UniNE at CLEF 2012

Mitra Akasereh, Nada Naji, Jacques Savoy

Computer Science Dept., University of Neuchatel,
Rue Emile Argand 11, 2000 Neuchatel, Switzerland
{mitra.akasereh, nada.naji, jacques.savoy}@unine.ch

Abstract. As participants in this CLEF evaluation campaign, our first objective is to propose and evaluate various indexing and search strategies for the CHiC corpus, in order to compare the retrieval effectiveness across different IR models. Our second objective is to measure the relative merit of various stemming strategies when used for the French and English monolingual task in the CH context. Our third objective is to assess the effectiveness of query translation methods in a bilingual retrieval. To do so we evaluated the CHiC test-collections using Okapi, various IR models derived from the *Divergence from Randomness* (DFR) paradigm together with the *dtu-dtn* vector-space model. We also evaluated different pseudo-relevance feedback approaches. In the bilingual task, we conducted our search on the English corpus using the French and German topics with two different translations for each of them. For both English and French languages, we find that word-based indexing with our light stemming procedure results in better retrieval effectiveness than with other strategies. When ignoring stemming, the performance variations were relatively small yet for the French corpus better than applying a light stemmer. In bilingual level results show that using a combination of translation resources gives better results than a single source.

Keywords: Probabilistic IR Models, Stemmer, Data Fusion, Cultural Heritage, bilingual IR

1   Introduction

Cultural heritage can be defined as any handmade substance or intangible feature remained from previous societies. It can refer to any artefacts, built or natural environments, traditions and languages, etc. The developing use of digital information challenges the cultural heritage organizations to provide cultural heritage collections in electronic format. The data may come from different sources (libraries, archives, museums, audiovisual archives, books, journals, etc.), in various languages and formats. These digital libraries should not only be created but also properly managed and assessed in order to bring the maximum utility to their users. As yet no proper evaluation approaches are available and there is work to be done in this area. The goal of Cultural Heritage in CLEF (CHiC) evaluation lab is thus providing a systematic and large-scale evaluation of cultural heritage digital libraries.

The IR group of university of Neuchâtel focuses, as one of its main tasks, on design, implementation and evaluation of various indexing and search strategies for a set of different natural languages. Up to this point we achieved to provide a groundwork for evaluation and comparison of different tools for monolingual IR, in different languages, using generic test-collections (e.g., newspaper articles). Our second goal is to evaluate different tools considering only a specific field of knowledge in order to integrate domain specific search into our system. The aim here is to be able to evaluate the impact of document structure and query formulation on retrieval effectiveness in order to study the possibilities to improve the search quality in a domain specific search. As a third objective we also want to integrate translation into the search process and adapting our system for bilingual and multilingual IR. Accordingly reaching these objectives has been our main motive to participate in CHiC evaluation lab at CLEF 2012.

The rest of this report is organized as follows: section 2 presents an introduction to our experiment setup. Section 3 describes the results obtained during the experiment and the related analysis. Section 4 shows our official runs and finally section 5 concludes the experiment.

2   Experiment Setup

2.1   Overview of the Task

In our participation in CHiC we worked on the *ad-hoc* retrieval task. This task is a standard retrieval task in which retrieval effectiveness for individual queries is assessed. At this level the only authorized user/system interaction would be blind-query expansion technics. The expected output is a ranked list of retrieved documents for each query. The task is covering monolingual, bilingual and multilingual subtasks in English, French and German. In our experiment we worked on monolingual English and French retrieval as well as bilingual retrieval in which we worked with French and German topics to be searched on the English corpus.

