Technical University of Lisbon CLEF 2008 Submission (TEL@CLEF Monolingual Task)

Jorge Machado, Bruno Martins and José Borbinha
Departmento de Engenharia Informática, Technical University of Lisbon, Portugal

Abstract

We describe our participation in the TEL@CLEF task of the CLEF 2008 ad-hoc track, where we measured the retrieval performance of the IR service that is currently under development as part of the DIGMAP project. DIGMAP's IR service is mostly based on Lucene, together with extensions for using query expansion and multinomial language modelling. In our runs, we experimented combinations of query expansion, Lucene's off-the-shelf ranking scheme and the ranking scheme based on multinomial language modelling. Results show that query expansion and multinomial language modelling both result in increased performance.

1 Introduction and background

One task of the ad-hoc track at the 2008 edition of the Cross Language Evaluation Forum (CLEF) addresses the problem of searching and retrieving relevant items from collections of bibliographic records from The European Library (TEL@CLEF). Three target collections were provided, each corresponding to a monolingual retrieval task where we participated:

- TEL Catalogue records in English. Copyright British Library (BL)
- TEL Catalogue records in French. Copyright Bibliothèque nationale de France (BnF)
- TEL Catalogue records in German. Copyright Austrian National Library (ONB)

The evaluation task aimed at investigating the best approaches for retrieval from library catalogues, where the information is frequently very sparse and often stored in unexpected languages.

This paper describes the participation of the Technical University of Lisbon at the TEL@CLEF task. Our experiments aimed at measuring the retrieval performance of the IR service that is currently under development as part of DIGMAP[1], an EU-funded project which addresses the development of services for virtual digital libraries of materials related to historical cartography (Pedrosa (2008)). DIGMAP collects bibliographic metadata from European national libraries and other relevant third-party providers (e.g. collections with descriptions available through OAI-PMH), aiming to provide advanced searching and browsing mechanisms that combine thematic, geographic and temporal aspects. In case of success, the ultimate goal of the project is to become fully integrated into The European Library.

The DIGMAP text retrieval service is mostly based on Lucene, together with extensions for using query expansion and multinomial language modelling. A previous version of the system was described in the MSc thesis of Machado (2005) and we are now in the process of developing extensions for geo-temporal information retrieval. In CLEF, we experimented combinations of query expansion, Lucene's off-the-shelf ranking scheme and the ranking scheme based on multinomial language modelling.

2 The experimental environment

The underlying IR system used in our submissions is based on Lucene[2], together with a multinomial language modelling extension developed at the University of Amsterdam and a query expansion extension developed by Neil Rubens. The following subsections detail these components.

2.1 Lucene's off-the-shelf retrieval model

We started with Lucene's off-the-shelf retrieval model. For a collection D, document d and query q, the ranking score is given by the next formula:

---

[1] http://www.digmap.eu

[2] http://lucene.apache.org

$$ranking(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where

$$tf_{t,X} = \sqrt{termFrequency(t,X)},$$

$$idf_t = 1 + \log \frac{|D|}{documentFrequency(t,D)},$$

$$norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t^2},$$

$$norm_d = \sqrt{|d|},$$

$$coord_{q,d} = \frac{|q \cap d|}{|q|}$$

Lucene has been extensively used in previous editions of the CLEF, NTCIR and TREC joint evaluation experiments.

2.2 Lucene extension based on multinomial language modelling

We experimented with a Lucene extension that implements a retrieval scheme based on estimating a language model (LM) for each document, using the formula described by Hiemstra (2001). This extension was developed at the Informatics Institute of the University of Amsterdam[3]. For any given query, it ranks the documents with respect to the likelihood that the document's LM generated the query:

$$ranking(d,q) = P(d|q) \propto P(d) \cdot \prod_{t \in q} P(t|d)$$

In the formula, *d* is a document and *t* is a term in query *q*. The probabilities are reduced to rank-equivalent logs of probabilities. To account for data sparseness, the likelihood *P(t|d)* is interpolated using Jelinek-Mercer smoothing.

$$P(d|q) = P(d) \cdot \prod_{t \in q} ((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d))$$

In the formula, *D* is the collection and λ is a smoothing parameter (in our experiments set to the default value of 0.15). The model needs to estimate three probabilities: the prior probability of the document, *P(d)*; the probability of observing a term in a document, *P(t|d)* and the probability of observing the term in the collection, *P(t|D)*. Assuming the query terms to be independent, and using a linear interpolation of a document model and a collection model to estimate the probability of a query term, the probabilities can be estimated using maximum likelihood estimates:

$$P(t|d) = \frac{termFrequency(t,d)}{|d|}$$

$$P(t|D) = \frac{documentFrequency(t,D)}{\sum_{t' \in D} documentFrequency(t',D)}$$

$$P(d) = \frac{|d|}{\sum_{d' \in D} |d'|}$$

This language modelling approach has been used in past experiments within the CLEF, NTCIR and TREC joint evaluation campaigns – see for example Ahn et. al (2005).

