10 to 1 and vice versa. The results that are shown in table 3 and 4 show that by boosting the title field of the query the results improve by 0.0144 MRR points on average. Applying this procedure, the performance of the multilingual run based on the StandardAnalyzer Index results in higher MRR values. The boosted multilingual run has a better result than any monolingual run and is the best run of all our experiments.

Table 3. Translation language English field Boost 10 to 1

	UHi3MuBo110	UHi3TiMuBo110	UHiScoMuBo110	UHiSMuBo110
Mean reciprocal rank	0.0063	0.0139	0.0677	0.0811
Average success at 1	0.0018	0.0091	0.053	0.0658
Average success at 5	0.0073	0.0165	0.0786	0.0987
Average success at 10	0.0146	0.0238	0.1079	0.1133
Average success at 20	0.0201	0.0293	0.1207	0.1316
Average success at 50	0.0402	0.0512	0.128	0.1444

Table 4. Title field Boost 10 to 1

	UHi3MuBo101	UHi3TiMuBo101	UHiScoMuBo101	UHiSMuBo101
Mean reciprocal rank	0.0172	0.0379	0.1307	0.1608
Average success at 1	0.0091	0.0219	0.1042	0.1298
Average success at 5	0.0256	0.053	0.1609	0.1974
Average success at 10	0.0329	0.0622	0.1883	0.2176
Average success at 20	0.0439	0.075	0.2285	0.245
Average success at 50	0.053	0.1042	0.2486	0.2724

4 Conclusion and Outlook

For the first web track at CLEF we intended to tune our system to be able to cope with a large amount of data. We suceeded in returning valid results for several runs.

In future experiments, we intend to step beyond the baseline runs and try to involve the metadata that is being provided by the WebCLEF topics. We also want to include advanced quality measures into consideration. Link based quality measures seem to be integral part of commercial search engines. They have been evaluated at the web track at TREC (Hawking 2000). Advanced quality measures take more features into account, especially information and design aspects (Mandl 2005).

References

Arasu, Arvind; Cho, Junghoo; Garcia-Molina, Hector; Paepcke, Andreas; Raghavan, Sriram (2001): *Searching the Web*. In: ACM Transactions on Internet Technology 1 (1) pp. 2-43.

Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): *Mono- and Cross-lingual Retrieval Experiments at the University of Hildesheim.* In: Peters, Carol; Clough, Paul; Gonzalo, Julio; Kluck, Michael; Jones, Gareth; Magnini, Bernard (eds): Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [Lecture Notes in Computer Science 3491] pp. 165-169.

Hawking, David (2000): *Overview of the TREC-9 Web Track*. In: The Ninth Text Retrieval Conference (TREC-9). NIST Special Publication 500-249. National Institute of Standards and Technology. Gaithersburg, Maryland. November 2000. http://trec.nist.gov/pubs/trec9/t9_proceedings.html

Hofman Miquel, Laura (2005) *Informationslinguistische Ressourcen für das Information Retrieval in der tschechischen Sprache im Rahmen des Cross Language Evaluation Forums (CLEF).* Master Thesis Information Science, University of Hildesheim.

Jensen, Niels (2005a) *Web Information Retrieval am Beispiel des WEB-GOV Korpus.* Master Thesis Information Science, University of Hildesheim.

Jensen, Niels (2005b) *Mehrsprachiges Information Retrieval mit einem WEB-Korpus.* In: Mandl, Thomas; Womser-Hacker, Christa (Eds.): Proceedings Vierter Hildesheimer Information Retrieval und Evaluierungsworkshop (HIER 2005) Hildesheim, 20.7.2005. Universitätsverlag Konstanz [Schriften zur Informationswissenschaft] to appear.

Mandl, Thomas (2005): The quest for the best pages on the web. In: Information Service & Use. To appear

McNamee, Paul; Mayfield, James (2004): *Character N-Gram Tokenization for European Language Text Retrieval*. In: Information Retrieval, vol. 7 (1/2). pp. 73-98.

Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005a): *Blueprint of a Cross-Lingual Web Retrieval Collection*. In: Journal of Digital Information Management, vol. 3 (1) pp. 9-13.

Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005b): *Overview of WebCLEF 2005*. In this volume.