Published in Lecture Notes in Computer Science 5706, 178-185, 2009
which should be used for any reference to this work

UniNE at CLEF 2008: TEL, and Persian IR

Ljiljana Dolamic, Claire Fautsch, and Jacques Savoy

Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2009 Neuchatel, Switzerland
{Ljiljana.Dolamic,Claire.Fautsch,Jacques.Savoy}@unine.ch

Abstract. In our participation in this evaluation campaign, our first objective was to analyze retrieval effectiveness when using The European Library (TEL) corpora composed of very short descriptions (library catalog records) and also to evaluate the retrieval effectiveness of several IR models. As a second objective we wanted to design and evaluate a stopword list and a light stemming strategy for the Persian (Farsi), a member of the Indo-European family of languages and whose morphology is more complex than of the English language.

1 Introduction

During the last few years, the IR group at University of Neuchatel has focused on designing, implementing and evaluating IR systems for various natural languages, including European [1] and popular Asian languages (namely, Chinese, Japanese, and Korean). The main objective of our work is still to promote effective monolingual IR in many different natural languages.

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the TEL corpus used in the CLEF-2008 ad hoc track. Section 3 outlines the main aspects of the various IR models used with TEL collections as well as an evaluation of our official runs and certain related experiments. Section 4 presents the principal features of the Persian (Farsi) language along with the stopword list and stemming strategy developed for this language, and describes our official runs for this task.

2 Overview of TEL Corpus

In a certain sense this first ad hoc task takes us back to our research roots, due to the need to look for relevant items among the card catalog on the collection located at The European Library (TEL) (see `www.TheEuropeanLibrary.org`). This collection includes three sub-collections, one in the English language (from British Library), the second in German (Austrian National Library) and the third in French (Bibliothèque nationale de France). The real challenge in our work is to retrieve pertinent records through relying on very short catalog descriptions on the information items involved. In many of these record items

the only information contained is the title (under the tag <title>) and author, plus manually assigned subject headings (tag <subject>). Other records may however contain a short description of the object (tags <description> and <alternative>). Each record may of course contain other fields not used during the indexing process such as language, document identification, author, publisher, location, issue, date, etc. For more information, see [2].

The average size of each topic description is relatively short (between 10 and 16 terms), and is similar for all three languages (perhaps a bit longer for the French corpus). The descriptors are subdivided into title (T), descriptive (D) and narrative (N) logical sections, and from them we automatically removed certain phrases such as "Relevant document report ..." or "Relevante Dokumente berichten ...", etc. All our runs were fully automatic.

The available topics cover various subjects (e.g., Topic #500: "Gauguin and Tahiti," Topic #468: "Modern Japanese Culture," Topic #471: "Watchmaking," or Topic #477: "Web Advertising", etc.). While topic descriptions do not generally contain many proper names (creators and their works), we found two topics containing personal names ("Henry VIII" and "Gauguin"), and 23 with geographical names (e.g., "Europe," "Eastern," "Bordeaux" or "Greek"). Expressions referring to the Untied States of America are not standardized and may for example take the form "USA," "North America," or "America." Also, time periods are infrequently used (in 7 topics only), with many including expressions that are fairly broad (e.g., "Modern," or "Roman"), while others are more precise ("World War I").

3 IR Models and Evaluation

An essential element in our indexing strategy was the use a stopword list to denote very frequent word forms having no important impact matching topic and document representatives (e.g., "the," "in," "or," "has," etc.). In our experiments the stopword list contained 589 English, 484 French and 578 German terms, and diacritics were replaced by their corresponding non-accented equivalent. Another element was the use of light stemmers developed for the French and German languages, wherein inflectional suffixes attached only to nouns and adjectives were removed. This resulted in more effective retrieval than do more aggressive stemmers that also remove derivational suffixes [3]. These stemmers and stopword lists are freely available at the Web site `www.unine.ch/info/clef`. For the English language we tried both a light stemmer (the S-stemmer proposed by Harman [4] to remove only the plural form '-s') and a more aggressive version [5] based on a list of around 60 suffixes.