---

[3] http://ilps.science.uva.nl/Resources/

2.3 Rocchio query expansion

The fact that there are frequently occurring spelling variations and synonyms for any query term degrades the performance of standard techniques for ad-hoc retrieval. To overcome this problem, we experimented with the method for pseudo feedback query expansion proposed by Rocchio (1971). The Lucene extension from the LucQE project[4] implements this approach. On test data from the 2004 TREC Robust Retrieval Track, LucQE achieved a MAP score of 0.2433 using Rocchio query expansion.

Assuming that the top D documents returned for an original query q_i are relevant, a better query q_{i+1} can be given by the terms resulting from the formula bellow:

$$q_{i+1} = \alpha \cdot q_i + \frac{\beta}{|D|} \cdot \sum_{d_r \in D} termWeight(d_r)$$

In the formula, α and β are tuning parameters. In our experiments, they were set to the default values of 1.0 and 0.75. The system was allowed to add up to *200* terms extracted from the *10* highest ranked documents (i.e. the $|D|$ parameter) from the original query q_i. The query expansion method was tuned through experiments with the ad-hoc collections and relevance judgements from previous CLEF editions.

2.4 Processing the topics and the document collections

Before the actual indexing, the document collections (i.e. the bibliographic records) were passed through the following pre-processing operations:

- **Field Weighting -** The bibliographic records composing the collections from the TEL@CLEF experiment contain structured information in the form of document fields such as *title* or *subject*. We use the scheme proposed by Robertson et. al (2004) to weight the different document field according to their importance. Instead of changing the ranking formulas in order to introduce boosting factors, we generate virtual documents in which the content of some specific fields is repeated. The combination used in our experiments is based on repeating the *title* field three times, the *subject* field twice and keeping the other document fields unchanged.
- **Normalisation –** The structured documents were converted to unstructured documents for the process of indexing, removing the XML tags and putting the element's contents in separate sentences.

Topic processing was fully automatic and the queries submitted to the IR engine were generated using all parts of the topics (i.e. title, description and narrative). The generation of the actual queries from the query topics was based on the following sequence of processing operations:

- **Parsing and Normalisation -** All characters were reduced to the lowercase unaccented equivalents (i.e. "Ö" reduced to "o" and "É" to "e" etc.) in order to maximise matching.
- **Stop Word Removal -** Stopword lists were used to remove terms that carry little meaning and would otherwise introduce noise. The considered stop words came from the minimized lists distributed with Lucene, containing words such as articles, pronouns, prepositions, conjunctions or interjections. For English, French and German, these lists contained 120, 155 and 231 terms, respectively.
- **Retrieval –** The resulting queries were submitted to the IR system, which had been used to index the document collections. In some of the submitted runs, variations of the Porter (1980) stemming algorithm specific to the language of the collection were used on both the queries and the documents. The stemming algorithms came from the Snowball package[5].

Lucene internally normalizes documents and queries to lower case, also removing stop-words. However, explicitly introducing these operations when processing the topics has the advantage of facilitating the development of more advanced topic processing (e.g. adding query expansion methods).

3 The experimental story

We submitted 12 official runs to the CLEF evaluation process, a total of 4 runs for each of the languages/collections under consideration in the monolingual task. The conditions under test for each of the submitted runs are as follows:

---

[4] http://lucene-qe.sourceforge.net/

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

1. Baseline run using the off-the-shelf retrieval model from Lucene.
2. Lucene with the language modelling extension.
3. Lucene with the language modelling extension and language-specific stemming algorithms.
4. Lucene's off-the-shelf retrieval model with the extension for doing Rocchio query expansion.

We also discuss here the results of some unofficial runs that resulted from experiments that we performed with our retrieval engine. The test conditions for these unofficial runs are:

5. Lucene with the language modelling extension and Rocchio query expansion.
6. Lucene with the language modelling extension, Rocchio query expansion and stemming.
7. Lucene's off-the-shelf retrieval model with Rocchio query expansion and stemming.

