Published in Lecture Notes in Computer Science 4730, 137-144, 2007
which should be used for any reference to this work

Experiments with Monolingual, Bilingual, and Robust Retrieval

Jacques Savoy and Samir Abdou

Computer Science Department, Rue Emile Argand 11, 2009 Neuchatel,
University of Neuchatel, Switzerland

Jacques.Savoy@unine.ch, Samir.Abdou@unine.ch

Abstract. For our participation in the CLEF 2006 campaign, our first objective was to propose and evaluate a decompounding algorithm and a more aggressive stemmer for the Hungarian language. Our second objective was to obtain a better picture of the relative merit of various search engines for the French, Portuguese/Brazilian and Bulgarian languages. To achieve this we evaluated the test-collections using the Okapi approach, some of the models derived from the *Divergence from Randomness* (DFR) family and a language model (LM), as well as two vector-processing approaches. In the bilingual track, we evaluated the effectiveness of various machine translation systems for a query submitted in English and automatically translated into the French and Portuguese languages. After blind query expansion, the MAP achieved by the best single MT system was around 95% for the corresponding monolingual search when French was the target language, or 83% with Portuguese. Finally, in the robust retrieval task we investigated various techniques in order to improve the retrieval performance of difficult topics.

1 Introduction

During the last few years, the IR group at University of Neuchatel has been involved in designing, implementing and evaluating IR systems for various natural languages, including both European [1] and popular Asian [2] languages. In this context, our first objective was to promote effective monolingual IR for these languages. Our second aim was to design and evaluate effective bilingual search techniques (using a query-based translation approach), and our third objective was to propose effective multilingual IR systems.

The rest of this paper is organized as follows: Section 2 explains the principal features of different indexing and search strategies, and then evaluates them using the available corpora. The data fusion approaches used in our experiments and our official results are outlined in Section 3. Our bilingual experiments are presented and evaluated in Section 4. Finally, Section 5 presents our first experiments in the robust track, limited however to the French language.

2 Indexing and Searching Strategies

In order to obtain a broader view of the relative merit of various retrieval models, we first adopted the classical $tf \cdot idf$ weighting scheme (with cosine normalization). We then computed the inner product to measure similarity between documents and requests. Although various other indexing weighting schemes have been suggested, in our study we will only consider the IR model "Lnu" [3].

In addition to these two IR models based on the vector-space paradigm, we also considered probabilistic approaches such as the Okapi model [4]. As a second probabilistic approach, we implemented several models derived from the *Divergence from Randomness* (DFR) family [5]. The GL2 approach for example is based on the following equations:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij}) \quad \text{with} \tag{1}$$

$$Prob^2_{ij} = tfn_{ij} \;/\; (tfn_{ij} + 1) \tag{2}$$

$$tfn_{ij} = tf_{ij} \cdot \log_2 [1 + ((c \cdot mean\ dl) \;/\; l_i)] \tag{3}$$

$$Inf^1_{ij} = -\log_2 [1/(1+\lambda_j)] - tfn_{ij} \cdot \log_2 [\lambda_j/(1+\lambda_j)], \quad \lambda_j = tc_j/n \tag{4}$$

where tc_j represents the number of occurrences of term t_j in the collection, l_i the length of document D_i, *mean dl* is the document mean length, and n the number of documents in the corpus. On the other hand, the PL2 model uses Equation 5 instead of Equation 4.

$$Inf^1_{ij} = -\log_2 \left[\frac{e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}}}{tf_{ij}!} \right] \tag{5}$$

Finally, we considered an approach based on the language model (LM) [6] known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus be estimated directly, based on occurrence frequencies in document D_i or corpus C. Within this language model paradigm, various implementation and smoothing methods might be considered, and in this study we adopted a model which combines an estimate based on document ($P[t_j|D_i]$) and on corpus ($P[t_j|C]$) [6].

$$Prob[D_i|Q] = Prob[D_i] \prod_{tj \in Q} [\lambda_j \cdot Prob[t_j|D_i] + (1-\lambda_j) \cdot Prob[t_j|C]] \tag{6}$$

$$Prob[t_j|D_i] = tf_{ij}/l_i \quad \text{and} \quad Prob[t_j|C] = df_j/lc \quad \text{with} \quad lc = \sum_k df_k \tag{7}$$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and fixed at 0.35) and lc the size of the corpus C.