2.2   Overview of the Test-collection

The corpus used in CHiC test-collection is extracted from Europeana (www.europeana.eu) and is offered in 3 major European languages, namely English (EN), French (FR) and German (DE). Europeana is a digitized collection of Europe's cultural and scientific heritage. It provides access over 23 million objects such as books, paintings, films, museum objects, etc. collected from more than 2200 institutions in 33 countries. Europeana collection is cross-domain and in multiple languages. The documents metadata is mapped to a single data model. Each document consists of elements providing brief descriptions of the objects (title, keywords, description, date, provider, etc.). It is worth-mentioning that some documents contain less of these tags than other ones which sometimes leaves them with very poor content. As far as our experiment is concerned, only human-readable informative texts are of use.

The English corpus consists of 1,106,426 documents while the French one has 3,635,388 ones. A sample of both French and English documents is shown in Figures 1 and 2.

```
<ims:metadata
    ims:identifier="http://www.europeana.eu/resolve/record/10105/662DC5085397837
    C8C8891836EA6431C4A477CB2"ims:namespace="http://www.europeana.eu/"
    ims:language="eng">
 <ims:fields>
    <dc:identifier>Orn.0446</dc:identifier>
    <dc:subject>Australian Pelican</dc:subject>
    <dc:title>Australian Pelican (Orn.0446)</dc:title>
    <dc:type>mounted specimen</dc:type>
    <europeana:country>malta</europeana:country>
    <europeana:dataProvider>Heritage Malta</europeana:dataProvider>
    <europeana:isShownAt>http://www.heritagemalta.org/sterna/orn.php?id=0446
    </europeana:isShownAt>
    <europeana:language>en</europeana:language>
    <europeana:provider>STERNA</europeana:provider>
    <europeana:type>IMAGE</europeana:type>
    <europeana:uri>
    http://www.europeana.eu/resolve/record/10105/662DC5085397837C8C8891
    836EA6431C4A477CB2</europeana:uri>
 </ims:fields>
</ims:metadata>
```

Fig. 1. Example of an English document

```
<ims:metadata
    ims:identifier="http://www.europeana.eu/resolve/record/91401/6BA09455082
C5E65E39C59DC30A0A966C235FBAE"ims:namespace="http://www.europeana.eu/"ims
    :language="fre">
 <ims:fields>
    <dc:identifier>oai:ircam.fr:grove:41112</dc:identifier>
     <dc:identifier>http://www.grovemusic.com/shared/views/article.html?section=music.
        41112
    </dc:identifier>
    <europeana:uri>
    http://www.europeana.eu/resolve/record/91401/6BA09455082C5E65E39C59DC30A0
        A966C235FBAE</europeana:uri>
    <dc:title>Biography. Lejla Agolli</dc:title>
    <dc:contributor>Macy, Laura (éditeur)</dc:contributor>
    <dc:type>text</dc:type>
    <dc:type>biography</dc:type>
    <dc:type>biographie</dc:type>
    <dc:type>document numérique</dc:type>
    <europeana:type>TEXT</europeana:type>
    <dc:publisher>Grove Music Online</dc:publisher>
    <dc:language>English</dc:language>
    <dc:language>eng</dc:language>
    <dc:description>Cette biographie n'est accessible qu'à partir de postes de consulta-
```

```
                  tion d'organismes abonnés au Grove Music Online.</dc:description>
         <dc:description>Biographie de Lejla Agolli</dc:description>
         <dc:subject>Agolli, Lejla (compositeur) :</dc:subject>
          <dc:source>http://www.musiquecontemporaine.fr/record/oai:ircam.fr:grove:41112</
             dc:source>
         <europeana:isShownAt>
          http://www.grovemusic.com/shared/views/article.html?section=music.41112
         </europeana:isShownAt>
         <dc:rights>internet</dc:rights>
         <dcterms:provenance>Ircam - Centre Pompidou</dcterms:provenance>
         <europeana:country>france</europeana:country>
         <europeana:provider>IRCAM-Institut de Recherche et Coordination
           Acoustique/Musique</europeana:provider>
         <europeana:language>fr</europeana:language>
       </ims:fields>
     </ims:metadata>
```