4 Results

Table 1 shows the obtained results for the official runs that make up our TEL@CLEF experiments. The results show that, in terms of the mean average precision (MAP), run 3 consistently outperforms our other submissions. The language modelling approach, complemented with the use of stemming, indeed seems beneficial to the retrieval task at study. Run 4 (i.e. query expansion) also consistently outperformed the baseline run with the off-the-shelf Lucene retrieval scheme, although run 3 (i.e. language modelling without stemming) failed to improve over the baseline.

	English				French				German			
	RUN 1	**RUN 2**	**RUN 3**	**RUN 4**	**RUN 1**	**RUN 2**	**RUN 3**	**RUN 4**	**RUN 1**	**RUN 2**	**RUN 3**	**RUN 4**
num_q	50	50	50	50	50	50	50	50	50	50	50	50
num_ret	50000	50000	50000	50000	50000	50000	50000	50000	48368	48368	49138	50000
num_rel	2533	2533	2533	2533	1339	1339	1339	1339	1637	1637	1637	1637
num_rel_ret	1858	1884	2056	2060	830	791	1028	891	736	752	943	921
map	0.2976	0.2969	**0.3623**	0.3048	0.2174	0.2020	**0.2341**	0.2190	0.1218	0.1404	**0.2298**	0.1605
gm_ap	0.2015	0.2008	0.2418	0.1939	0.0746	0.0648	0.0941	0.0553	0.0427	0.0534	0.0964	0.0475
R-prec	0.3106	0.3118	**0.3649**	0.3130	0.2463	0.2297	**0.2547**	0.2406	0.1446	0.1606	**0.2432**	0.1838
bpref	0.3126	0.3095	0.3619	0.3415	0.2215	0.2068	0.2315	0.2427	0.1203	0.1374	0.2346	0.1759
recip_rank	0.8263	0.8271	0.8318	0.7936	0.6143	0.5984	0.6309	0.5768	0.5069	0.5950	0.7007	0.5382
ircl_prn.0.00	0.8474	0.8580	0.8580	0.8259	0.6386	0.6224	0.6564	0.6120	0.5368	0.6431	0.7292	0.5764
ircl_prn.0.10	0.6917	0.6470	0.6912	0.6305	0.4804	0.4428	0.4730	0.4800	0.3512	0.3918	0.5392	0.3349
ircl_prn.0.20	0.4997	0.4979	0.5527	0.4829	0.3680	0.3450	0.3520	0.3636	0.2411	0.2562	0.4381	0.2658
ircl_prn.0.30	0.3753	0.3858	0.4537	0.3976	0.3035	0.3010	0.3057	0.2974	0.1505	0.1687	0.3102	0.2268
ircl_prn.0.40	0.3160	0.3166	0.3824	0.3127	0.2236	0.2134	0.2644	0.2318	0.1109	0.1348	0.2417	0.1880
ircl_prn.0.50	0.2654	0.2775	0.3439	0.2611	0.1812	0.1774	0.2265	0.1962	0.0749	0.0861	0.1839	0.1613
ircl_prn.0.60	0.1935	0.2093	0.2870	0.2245	0.1453	0.1331	0.1857	0.1553	0.0581	0.0741	0.1583	0.1251
ircl_prn.0.70	0.1351	0.1448	0.2464	0.1803	0.1089	0.0896	0.1285	0.1107	0.0408	0.0571	0.0879	0.0723
ircl_prn.0.80	0.1106	0.1170	0.1937	0.1362	0.0713	0.0634	0.0978	0.0825	0.0336	0.0354	0.0690	0.0483
ircl_prn.0.90	0.0668	0.0752	0.1153	0.0806	0.0403	0.0456	0.0734	0.0463	0.0154	0.0223	0.0236	0.0227
ircl_prn.1.00	0.0149	0.0177	0.0345	0.0320	0.0099	0.0130	0.0391	0.0124	0.0044	0.0072	0.0041	0.0034
P@5	0.6000	0.5720	0.6160	0.5920	0.3720	0.3560	0.3640	0.3680	0.3040	0.3640	0.4800	0.2960
P@10	0.4840	0.4920	0.5160	0.5020	0.2900	0.2800	0.3020	0.3160	0.2440	0.2680	0.4040	0.2560
P@15	0.4347	0.4293	0.4667	0.4373	0.2520	0.2427	0.2680	0.2600	0.2213	0.2373	0.3547	0.2453
P@20	0.4000	0.3930	0.4250	0.3910	0.2360	0.2270	0.2430	0.2270	0.2020	0.2110	0.3150	0.2260
P@30	0.3500	0.3373	0.3800	0.3333	0.2067	0.2020	0.2147	0.1853	0.1793	0.1847	0.2540	0.1973
P@100	0.2072	0.2124	0.2442	0.2048	0.1102	0.1036	0.1230	0.1064	0.0850	0.0892	0.1204	0.1096
P@200	0.1308	0.1330	0.1559	0.1396	0.0638	0.0626	0.0780	0.0664	0.0496	0.0518	0.0729	0.0686
P@500	0.0663	0.0681	0.0758	0.0728	0.0304	0.0292	0.0374	0.0322	0.0242	0.0246	0.0344	0.0333
P@1000	0.0372	0.0377	0.0411	0.0412	0.0166	0.0158	0.0206	0.0178	0.0147	0.0150	0.0189	0.0184