In the German language compound words are widely used and present some specific challenges. For example the compound noun "Forschungsprojekt" can be divided into "Forschung" + 's' + "Projekt" (research + project), and the augment (i.e. the letter 's' in our example) is not always present (e.g., "Bankangestelltenlohn" combines "Bank" + "Angestellten" + "Lohn" (salary)). Given the fairly wide use of compound constructions in German and their many different forms,

an effective IR system must include an automatic decompounding procedure. The automatic one used in our experiments [1] leaves both the compound form and its composite parts in both the topic and document representatives.

In an effort to obtain high MAP values we considered adopting different weighting schemes for the terms found in documents or queries. This would thus allow us to account for term occurrence frequency (denoted tf), inverse document frequency (denoted *idf*) as well as the document length. In the following experiments we considered the classical $tf \cdot idf$ formulation (with the cosine normalization), as well as probabilistic models such as the Okapi (or BM25) and variants derived from the DFR (*Divergence from Randomness*) family of models. Finally we also implemented a statistical language model (LM) known as a non-parametric probabilistic model (Okapi and DFR are considered as parametric models). For specific details on these IR models, see [6].

To measure retrieval performance we used the mean average precision (MAP) obtained from 50 queries. The best performance obtained under a given condition is shown in bold type in the following tables. We then applied the bootstrap methodology in order to statistically determine whether or not a given search strategy would be better than the performance depicted in bold. Thus, in the tables in this paper we added an asterisk to indicate any statistically significant differences resulting from the use of a two-sided non-parametric bootstrap test ($\alpha = 5\%$).

Table 1 shows the MAP obtained by various probabilistic models for the English collection, using two different query formulations (T or TD) and two stemmers. The last two columns show the MAP obtained when applying our light stemmer to the French corpus. An analysis of this data shows that the best performing IR model was usually the DFR-$I(n_e)B2$ or DFR-PB2 formulation (English corpus, T queries). For the English corpus with the Porter stemmer and TD query formulation, the LM model performed slightly better (0.3701 vs. 0.3643, a statistically non-significant difference).

Table 1. MAP of Various IR Models and Query Formulations (English & French TEL Corpus)

	Mean Average Precision					
Query Stemmer	English T S-stem.	English TD S-stem.	English T Porter	English TD Porter	French T	French TD
Okapi	0.2795*	0.3171*	0.3004*	0.3329*	0.2659*	0.2998*
DFR-PB2	**0.3076**	0.3540	**0.3263**	0.3646	0.2734	0.3103*
DFR-GL2	0.2935*	0.3300*	0.3125*	0.3478*	0.2734	0.3117*
DFR-$I(n_e)B2$	0.3075	**0.3541**	0.3258	0.3643	**0.2825**	**0.3291**
LM	0.3029	0.3527	0.3180	**0.3701**	0.2747	0.3201
$tf\ idf$	0.1420*	0.1783*	0.1600*	0.1871*	0.1555*	0.1821*
% over T		+14.6%		+12.4%		+14.7%
% over S-stem.			+6.2%	+4.2%		

Table 2. MAP of Various IR Models and Query Formulations (German TEL Corpus)

	Mean Average Precision			
Query	German T	German TD	German T	German TD
Decompounding?	no	no	yes	yes
Okapi	0.1462*	0.1872*	0.2188*	0.2522*
DFR-PB2	**0.1635**	**0.2097**	0.2193	0.2555
DFR-GL2	0.1462*	0.1878*	0.2309	0.2615*
DFR-$I(n_e)B2$	0.1606	0.2071	0.2248	0.2615
LM	0.1529	0.1972*	**0.2361**	**0.2697**
tf idf	0.1105*	0.1382*	0.1312*	0.1598*
% over T		+28.5%		+15.1%
% over no decomp.			+46.8%	+31.5%

The second last line shows the percentage variations derived from comparing results with the short (T) query formulation, and the last line the performance difference obtained using the S-stemmer. As indicated, increasing query size improves the MAP (around +12.4% to +14.7%). Statistically, when using the MAP obtained by T query formulation as baseline, the TD query format always improves retrieval performance significantly.