Fig. 2. Example of a French document

For the *ad-hoc* task there are 50 very short topics. These topics are mostly named entities (people, places and works) and they mainly extracted from Europeana queries logs. Thus they convey the real user's information needs in a cultural heritage search context. Among the 50 French topics, 11 have no relevant documents in the collection. This number grows to 14 for the English topics. One topic from each language is shown in Figure 3. As shown in the sample below each topic consists of a title and, sometimes, a description of the content. Even though, the only field that should be used for retrieval is the title.

```
<topic lang="en">
  <identifier>CHIC-006</identifier>
  <title>esperanto</title>
  <description>Constructed international auxiliary language</description>
</topic>

<topic lang="fr">
  <identifier>CHIC-004</identifier>
  <title>film muet</title>
  <description />
</topic>

 <topic lang="de">
  <identifier>CHIC-025</identifier>
  <title> amerikanische sklaverei </title>
  < description />
 </topic>
```

Fig. 3. Sample of English, French and German topics

2.3 Indexing Strategies

In our experiment we applied a stopword removal along with a light stemmer for both English and French corpora. Our stopword list for English contains 571 terms while the French one has 464 terms. These tools are freely available at members.unine.ch/jacques.savoy/clef/. These lists are composed of terms having a high frequency such as determinants, prepositions, conjunctions, pronouns, and some verbal forms which convey no important meaning. The light stemmer that we used for English removes only the plural '-s' and is called S-stemmer [1]. The stemmer for French removes the inflectional suffixes from plural and feminine forms of the words [2].

Our choice of these light stemmers is based on previous experiments which show that light stemmers tend to be as effective as stemmers based on morphological analysis [1], [3], [4]. Moreover applying stemming would not be a good manner to achieve high precision which is the aim in this experiment [5].

2.4 IR Models

In our experiments we tried different weighting schemes in order to compare them and define the most effective ones in terms of achieving a high precision. First we picked the *dtu-dtn* model [6] as an effective vector-space model. Second, as probabilistic models, we used the Okapi (BM25) [7]. Then we tried three other probabilistic models extracted from the *Divergence from Randomness* (DFR) family [8], namely DFR-PL2, DFR-I(n_e)C2, and DFR-I(n_e)B2. The indexing weight (weight of term t_j in document d_i) in these models is computed as shown in Table 1.

Table 1. Formulas used in different models for assigning indexing weight

Okapi	$w_{ij} = \dfrac{((k_1 + 1) \cdot tf_{ij})}{(k + tf_{ij})}$ $\qquad$ $K = k_1 \cdot \left[(1-b) + b \cdot \dfrac{l_i}{avdl} \right]$
	l_i is the length of document d_i and *avdl* is the average document length.
dtu-dtn	Indexing weight for document terms (dtu): $$w_{ij} = \frac{(1 + \ln(1 + \ln(tf_{ij}))) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$$ Indexing weight for query terms (dtn): $$w_{ij} = (1 + \ln(1 + \ln(tf_{ij}))) \cdot idf_j$$

DFR	$w_{ij} = Inf_{ij}^1 \cdot Inf_{ij}^2 = -\log_2(\text{Prob}_{ij}^1(tf_{ij})) \cdot (1 - \text{Prob}_{ij}^2(tf_{ij}))$

DFR- I(n_e)C2:

$$Inf_{ij}^1 = tfn_{ij} \cdot \log\left[\frac{n+1}{n_e + 0.5}\right] \qquad \text{Prob}_{ij}^2 = 1 - \frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}$$

DFR-I(n_e)B2:

$$Inf_{ij}^1 = tfn_{ij} \cdot \log_2[(n+1)/(n_e + 0.5)]$$
$$Prob_{ij}^2 = 1 - \left[(tc_j + 1)/(df_j \cdot (tfn_{ij} + 1))\right]$$

DFR-PL2:

$$\text{Pr}ob_{ij}^1 = \frac{e^{-\lambda_j} \cdot \lambda_j^{tfn_{ij}}}{tfn_{ij}!} \qquad \text{Pr}ob_{ij}^2 = \frac{tfn_{ij}}{tfn_{ij} + 1}$$

— tf_n is the normalized term frequency.