During this evaluation campaign, we applied the stopword lists and stemmers that were used in our CLEF 2005 participation [7]. In our Bulgarian stopword list however we corrected certain errors (including the removal of words having a

clear meaning and introduced by mistake in the suggested stopword list). For the Hungarian collection, we automatically decompounded long words (more than 6 characters) using our own algorithm [8]. In this experiment, compound words were replaced by their components in both documents and queries. This year, we tried to be more aggressive, adding 17 rules to our Hungarian stemmer in order to also remove certain derivational suffixes (e.g., "jelent" (to mean) and "jelentés" (meaning), or "tánc" (to dance) and "táncol" (dance)).

To measure the retrieval performance, we adopted the mean average precision (MAP) computed by the `trec_eval` system. Then we applied the bootstrap methodology [9] in order to statistically determine whether or not a given search strategy would be better than another. Thus, in the tables included in this paper we underline statistically significant differences resulting from the use of a two-sided non-parametric bootstrap test (significance level fixed at 5%).

We indexed the various collections using words as indexing units. Table 1 shows evaluations on our four probabilistic models, as well as two vector-space schemes. The best performances under given conditions are shown in bold type in this table, and they are used as a baseline for our statistical testing. The underlined results therefore indicate which MAP differences can be viewed as statistically significant when compared to the best system value. As shown in the top part of Table 1, the Okapi model was the best IR model for the French and Portuguese/Brazilian collections. For these two corpora however, MAP differences between the various probabilistic IR models were not always statistically significant. The DFR-GL2 model provided the best results for the Bulgarian collection, while for the Hungarian corpus, the DFR-PL2 approach resulted in the best performance.

Table 1. MAP of Single Searching Strategies

	Mean average precision			
Query Model	French TD 49 queries	Portuguese TD 50 queries	Bulgarian TD 50 queries	Hungarian TD 48 queries
Okapi	**0.4151**	**0.4118**	0.2614	0.3392
DFR-PL2	0.4101	0.4147	N/A	**0.3399**
DFR-GL2	0.3988	0.4033	**0.2734**	0.3396
LM	0.3913	0.3909	0.2720	0.3344
Lnu-ltc	0.3738	0.4212	0.2663	0.3303
tf idf	0.2606	0.2959	0.1898	0.2623

It was observed that pseudo-relevance feedback (PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [3] with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add m terms extracted from the k best ranked documents from the original query. To evaluate this proposition, we used three IR

Table 2. MAP Before and After Blind-Query Expansion

	Mean average precision			
Query TD Model	French 49 queries	Portuguese 50 queries	Bulgarian 50 queries	Hungarian 48 queries
Okapi	0.4151	0.4118	0.2614	0.3392
k docs/ m terms	10/20 0.4222	10/20 0.4236	3/50 0.2833	3/10 0.3545
	10/30 0.4269	10/30 0.4361	5/50 0.2798	5/10 0.3513
	10/40 0.4296	10/40 0.4362	3/90 0.2854	5/15 0.3490
	10/50 0.4261	10/50 **0.4427**	5/90 0.2809	10/15 0.3492
DFR-GL2	0.3988	0.4033	0.2734	0.3396
k / m	10/50 0.4356	10/20 0.4141	5/10 **0.3327**	5/50 **0.4059**
LM	0.3913	0.3909	0.2720	0.3344
k / m	10/50 **0.4509**	10/30 0.4286	10/20 0.3305	3/40 0.3855

models and enlarged the query by the 10 to 90 terms extracted from the 3 to 10 best-ranked articles (see Table 2).

For the French corpus, the percentage of improvement varied from +3.5% (Okapi model, 0.4151 vs. 0.4296) to +15.2% (LM model, 0.3913 vs. 0.4509). For the Portuguese/Brazilian corpus, the increased enhancement was +2.7% (DFR-GL2 model, 0.4033 vs. 0.4141) to +9.6% (LM model, 0.3909 vs. 0.4286). For the Bulgarian language, the use of a blind query expansion improved the MAP from +9.2% (Okapi model, 0.2614 vs. 0.2854) to +21.7% (DFR-GL2 model, 0.2734 vs. 0.3327). Finally, with the Hungarian language, a blind query expansion might enhance retrieval effectiveness from +4.5% (Okapi model, 0.3392 vs. 0.3545) to +19.5% (DFR-GL2 model, 0.3396 vs. 0.4059).

3 Data Fusion and Official Results

It is assumed that combining different search models should improve retrieval effectiveness, due to the fact that different document representations might retrieve different pertinent items, thus increasing overall recall [10]. In this current study we combined two or three probabilistic models, representing both the parametric (Okapi and DFR) and non-parametric (LM) probabilistic approaches. To achieve this we evaluated various fusion operators (see [7] for a list of their precise descriptions) such as the Norm RSV, where a normalization procedure is applied before adding document scores computed by different search models.

