The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier

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

Abstract

This paper describes our participation in the CLEF 2004 French monolingual task. We used our Terrier Information Retrieval platform and experimented with query expansion and query length normalisation.

1 Introduction

Terrier (http://ir.dcs.gla.ac.uk/terrier) is a toolkit for the rapid development of large-scale Information Retrieval (IR) applications. It is based on a framework for deriving non-parametric probabilistic models for IR. The framework deploys more than 50 Divergence From Randomness (DFR) models for term weighting [1]. The term weighting models are derived by measuring the divergence of the actual term distribution from that obtained under a random process. Terrier was demonstrated to be highly effective at retrieving Web documents at the recent TREC-11 and TREC-12, and is currently available as the search engine of the Web site of the Department of Computing Science at the University of Glasgow (http://www.dcs.gla.ac.uk/search).

In this paper, we report on our participation in the French Monolingual task. Our main aim was to test to which extent our existing English monolingual Terrier retrieval system could perform French retrieval, simply by changing the stemmer and stopword list from English into French. We chose French in order to test our system on new unfamiliar grounds. We opted for minimal language-specific normalisation changes, namely the use of a French stemmer and stopword list, and chose to exclude other performance enhancing options, such as POS-taggers and morphological analysers. Our secondary aim was to continue and complement our earlier work (TREC-11, TREC-12) on studying the effect of length normalisation on the retrieval performance, through the investigation of its impact on French IR. The outcome of this experimentation is being put to practical use, as we are currently working towards merging our existing English and French monolingual retrieval systems into one, thus extending our system to accommodate Cross Language Information Retrieval.

This paper is organised as follows. Section 2 presents a brief overview of the retrieval approaches adopted for our participation in CLEF 2004. Section 3 presents our official retrieval runs for the French monolingual task. Section 4 analyses the obtained results, along with a series of unofficial runs for the said task. Section 5 concludes with a brief summary of our participation in CLEF 2004 and the direction of our future research work.

2 System Setup

The following preprocessing steps were applied both to documents and queries. All input was tokenized. Punctuation marks and numbers of more than 4 digits were omitted. Proper nouns, abbreviations, acronyms, multi word units and compounds were not extracted or processed. Accents were preserved. Both queries and documents were stopped using the standard French stopword list available with the Snowball stemming algorithm for French [5]. We did not eliminate topic specific phrases such as “Les documents pertinents devront mentionner/parler de/donner des details sur…” from the queries. We did not use a stop stem list, as we used the stopword list before the stemming stage. We used the French stemmer from the Snowball family of stemmers, developed by Martin Porter [5]. The stemmer striped affixes from the index words in a specific order and applied repair strategies, where applicable, in order to reduce the input into clusters of words sharing the same stem.

We experimented with the PL2 weighting model, one of the Divergence From Randomness (DFR) term weighting models developed within Amati & Van Rijsbergen’s probabilistic framework for IR [1]. Using the PL2 model, the relevance score of a document d for a query q is given by:

$$\sum_{t \in q} qtf \cdot w(t,d)$$

where

- *qtf* is the frequency of term *t* in the query *q*,
- *w(t,d)* is the relevance score of a document *d* for the query term *t*, given by:

$$w(t,d) = (tfn \cdot \log_2 \frac{tfn}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tfn} - tfn \right) \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn)) \cdot \frac{1}{tfn+1} \qquad (1)$$

where

- λ is the mean and variance of a Poisson distribution. λ is given by $\frac{F}{N}$, ($F << N$), where F is the term frequency of the term *t* in the whole collection and N is the number of documents in the collection.

- *tfn* is the normalised within-document frequency of the term *t* in the document *d*. It is given by the normalisation 2 [1, 3]:

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

where

- *c* is a parameter.
- *tf* is the within-document frequency of the term *t* in the document *d*.
- l is the document length and avg_l is the average document length in the whole collection.