According to the MAP, the best indexing seemed to be the stemming technique using Porter's approach. In this case, the MAP with TD query formulation and Porter's stemmer increased by about 4.2% compared to the S-stemmer. Applying our statistical test when comparing the S-stemmer with Porter's approach, only three cases had statistically significant performance differences (underlined in Table 1).

Table 2 shows the MAP obtained with the probabilistic models and with two query formulations (T or TD) to the German collection, and comparing performances with and without our automatic decompounding approach. The best IR models seemed to be the DFR-PB2 (without decompounding) or the LM with our decompounding scheme. By adding terms to the topic descriptions, we could improve MAP (between 15.1% to 28.5%), although the performance differences were never statistically significant. Comparing the average performances shows that applying an automatic decompounding approach improved retrieval effectiveness, on average by 46.8% for short query formulations compared to +31.5% for TD queries) (see last line of Table 2). When analyzing the performance of various models, the differences were usually statistically significant (MAP underlined in Table 2).

An analysis showed that pseudo-relevance feedback (or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [7] (denoted "Roc." in the following tables 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. From previous experiments we learned that this type of blind query expansion strategy does not always work well. More particularly, we believe that including terms

Table 3. Description and MAP of Our Best Official TEL Runs

Language	Index	Query	Model	Query expansion	MAP	MAP
English	Porter	TD	Okapi		0.3329	Z-score
UniNEen3	S-stem	TD	$I(n_e)B2$		0.3541	**0.3754**
	Porter	TD	LM	Roc. 5 docs/10 terms	0.3913	
French	stem	TD	Okapi		0.2998	Z-score
UniNEfr3	stem	TD	$I(n_e)B2$		0.3291	**0.3327**
	stem	TD	LM	Roc. 5 docs/10 terms	0.3150	
German	decomp.	TD	Okapi	idf 5 docs/10 terms	0.2302	Z-score
UniNEde1	word	TD	GL2	Roc. 5 docs/20 terms	0.2356	**0.3013**
	decomp.	TD	$I(n_e)B2$	Roc. 5 docs/50 terms	0.2757	

occurring frequently in the corpus (because they also appear in the top-ranked documents) may introduce additional noise, and thus be ineffective in discriminating between relevant and non-relevant items. We thus decided to also apply our idf-based query expansion model [8] (denoted "idf" in following tables).

It is usually assumed that combining result lists computed by different search models (data fusion) could improve retrieval effectiveness [9]. Thus in this study we combined three probabilistic models representing both the parametric (Okapi and DFR) and non-parametric (language model or LM) approaches. To produce a combination such as this we evaluated various fusion operators and thus we suggest the "Z-score" approach which applies a normalization procedure to each result list before combining the different document scores (see details in [1]).

4 IR with Persian (Farsi) Language

The Persian (or Farsi) language is a member of the Indo-European family and has relatively few morphological variations. This year we used a corpus comprising Hamshahri newspapers from 1996 to 2002 (611 MB). It contains exactly 166,774 documents covering various subjects (politics, literature, art and economics, etc.) and comprises 448,100 different words. Article size varies between 1 KB and 140 KB and include on average about 202 tokens (127 when counting the number of distinct word types). The corpus is coded in UTF-8 and its alphabet has 28 Arabic letters plus an additional 4 letters used in Persian ("گ" "چ" "ژ" "پ").