Table 3 shows the exact specifications of our best performing official monolingual runs. In these experiments, we combined the Okapi, and the LM probabilistic models using the Z-score data fusion operator [7] for the French and Portuguese/Brazilian corpora. We obtained the best results when using the LM model combined with the DFR-GL2 model for the Bulgarian corpus or when combining the Okapi, DFR-PL2 and LM models for the Hungarian language.

Table 3. Description and MAP of our Best Official Monolingual Runs

Language	Index	Query	Model	Query exp.	MAP	comb. MAP
French	word	TD	Okapi	10 docs/60 terms	0.4275	Norm RSV
UniNEfr3	word	TD	LM	10 docs/30 terms	0.4460	**0.4559**
Portuguese	word	TD	Okapi	10 docs/80 terms	0.4276	Z-score
UniNEpt1	word	TD	LM	10 docs/50 terms	0.4403	**0.4552**
Bulgarian	word	TD	LM	5 docs/40 terms	0.3201	Z-score
UniNEbg2	4-gram	TD	GL2	10 docs/90 terms	0.2941	**0.3314**
Hungarian	word	TD	PL2	3 docs/40 terms	0.3794	Z-score
	word	TD	LM	3 docs/70 terms	0.3815	
UniNEhu2	4-gram	TD	Okapi	3 docs/100 terms	0.3870	**0.4308**

4 Bilingual Information Retrieval

Due to a time constraint, we limited our participation in the bilingual track to the French and Portuguese/Brazilian languages. Moreover, as the query submission language we chose English, which was automatically translated into the two other languages, using ten freely available machine translation (MT) systems (listed in the first column of Table 4).

Table 4. MAP of Various Translation Devices (Okapi model)

	Mean average precision (% monolingual)	
TD queries Model	French 49 queries	Portuguese 50 queries
Manual & PRF (10/40)	0.4296	0.4389
AlphaWorks	0.3378 (78.6%)	N/A
AppliedLanguage	0.3726 (86.7%)	0.3077 (70.1%)
BabelFish	0.3771 (87.8%)	0.3092 (70.4%)
FreeTranslation	0.3813 (88.8%)	0.3356 (76.5%)
Google	0.3754 (87.4%)	0.3070 (69.9%)
InterTrans	0.2761 (64.3%)	0.3343 (76.2%)
Online	0.3941 (91.8%)	**0.3677** (83.3%)
Reverso	**0.4081** (95.0%)	0.3531 (80.5%)
Systran	N/A	0.3077 (70.1%)
WorldLingo	0.3832 (89.2%)	0.3091 (70.4%)

The results of experiments are shown in Table 4, indicating that Reverso provided the best translation for the French collection and Online for the Portuguese corpus. With the FreeTranslation system, these three MT systems usually obtained satisfactory retrieval performances for both languages. For French, the WorldLingo, BabelFish, or Google translation systems also worked well.

Finally, Table 5 lists the parameter settings used for our best performing official runs in the bilingual task. Based on our previous experiments [7], we

Table 5. Description and MAP of our Best Official Bilingual Runs

From EN to ...	French 49 queries	French 49 queries	Portuguese 50 queries	Portuguese 50 queries
IR 1 (doc/term)	PL2 (10/30)	PL2 (10/30)	I(n)L2 (10/40)	GL2 (10/40)
IR 2 (doc/term)	LM (10/30)	Okapi (10/60)	LM (10/30)	Okapi (10/80)
IR 3 (doc/term)		LM (10/50)		LM (10/40)
Data fusion	Z-score	Z-score	Round-robin	Z-score
Translation tools	BabelFish & Reverso	Reverso & Online	Promt & Free & Online	Promt & Free
MAP	**0.4278**	0.4256	0.4114	**0.4138**
Run name	UniNEbifr1	UniNEbifr2	UniNEbipt1	UniNEbipt2

first concatenated two or three query translations obtained by different freely available translation tools. Before combining the result lists obtained by various search models, we automatically expanded the translated queries using a pseudo-relevance feedback method (Rocchio), as described in Table 5.

5 French Robust Track

The aim of this track is to analyze and to improve IR system performance when processing difficult topics [11], or queries from previous CLEF evaluation campaigns that have poor MAP. The goal of the robust track is therefore to explore various methods of building a search system that will perform "reasonably well" for all queries. In real systems this is an important concern, particularly when evaluating situations where the search engine returns unexpected results or "silly" responses to users.