— $\lambda_j = \dfrac{tc_j}{n}$ (tc_j is the number of occurrence of term t_j in the collection and n is the size of the elite set.)

— $n_e = n \cdot (1 - (\dfrac{n-1}{n})^{tc_j})$

— $tfn_{ij} = tf_{ij} \cdot \log_2(1 + \dfrac{c \cdot mean\_dl}{l_i})$ (c and $mean\_dl$ (average document length)

2.5   Evaluation

For evaluating the retrieval performance we chose the MAP (mean average precision) measure. This is computed with the TREC_EVAL program where MAP value is computed based on, maximum, 1000 retrieved items per query. It is important to mention that when computing the MAP, the topics with no relevant items are not taken into account (14 topics among the English topics and 11 French ones).

2.6   Pseudo-Relevance Feedback

In order to enhance the retrieval effectiveness we also applied a blind-query expansion to our test. Our previous experiments on other corpora show that pseudo-relevance feedback (PRF or blind-query expansion) tends to improve the retrieval effectiveness [9]. As a first approach we tried the Rocchio's approach [10] with α = 0.75, β = 0.75. In this method the system expands the query by adding *m* terms selected from the *k* best ranked documents retrieved for the original query. As a second

approach we tried an *idf*-based query expansion model [11]. The reason for trying both approaches is that in some cases adding frequently occurring terms produces noise and consequently Rocchio's approach does not give good results [12].

2.7 Data Fusion

In our experiment we tried to see whether combining different indexing schemes and IR models improves the retrieval effectiveness, as it is supposed to, or not [13]. It is probable that different strategies retrieve the same relevant items in their top ranks rather than the same non-relevant ones. Therefore we consider that by combining different ranked lists, resulting from different IR models, we will gain a list with relevant documents in higher ranks and the non-relevant items in lower ones [14]. In order to produce this combination of ranked lists, different fusion operators can be used. In our study we chose the Z-score scheme which tends to perform the best [14], [15]. More details about the Z-score strategy can be found in [16].

3 Results & Analysis

3.1 Monolingual Retrieval

At monolingual *ad-hoc* task, we test our system using the English and French corpora. Tables 2 and 3 show the Mean Average Precision (MAP) for, respectively, English and French corpora. For both languages, we tried different IR models while applying a light stemmer (Section 2.3) and compared these results with the ones obtained when stemming is ignored. In using the Okapi model the *avdl* (average document length) is set to 181 for English corpus and 169 for the French one, the constant k_1 to 1.2, for both languages, and we tried three different values for the constant *b*: 0.5, 0.7 & 0.9.

Table 2. MAP of different IR models, English corpus

	DFR I(n_e)C2	DFR I(n_e)B2	DFR PL2	Okapi (*b*=0.5)	Okapi (*b*=0.7)	Okapi (*b*=0.9)	dtu-dtn	Avg.
NoStem.	0.4244	0.4524	0.4354	0.4289	0.4207	0.4032	0.4320	0.4281
SStem.	0.4487	**0.4752**	0.4628	0.4560	0.4429	0.4229	0.4484	0.4510
% Change	+5.7%	+5.0%	+6.3%	+6.3%	+5.3%	+4.9%	+3.8%	+5.3%

Table 3. MAP of different IR models, French corpus

	DFR I(n_e)C2	DFR I(n_e)B2	DFR PL2	Okapi (*b*=0.5)	Okapi (*b*=0.7)	Okapi (*b*=0.9)	dtu-dtn	Avg.
NoStem.	0.3520	0.3582	0.3623	**0.3627**	0.3602	0.3497	0.3413	0.3552
LStem.	0.3290	0.3360	0.3392	0.3402	0.3348	0.3253	0.3197	0.3320
% Change	-6.6%	-6.2%	-6.4%	-6.2%	-7.1%	-7.0%	-6.3%	-6.5%