Table 1. Results for the official runs submitted to TEL@CLEF.

The charts at Figure 1 show precision-recall curves for the official runs, separating the results according to the language (i.e. English, French and German submissions, from left to right).

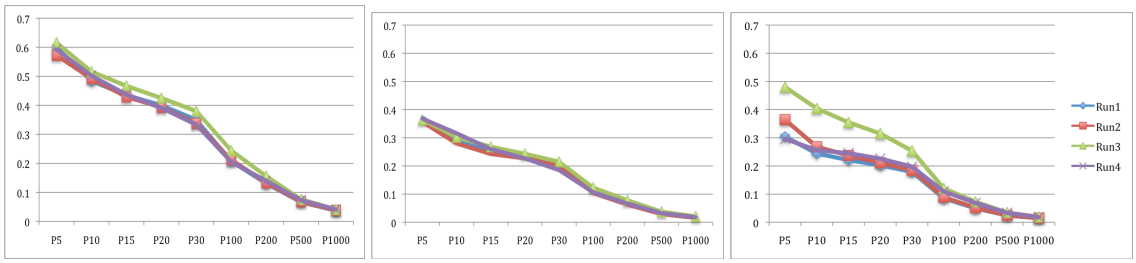

Figure 1. Precision vs. Recall curves for the official runs submitted to TEL@CLEF.

Table 2 shows the results obtained from the unofficial runs that were described in the previous section. The values show that, in terms of MAP, naively combining the language modelling approach with query expansion results in a poor retrieval performance. Results also show that complementing run 4 (Lucene's standard retrieval model, plus Rocchio query expansion) with stemming can be beneficial, particularly in the case of the English collection.