We estimated the parameter *c* of the normalisation 2 by measuring the normalisation effect on the term frequency distribution with respect to the document length distribution [4]. More specifically, our tuning approach automatically adjusted the parameter *c* to a value dependent on the topic fields used. For the runs submitted to CLEF 2004, we obtained the following values: *c*=4.83 for short queries (only Title field was used), *c*=1.56 for long queries (all the three fields were used), *c*=3.1 for queries using the Title and Description fields, and *c*=2.6 for queries using the Title and Narrative fields.

We have also used a query expansion mechanism, which follows the idea of measuring divergence from randomness. The approach can be seen as a generalisation of the approach used by Carpineto and Romano in which they applied the Kullback-Leibler divergence to the un-expanded version of BM25 [2, 3]. The query expansion formula is given by:

$$w = P_{Eq} \cdot \log_2 \frac{P_{Eq}}{P_D} \qquad (2)$$

where

$$P_{Eq} = \frac{withinDocumentFrequency}{totalDocumentLength}$$

and

$$P_D = \frac{termFrequency}{collectionLength}$$

where

- *withinDocumentFrequency* is the term frequency in the X top-retrieved documents. X depends on the setting we applied as indicated below.

- *termFrequency* is the term frequency of the given term in the collection.
- *totalDocumentLength* is the sum of the length of the X top-retrieved documents.
- *collectionLength* is the number of tokens in the whole collection.

For short queries, we extracted the 10 most informative terms from the top 3 retrieved documents as the expanded terms. For long queries, we extracted the 100 most informative terms from the top 25 retrieved documents as the expanded terms. For queries using the Title and Description fields we extracted the 10 most informative terms from the top 15 retrieved documents, and for queries using the Title and Narrative fields we extracted the top 15 informative terms from the top 3 retrieved documents. We added these terms to the query and repeated the retrieval stage.

3 Runs

This section presents our French monolingual retrieval runs submitted to CLEF 2004. We realised our runs on the CLEF 2004 document collection for the French Monolingual task, which consists of 90,261 newswire and newspaper articles published in 1995 (42,615 SDA and 47,646 Le Monde). There were 50 test topics. We submitted a total of 4 runs for the French monolingual task (Table 1), namely UOGLQ, UOGSQ, UOGLQQE, and UOGSQQE. The second column gives information on the topic fields selected for each run, namely T[itle], D[escription] and N[arrative]. The last column clarifies which runs used query expansion and which did not.

Run id	*Topic fields*	*Query Expansion*
UOGLQ	TDN	No
UOGSQ	T	No
UOGLQQE	TDN	Yes
UOGSQQE	T	Yes

Table 1: Runs submitted to the CLEF 2004 French Monolingual task.

In addition to the above runs, we also undertook further experiments, in order to test additional query length and query expansion settings, varying the number of expanded terms and the used number of top retrieved documents.

4 Results

This section summarises and discusses the results of our CLEF 2004 participation and of our additional runs. Table 2 reports on the main settings and scores of our collective runs. The submitted runs are in boldface. Column 2 presents the topic fields used for each run. The last column presents the Mean Average Precision (*MAP*) figures achieved.

Run id	*Topic Fields*	*c*	*MAP*
UOGSQ	**T**	**4.83**	**0.4237**
UOGSQQE	**T**	**4.83**	**0.3400**
UOGTD	TD	3.1	0.4485
UOGTDQE	TD	3.1	0.4222
UOGTN	TN	2.6	0.4431
UOGTNQE	TN	2.6	0.3711
UOGLQ	**TDN**	**1.56**	**0.4244**
UOGLQQE	**TDN**	**1.56**	**0.4186**

Table 2: Overview of our collective runs for CLEF 2004. Submitted runs are in boldface.

The best run was the one combining the topic fields of Title and Description (UOGTD), which slightly exceeded our best submitted run (UOGLQ). Overall, query length had little impact on the performance of the runs (*MAP* varied from 0.4237 to 0.4485).

In general, query expansion decreased the mean average precision of all the runs (see Table 2). Table 3 shows that query expansion does not work, independently of the length of the query.