As the results show, for English the corpus with DFR-I(n_e)B2 model we achieve the highest MAP while the best performing model for French is Okapi model (with b=0.5). The results show that applying the light stemmer for the English language improves the effectiveness of the search which is not the case for the French collection. As can be seen in Table 3 we achieved higher MAP while ignoring the stemming phase for the French language. By making a query-by-query analysis on the results we can find some examples where stemming misleads the retrieval. In Topic #21 the title "chanrdonne" (Jacques Chardonne, Writer (F.) Or place in Switzerland) is indexed as "chardon" (after applying the light stemmer) which leads the system to retrieve in its top ranks non-relevant documents (in which "chardon" refers to a flower) such as:

  — Etude de feuilles de echirops, de sphoerophalus, *chardon* cultivé, de *chardon* sauvage de la mer, de fleur lilas, de *chardon* sauvage
  — Sujet ou décor : représentation végétale (fleur, *chardon*) ; *chardon* bleu ; Etude de *chardon* fleuri
  — *Chardon*s sur la côte rocheuse

As another example we can mention Topic #9 for which the title "îles malouines" changes to "malouin" after stemming and results in the retrieval of non-relevant documents (where "Malouin" is a proper name) such as follows in the top ranks:

  — L'Avare, comédie de Molière en 5 actes, mise en vers, par A. *Malouin*
  — villas de la *Malouin*e

Table 4 contains the MAP obtained when applying pseudo-relevance feedback. These results reveal that in this experiment the PRF technic did not help to enhance the retrieval performance. The reason should be due to the fact that in this experiment we are dealing with relatively short documents (having the average number of distinct indexing terms per document at ~54 for English and ~56 for French).

Table 4. MAP of *idf*-based blind-query expansion, English and French queries

		Mean Average Precision	
		English	French
		DFR_I(n_e)B2 SStemmer	Okapi NoStem
		0.4752	0.3627
5 documents	5 terms	0.4382	**0.3488**
	10 terms	0.4315	0.3483
	30 terms	0.3864	0.3428
	50 terms	0.3656	0.3241
	70 terms	0.3606	0.3110
10 documents	5 terms	**0.4557**	0.3432
	10 terms	0.4250	0.3472
	30 terms	0.3923	0.3300
	50 terms	0.3875	0.3283
	70 terms	0.3913	0.3272
15 documents	5 terms	0.4545	0.3329
	10 terms	0.4432	0.3166
	30 terms	0.3981	0.2971
	50 terms	0.3878	0.2947
	70 terms	0.3764	0.2916
20 documents	5 terms	0.4519	0.3404
	10 terms	0.4338	0.3181
	30 terms	0.3962	0.2900
	50 terms	0.3850	0.2864
	70 terms	0.3798	0.2876
25 documents	5 terms	0.4456	0.3439
	10 terms	0.4346	0.3231
	30 terms	0.3901	0.3031
	50 terms	0.3789	0.2641
	70 terms	0.3723	0.2608

In Table 5 we can see the results for our data fusion approach for the English corpus. We can see that the MAP obtained by combining different result lists enhances, in some cases, slightly the performance. However the difference between the MAP obtained for each model separately and the combined one is rather small.