	English			French			German		
	RUN 5	**RUN 6**	**RUN 7**	**RUN 5**	**RUN 6**	**RUN 7**	**RUN 5**	**RUN 6**	**RUN 7**
num_q	50	50	50	50	50	50	50	50	50
num_ret	50000	50000	50000	50000	50000	50000	50000	50000	50000
num_rel	2533	2533	2533	1339	1339	1339	1637	1637	1637
num_rel_ret	1343	1448	2124	583	598	976	734	842	1065
map	**0.1776**	**0.2301**	**0.3527**	**0.1175**	**0.1404**	**0.2258**	**0.1035**	**0.1591**	**0.2437**
gm_ap	0.0942	0.1270	0.2499	0.0263	0.0347	0.0858	0.0255	0.0440	0.0844
R-prec	**0.2241**	**0.2733**	**0.3576**	**0.1519**	**0.1893**	**0.2358**	**0.1395**	**0.1996**	**0.2616**
bpref	0.2923	0.3328	0.3768	0.1768	0.2038	0.2496	0.1862	0.2616	0.2685
recip_rank	0.7107	0.7434	0.8324	0.4470	0.5429	0.5721	0.4676	0.6222	0.6221
ircl_prn.0.00	0.7591	0.7882	0.8717	0.5070	0.5866	0.6158	0.5101	0.6560	0.6751
ircl_prn.0.10	0.4847	0.5838	0.7180	0.3435	0.3859	0.4937	0.3414	0.4677	0.5178
ircl_prn.0.20	0.3039	0.4375	0.5532	0.2449	0.2810	0.3488	0.1773	0.3297	0.4183
ircl_prn.0.30	0.2362	0.3227	0.4487	0.1418	0.1698	0.2882	0.1176	0.2196	0.3491
ircl_prn.0.40	0.1815	0.2441	0.3711	0.1041	0.1274	0.2477	0.0895	0.1521	0.2991
ircl_prn.0.50	0.1363	0.1829	0.3155	0.0873	0.1073	0.2137	0.0720	0.1013	0.2381
ircl_prn.0.60	0.0779	0.1163	0.2596	0.0519	0.0632	0.1681	0.0454	0.0570	0.1901
ircl_prn.0.70	0.0438	0.0735	0.2092	0.0191	0.0326	0.1161	0.0140	0.0198	0.1109
ircl_prn.0.80	0.0220	0.0361	0.1616	0.0063	0.0241	0.0873	0.0053	0.0076	0.0666
ircl_prn.0.90	0.0110	0.0114	0.1048	0.0033	0.0058	0.0498	0.0007	0.0014	0.0265
ircl_prn.1.00	0.0004	0.0017	0.0503	0.0006	0.0014	0.0229	0.0007	0.0014	0.0063
P@5	0.5160	0.5920	0.6600	0.3080	0.3640	0.3880	0.3360	0.4640	0.4640
P@10	0.4220	0.4860	0.5460	0.2420	0.2520	0.3280	0.2520	0.3580	0.4060
P@15	0.3547	0.4107	0.4720	0.1853	0.2067	0.2733	0.2027	0.2787	0.3427
P@20	0.3120	0.3620	0.4360	0.1480	0.1740	0.2490	0.1730	0.2400	0.3110
P@30	0.3500	0.3027	0.3760	0.2067	0.1360	0.2147	0.1793	0.1940	0.2540
P@100	0.2072	0.1572	0.2350	0.1102	0.0648	0.1230	0.0850	0.0904	0.1204
P@200	0.1308	0.1006	0.1548	0.0638	0.0389	0.0780	0.0496	0.0559	0.0729
P@500	0.0663	0.0498	0.0760	0.0304	0.0292	0.0374	0.0242	0.0246	0.0344
P@1000	0.0372	0.0290	0.0425	0.0166	0.0158	0.0206	0.0147	0.0150	0.0189

Table 2. Results for the unofficial runs using the TEL@CLEF collections.

5 Conclusions

The obtained results support the support the hypotheses that using Rocchio query expansion and a ranking scheme based on language modelling can be beneficial to the CLEF ad-hoc task. Our official runs only made use of relatively simple techniques, but we're now in the process of implementing additional features into our retrieval engine. These include geographic information retrieval extensions with basis on Local Lucene[6] and advanced query expansion methods using bibliographic information.

6 Bibliography

Porter, M. F. (1980). "An algorithm for suffix stripping". In: Sparck Jones, K. & Willett, P. (eds.), *(1997) Readings in Information Retrieval.*, pp. 313 - 316. San Francisco: Morgan Kaufmann.

Hiemstra, D. (2001) "Using Language Models for Information Retrieval", Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente.

[6] http://sourceforge.net/projects/locallucene

Rocchio, J. J. (1971) "Relevance Feedback in Information Retrieval". In: The SMART Retrieval System: Experiments in Automatic Document Processing., pp 313 - 323. Prentice Hall

Machado, J. and Borbinha, J. (2008) "Mitra: A Metadata Aware Web Search Engine for Digital Libraries", M.Sc. Thesis, Departamento de Engenharia Informática, Technical University of Lisbon

Robertson, S., Zaragoza, H., and Taylor, M. (2004). "Simple BM25 extension to multiple weighted fields". In Proceedings of the Thirteenth ACM international Conference on information and Knowledge Management (Washington, D.C., USA, November 08 - 13, 2004). CIKM '04. ACM, New York, NY, 42-49. DOI= http://doi.acm.org/10.1145/1031171.1031181

Ahn, D. D., Azzopardi, L., Balog, K., Fissaha, A. S., Jijkoun, V., Kamps, J., Müller, K., de Rijke, M. and Erik Tjong Kim Sang (2005) "The University of Amsterdam at TREC 2005". Working Notes for the 2005 Text Retrieval Conference

Pedrosa, G., Luzio, J., Manguinhas, H., and Martins, B. (2008) "DIGMAP: A service for searching and browsing old maps". In Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (Pittsburgh PA, PA, USA, June 16 - 20, 2008). JCDL '08. ACM, New York, NY, 431-431. DOI= http://doi.acm.org/10.1145/1378889.1378978