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

Christina Lioma, Ben He, Vassilis Plachouras, and 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 Information Retrieval platform, Terrier, and experimented with query expansion and query length normalisation.

1 Introduction

Terrier (http://ir.dcs.gla.ac.uk/terrier) is a platform 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 document weighting [1]. The document 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 has been put to practical use, as we have merged our existing English and French monolingual retrieval systems into one. Some unofficial results on our bilingual runs are briefly presented here.

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 and our English-French

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 253 – 259, 2005.

unofficial runs. Section 4 analyses the obtained results, along with a further series of unofficial runs for the said tasks. 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. We used the standard French stopword list, which is 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 stripped 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 Terrier's DFR-based document weighting models. 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:

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

where

- λ is the mean and variance of a Poisson distribution. λ is given by $\frac{F}{N}$, $(F \ll N)$, 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 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 (KL) divergence to the un-expanded version of BM25 [2, 3]. In our experiments, we applied the KL model for query expansion. It is one of the Terrier DFR-based term weighting models. Using the KL model, the weight of a term t in the *#documents* top-ranked documents is given by:

$$w(t) = P_x \cdot \log_2 \frac{P_x}{P_c}$$

In the above formula,

$$P_x = \frac{tf_x}{l_x}$$

and

$$P_c = \frac{F}{token_c},$$

where tf_x is the frequency of the query term in the top-ranked documents. l_x is the sum of the length of the *#documents* top-ranked documents, and *#documents* is a parameter of the query expansion methodology. F is the term frequency of the query term in the whole collection. $token_c$ is the total 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, and additional French monolingual and English-French bilingual runs. We realised our runs on the CLEF 2004 document collection for the French monolingual ad-hoc 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.

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

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

In addition to the above runs, we also undertook further experiments, in order to test additional query length and query expansion settings. Specifically, we varied the number of expanded terms and the number of top retrieved documents used, both on French monolingual and on English-French bilingual retrieval. In the case of English-French bilingual runs, we manually translated the French queries into English, since we did not have the corresponding CLEF English queries available at the time. We then used the freely available Babelfish machine translation technology [6] to convert our English queries to French, and repeated the procedure described above in order to retrieve relevant documents from the French collection only.

4 Results

This section summarises and discusses the results of our CLEF 2004 participation and of our additional runs. Table 2 reports the main settings and scores of our collective runs. The submitted runs are in boldface. The second column presents the topic fields

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

Run id	*Topic Fields*	*c*	*MAP French*	*MAP English-French*
UOGSQ	**T**	**4.83**	**0.4237**	0.3456
UOGSQQE	**T**	**4.83**	**0.3400**	0.2754
UOGTD	TD	3.1	0.4485	0.3770
UOGTDQE	TD	3.1	0.4222	0.3425
UOGTN	TN	2.6	0.4431	0.3698
UOGTNQE	TN	2.6	0.3711	0.3024
UOGLQ	**TDN**	**1.56**	**0.4244**	0.3867
UOGLQQE	**TDN**	**1.56**	**0.4186**	0.3339

used for each run. The last two columns present the Mean Average Precision (*MAP*) figures achieved for French and for English-French retrieval accordingly.

The best French monolingual run was the one combining the topic fields of Title and Description (UOGTD), which slightly exceeded our best submitted run (UOGLQ). The best English-French bilingual run was the one combining the fields of Title, Description and Narrative. Overall, query length had little impact on the performance of the runs (*MAP* varied from 0.4237 to 0.4485 for French, and from 0.3456 to 0.3867 for English-French). It should be noted that the English-French retrieval performance of our system is highly correlated with the French monolingual retrieval performance. The rank correlation coefficient of the two lists of *MAP* values is R = 0.9286, with a p-value of 0.002232. This shows that Terrier has performed consistently.

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

Run id	*c*	*#terms/ #documents*	*MAP French*
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 4. Overview of our collective runs varying expanded terms (*#terms*)/top retrieved documents (*#docs*). Submitted runs are in boldface

Official Runs				*Unofficial Runs*			
Run id	*c*	*#terms/ #docs*	*MAP*	*Run id*	*c*	*#terms/ #docs*	*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

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.

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 number of top retrieved documents used (*#documents*) compared to the setting mentioned in Section 2. Table 4 shows the effect of that parameter tuning on the performance of the system. Overall, query expansion deteriorated performance, independently of the parameters used. The parameter settings in the official submitted runs were actually 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 (see Table 5). This shows that the parameter tuning approach for term frequency normalisation [4] which we adopted is robust and efficient, performing as well on both French and English document collections [4].

Table 5. Overview of our collective runs varying the *c* value. Submitted runs are in boldface

Official Runs			*Unofficial Runs*		
Run id	*c*	*MAP*	*Run id*	*c*	*MAP*
UOGSQ	4.0	0.4222	UOGTD	2.5	0.4454
UOGSQ	4.50	0.4232	UOGTD	3.0	0.4442
UOGSQ	**4.83**	**0.4237**	UOGTD	3.1	0.4485
UOGSQ	5.0	0.4231	UOGTD	3.5	0.4481
UOGSQ	5.5	0.4199	UOGTD	4.0	0.4480
UOGLQ	1.0	0.4173	UOGTN	2.0	0.4425
UOGLQ	1.25	0.4227	UOGTN	2.5	0.4427
UOGLQ	**1.56**	**0.4244**	UOGTN	2.6	0.4431
UOGLQ	2.0	0.4240	UOGTN	3.0	0.4430
UOGLQ	2.5	0.4239	UOGTN	3.5	0.4422

5 Conclusions and Future Work

This paper presented a French monolingual IR system and an English-French bilingual IR system, both of which were developed at the University of Glasgow. The French monolingual IR system was evaluated in the French monolingual ad-hoc 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, which is in agreement with a number of other retrieval systems participating at CLEF 2004, thus indicating that the specific data collection may be partly responsible for the bad

performance of query expansion. We have now merged our French and English monolingual retrieval systems into a single bilingual retrieval platform. We are currently working towards improving and enhancing this bilingual platform, the performance of which will be tested in the CLEF 2005 multilingual track.

Acknowledgments

This project is funded by a UK Engineering and Physical Sciences Research Council (EPSRC) grant, number GR/R90543/01. The project funds the development of the Terrier Information Retrieval framework (http://www.ir.dcs.gla.ac.uk/terrier).

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/

[6] http://www.babelfish.altavista.com/