Run id	*c*	*# terms/#documents*	*MAP*
UOGSQQE	**4.83**	**10/3**	**0.3400**
UOGTDQE	3.1	10/15	0.4222
UOGTNQE	2.6	15/3	0.3711
UOGLQQE	**1.56**	**100/25**	**0.4186**

Table 3: Query expansion deteriorated the retrieval performance independently of the query length.

In order to analyse the low performance of query expansion, we ran additional experiments with query expansion varying the number of expanded terms (#terms) and the used number of top retrieved documents (#documents) compared to the setting presented in Section 2. Table 4 shows the effect of the said parameter tuning on the performance of the system. Overall, query expansion deteriorated performance, independently of the used parameters. The used parameters in the official submitted runs were the optimal ones.

Finally, subsequent experiments revealed that the parameter *c* of the normalisation which was estimated by our tuning approach automatically (see Section 2) was indeed optimal, proving thus that the parameter tuning approach for term frequency normalisation adopted [4] is robust and efficient.

Official Runs				*Unofficial Runs*			
Run id	*c*	*#terms/#documents*	*MAP*	*Run id*	*c*	*#terms/#documents*	*MAP*
UOGSQQE	4.83	10/2	0.2876	UOGTDQE	3.1	50/3	0.3574
UOGSQQE	**4.83**	**10/3**	**0.3400**	UOGTDQE	3.1	20/3	0.4021
UOGSQQE	4.83	10/5	0.2998	UOGTDQE	3.1	10/3	0.4098
UOGSQQE	4.83	10/10	0.3113	UOGTDQE	3.1	10/10	0.4106
UOGSQQE	4.83	10/15	0.2981	UOGTDQE	3.1	15/10	0.3993
UOGSQQE	4.83	15/10	0.2780	UOGTDQE	3.1	10/13	0.4114
UOGSQQE	4.83	20/10	0.2882	UOGTDQE	3.1	10/15	0.3971
UOGLQQE	1.56	100/10	0.3745	UOGTNQE	2.6	10/2	0.3475
UOGLQQE	1.56	100/15	0.3889	UOGTNQE	2.6	10/3	0.3661
UOGLQQE	1.56	100/20	0.4088	UOGTNQE	2.6	10/10	0.3514
UOGLQQE	**1.56**	**100/25**	**0.4186**	UOGTNQE	2.6	15/3	0.3711
UOGLQQE	1.56	90/25	0.3550	UOGTNQE	2.6	20/3	0.3698
UOGLQQE	1.56	80/25	0.3401	UOGTNQE	2.6	50/3	0.3291
UOGLQQE	1.56	70/25	0.3228	UOGTNQE	2.6	100/25	0.2983

Table 4: Overview of our collective runs varying expanded terms (#terms)/top retrieved documents (#documents). Submitted runs are in boldface.

5 Conclusions and Future Work

This paper presented a French monolingual IR system developed at the University of Glasgow. The system was evaluated in the French monolingual track of CLEF 2004.

The experiments on which we briefly reported indicated the following. Our existing Terrier retrieval platform was shown to be truly modular, as it was extended to perform French monolingual IR successfully, simply by changing the stemming and stopword components from English into French, therefore with a very low overhead. Moreover, we found that query expansion performed poorly.

We are now investigating optimal ways to merge the French monolingual retrieval system described in this paper, with our existing English monolingual retrieval system, into a single crosslingual retrieval platform, the performance of which is to be tested in the CLEF 2005 multilingual track.

6 Acknowledgments

This project is funded by a UK Engineering and Physical Sciences Research Council (EPSRC) grant, number GR/R90543/01.

7 References

[1] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), volume 20(4), pages 357-389, October 2002.

[2] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS), 19(1), pages 1-27, January 2001.

[3] G. Amati. Probability Models for Information Retrieval based on Divergence from Randomness. Thesis of the degree of Doctor of Philosophy, Department of Computing Science, University of Glasgow, June 2003.

[4] B. He and I. Ounis. A study of parameter tuning for term frequency normalization. Proceedings of the Twelfth ACM CIKM International Conference on Information and Knowledge Management (CIKM), pages 10-16, New Orleans, LA, November 2003.

[5] http://www.snowball.tartarus.org/