We began by building a Persian stopword list containing 884 terms. Unlike most others lists, it contains the collection's most frequently occurring words (determinants, prepositions, conjunctions, pronouns or certain auxiliary verb forms), plus a large number of suffixes already separated from word stems in the collection. Note that that the Persian language does not include definite (the) or indefinite (a, an) articles (indefinite articles are indicated by a suffix ("ی" or simply by "one").

As a stemming strategy we used either a morphological analysis [10] or our simple, fast and light stemmer. It removes only nouns and adjective inflections

(number and case only, since Persian does distinguish gender). The general pattern is the following: <possessive> <plural> <other-suffix> <stem>.) In our light stemmer we usually remove possessive, plural and certain suffixes marked as others. The following examples from our light stemmer illustrate certain aspects of the Persian morphology. From the plural form "د ر خ ت ا ن" ("trees"), we can obtain "د ر خ ت" ("tree"). The plural is usually denoted by either "ا ه" (inanimate) or by "ا ن" or "ا ه" (animate nouns). The plural forms for words borrowed from Arabic usually apply the language's own plural formation rule, and in Persian there are certain irregular formations similar to "mouse/mice". For the possessive form, "د س ت م" ("my hand"), our stemmer returns "د س ت" ("hand"). For the form "ا ي ر ا ن ي ا ن" ("Iranians") we remove both the plural and the derivational suffixes to obtain "ا ي ر ا ن" ("Iran"). In this corpus we saw certain circumstances where suffixes might be written together or separated from the word (e.g., "ا ه"). Adjectives are usually indeclinable whether used attributively or as a predicate. When used as substantives, adjectives take the normal plural endings, while comparative and superlative forms use the endings "ر ت" and "ن ی ر ت" .

Unlike the Latin, German or Hungarian languages, Persian uses few case markers (other than the accusative case and certain specific genitive cases). The genitive case may also be expressed by coupling two nouns using the particle known as ezafe (e.g., "د ر م س پ" "man's son"). As usually done in English, other relations are expressed using prepositions (e.g., in, with, etc.).

Table 4 shows the MAP obtained by various probabilistic models using the Persian collection, and two different query formulations (T or TD), two stemmers and two indexing strategies (word or 4-gram). Since in documents (and queries) inflectional suffixes are usually clearly delimited (presence of a small space), applying our light stemmer or ignoring stemming does not lead to significantly different retrieval performances. Adding more terms in query formulations improves the MAP (between 4.8% to 14.6%) and the performance differences are usually statistically significant. The use of words as indexing units tends to

Table 4. MAP of Various IR Models and Query Formulations (Persian Corpus)

	Mean Average Precision					
Query	T	TD	T	TD	T	TD
Stemmer	none	none	light	light	4-gram	4-gram
Okapi	0.4065*	0.4266	0.4092*	0.4292*	0.3965*	0.4087*
DFR-PL2	0.4078*	0.4274	0.4120	0.4335	0.3815*	0.4005*
DFR-$I(n_e)C2$	**0.4203**	**0.4351**	**0.4204**	**0.4376**	**0.4127**	**0.4235**
LM	0.3621*	0.3839*	0.3607*	0.3854*	0.3248*	0.3518*
tf idf	0.2727*	0.2824*	0.2717*	0.2838*	0.2608*	0.2700*
% over T		+4.8%		+5.2%		+14.6%
% over "none"			+0.4%	+0.8%	-5.1%	-5.3%