In this track we reused queries created and evaluated during the CLEF 2001, 2002, and 2003 campaigns, with topic collections being the same for the most part. Moreover, the organizers arbitrarily divided this query set into a training set (60 queries) and a test set (100 queries). In the latter, 9 queries did not in fact obtain any relevant items and thus the test set only contained the 91 remaining queries. When analyzing this sample, we found that the mean number of relevant items per query was 24.066 (median: 14, minimum: 1, maximum: 177, standard deviation: 30.78).

When using the MAP to measure the retrieval effectiveness, all observations (queries) had the same importance. This arithmetic mean thus did not really penalize incorrect answers. Thus, Voorhess [11] and other authors suggested replacing the arithmetic mean (MAP) with the geometric mean (GMAP), in order to assign more importance to the poor performances obtained by difficult topics (both measures are depicted in Table 6).

Given our past experience, we decided to search the French collection using the three probabilistic models described in Section 2, as well as blind query expansion. As depicted in Table 6, the MAP resulting from these three models when applying the TD or T query formulations are relatively similar. These

Table 6. Description of our Official Robust Runs (French corpus, 91 queries)

Run name	Query	Model	Query exp.	MAP	comb. MAP	GMAP
UniNEfrr1	TD	Okapi	5 docs / 15 terms	0.5035	Round-Robin	
	TD	GL2	3 docs / 30 terms	0.5014		
	TD	LM	10 docs / 15 terms	0.5095	0.5227	**0.3889**
UniNEfrr2	T	Okapi	3 docs / 10 terms	0.4058	Z-score	
	T	GL2	5 docs / 30 terms	0.4029		
	T	LM	5 docs / 10 terms	0.4137	0.4396	**0.2376**
UniNEfrr3	TD	GL2	3 docs / 30 terms & Yahoo!.fr	0.4607		0.2935

result lists were then combined using a data fusion operator. This procedure was applied to two of our official runs, namely UniNEfrr1 with TD query, and UniNEfrr2 with T query (complete description given in Table 6).

For the last run (UniNEfrr3) we submitted the topic titles to Yahoo.fr search engine. The response page contained ten references, plus a short description. We then extracted these ten short textual descriptions and added them to the original query. The expanded query was then sent to our search model in order to hopefully obtain better results. With the TD query formulation, the mean number of distinct search terms was 7.51, and when including the first ten references retrieved by Yahoo!, the average value increased to 115.46 (meaning that we have added, in mean, 108 new search terms). This massive query expansion did not prove to be effective (MAP: 0.5014 before, and 0.4607 after query expansion using Yahoo! snippets) and in our efforts to improve retrieval performance we would most certainly need to include a term selection procedure.

6 Conclusion

During the CLEF 2006 evaluation campaign, we proposed a more effective search strategy for the Hungarian language. In an attempt to remove the more frequent derivational suffixes were applied a more aggressive stemmer, and we also evaluated an automatic decompouding scheme. Combining different indexing and retrieval schemes seems to be really effective for this language, although more processing time and disk space are required.

For the French, Portuguese/Brazilian and Bulgarian languages, we used the same stopword lists and stemmers developed during the previous years. In order to enhance retrieval performance, we implemented an IR model based on the language model and suggested a data fusion approach based on the Z-score, after applying a blind query expansion.

In the bilingual task, the freely available translation tools performed at a reasonable level for both the French and Portuguese languages (based on the best translation tool, compared to the monolingual search the MAP is around 95% for French and 83% for Portuguese/Brazilian). Finally, in the robust retrieval task,

we investigated some of the difficult topics, plus various methods that might be implemented to improve retrieval effectiveness.

Acknowledgments. The authors would like to thank Pierre-Yves Berger for his help in translating the English topics and in using the Yahoo.fr search engine. This research was supported by the Swiss NSF under Grants #200020-103420 and #200021-113273.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal 7, 121–148 (2004)
2. Savoy, J.: Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. ACM Transactions on Asian Languages Information Processing 4, 163–189 (2005)
3. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings TREC-4, Gaithersburg, pp. 25–48 (1996)
4. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management 36, 95–108 (2002)
5. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20, 357–389 (2002)
6. Hiemstra, D.: Using Language Models for Information Retrieval. Ph.D. Thesis (2000)
7. Savoy, J., Berger, P.-Y.: Monolingual, Bilingual, and GIRT Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 131–140. Springer, Heidelberg (2006)
8. Savoy, J.: Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 322–336. Springer, Heidelberg (2004)
9. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management 33, 495–512 (1997)
10. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. IR Journal 1, 151–173 (1999)
11. Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track. In: Proceedings TREC-2004, Gaithersburg, pp. 70–79 (2005)