Table 5. Description and MAP of Our Official Persian Monolingual Runs

Language	Index	Query	Model	Query expansion	MAP	MAP
UniNEpe1	word	T	PL2		0.4078	Z-score
	4-gram	T	LM	idf 10 docs/20 terms	0.3783	0.4675
	word	T	Okapi	Roc. 10 docs/20 terms	0.4376	
UniNEpe2	4-gram	TD	$I(n_e)C2$		0.4235	Z-score
	word	TD	PL2		0.4274	**0.4898**
	stem	TD	PL2	idf 10 docs/20 terms	0.4513	
	word	TD	PL2	Roc. 10 docs/20 terms	0.4311	
UniNEpe3	4-gram	TD	Okapi	Roc. 5 docs/100 terms	0.4335	Z-score
	word	TD	LM	idf 10 docs/70 terms	0.4141	0.4814
	word	TD	PL2		0.4274	
UniNEpe4	4-gram	TDN	LM	idf 10 docs/100 terms	0.3738	Z-score
	word	TDN	LM	Roc. 10 docs/20 terms	0.4415	**0.4807**
	word	TDN	PL2		0.4425	

produce better MAP and in certain cases, as underlined in Table 4, performance differences are fairly significant.

Table 5 shows the exact specifications of our four official monolingual runs for IR evaluation task for Persian, based mainly on the three probabilistic models (Okapi, DFR and statistical language model (LM)). The strategy we followed consisted of combining different indexing units (words, stems, and 4-grams), based on various probabilistic IR models (Okapi or DFR) and using three different blind-query expansion techniques (Rocchio, idf-based or none). As for the TEL runs (see Table 3) we suggest combining these probabilistic models using the "Z-score" approach (see details in [1]). Of course other methods can be applied to combine these ranked lists as for example the round-robin (RR), taking the sum of the different document scores (SUM) or sum these scores after normalizing (NormMax) (e.g., divided them by the max). If we consider our first official run (UniNEpe1), the MAP achieved with the "RR" approach is 0.4376, the "SUM" method produces a MAP of 0.4413, the "NormMax" 0.4639 and the "Z-score" 0.4675. The performance differences with the "Z-score" are always significant, at least for this run.

5 Conclusion

In this ninth CLEF campaign we evaluated various probabilistic IR models using two different test collections. The first was composed of short bibliographic notices extracted from the TEL corpora (written in the English, German and French) and the second containing newspapers articles written in Persian. For the latter we also suggested a stopword list and a light stemming strategy.

The results of our various experiments demonstrate that the $I(n_e)B2$ or PB2 models (or $I(n_e)C2$ for the Persian language) derived from the Divergence from Randomness (DFR) paradigm and the LM model seem to provide the best overall

retrieval performances (see Tables 1, 2, and 4). The Okapi model used in our experiments usually results in retrieval performances inferior to those obtained with the DFR or LM approaches. A data fusion strategy may however enhance the retrieval performance for the French and German (see Tables 3) or Persian languages (see Table 5), but not for the English corpus.

For the Persian language (Table 4), our light stemmer tends to produce better MAP than does the 4-gram indexing scheme (relative difference of 5.5%). For an approach ignoring a stemming stage the performance difference is however is rather small. Finally Persian new words can be formed using compound construction (e.g., handgun), yet retrieval effectiveness obtained by applying automatic decompounding procedures remains unknown.

Acknowledgments. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal 7, 121–148 (2004)
2. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2008: Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 15–37. Springer, Heidelberg (2009)
3. Savoy, J.: Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages. In: Proceedings ACM-SAC, pp. 1031–1035. ACM Press, New York (2006)
4. Harman, D.K.: How Effective is Suffixing? Journal of the American Society for Information Science 42, 7–15 (1991)
5. Porter, M.F.: An algorithm for Suffix Stripping. Program 14, 130–137 (1980)
6. Dolamic, L., Savoy, J.: Stemming Approaches for East European Languages. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 37–44. Springer, Heidelberg (2008)
7. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings TREC-4, pp. 25–48. Gaithersburg (1996)
8. Abdou, S., Savoy, J.: Searching in Medline: Stemming, Query Expansion, and Manual Indexing Evaluation. Information Processing & Management 44, 781–789 (2008)
9. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. IR Journal 1, 151–173 (1999)
10. Miangah, T.M.: Automatic Lemmatization of Persian Words. Journal of Quantitative Linguistics 13, 1–15 (2006)