Table 5. MAP of different combinations of IR models, English corpus

English / SStemmer			
Model	Query Expansion (*idf*-based)	Single MAP	Combined MAP Z-Score
DFR-I(n_e)B2		0.4752	0.4715
DFR-PL2		0.4628	
DFR-I(n_e)B2		0.4752	0.4611
DFR-I(n_e)C2	5 documents /10 terms	0.3918	
DFR-I(n_e)B2		0.4752	**0.4758**
dtu-dtn		0.4484	
dtu-dtn		0.4484	**0.4667**
DFR-PL2		0.4628	
DFR-I(n_e)C2		0.4487	**0.4518**
dtu-dtn		0.4484	
DFR-I(n_e)B2	20 documents /10terms	0.4338	**0.4378**
Okapi(*b*=0.9)		0.4229	
dtu-dtn		0.4484	0.4301
DFR-PL2	5 documents /10terms	0.3834	
DFR-I(n_e)C2	20 documents /10terms	0.4074	**0.4238**
dtu-dtn	10 documents /10terms	0.3677	
Okapi(*b*=0.9)		0.4229	
DFR-I(n_e)C2	20 documents /10terms	0.4074	0.4171
dtu-dtn	10 documents /30terms	0.3376	
Okapi(*b*=0.9)		0.4229	

3.2 Bilingual retrieval

In our bilingual retrieval we used the German and French topics to search the English corpus. Our approach was based on query translation (QT). Thus we produced the English translations for German and French topics and then we launched the search on English corpus. To translate the queries we first used two different strategies. First we used Google translation which seems to give reasonable results when dealing with very short query formulation [17]. As a second approach we used the combination of Wikipedia and Google considering that a combination of translation strategies slightly improves the retrieval performance [16]. The results for the bilingual retrieval are shown in Tables 6 and 7. We can see that using the combination of Google and Wikipedia results a better performance even though the difference is not remarkable.

The topics used in this collection are mostly name entities and only the title is used for the search which makes the translation less critical and easier. As a result there are not many differences between translations produced with the two strategies. However, by inspecting the results in details we can find some cases for which a better translation led to better retrievals. In translating Topic #5 ("briefmarke"), from German to

English, Google gives us the word "stamp" versus "postage stamp" which resulted from the Google and Wikipedia combination. As a result the system returns 9 relevant documents among its first 10 ranks when searching "postage stamp" while by searching "stamp" the first relevant document only appears at rank 82. Using the French topics for the same topic ("timbre poste"), Google gives us "stamp post" versus "postage stamp" using the combination method. Here again the system retrieves 9 relevant documents among its first 10 ranks using "postage stamp" while by searching "stamp post" it retrieves 5 relevant documents among its first 10 having the first relevant at rank 5.

Table 6. MAP of different IR models, German topics on English corpus

	DFR I(n_e)C2	DFR I(n_e)B2	DFR PL2	Okapi (b=0.5)	Okapi (b=0.7)	Okapi (b=0.9)	dtu-dtn	Avg.
Google	0.4181	**0.4462**	0.4309	0.4255	0.4101	0.3910	0.4223	0.4206
Google+ Wikipedia	0.4403	**0.4691**	0.4478	0.4322	0.4144	0.4580	0.4459	0.4440
% Change	+5.3%	+5.1%	+3.9%	+1.6%	+1.0%	+17.1%	+5.6%	+5.6%

Table 7. MAP of different IR models, French topics on English corpus

	DFR I(n_e)C2	DFR I(n_e)B2	DFR PL2	Okapi (b=0.5)	Okapi (b=0.7)	Okapi (b=0.9)	dtu-dtn	Avg.
Google	0.3960	**0.4214**	0.4053	0.4006	0.3908	0.3705	0.4100	0.3992
Google+ Wikipedia	0.4096	**0.4346**	0.4197	0.4137	0.4051	0.3861	0.4218	0.4129
% Change	+3.5%	+3.1%	+3.6%	+3.3%	+3.7%	+4.2%	+2.9%	+3.4%

4 Official Results

Table 8 summarizes our twelve official runs. We have submitted four runs for the English monolingual *ad-hoc* task and four French monolingual *ad-hoc* runs. For bilingual *ad-hoc* we submitted two runs using French topics to retrieve English documents and two runs using German topics again on the English corpus. In each run we used our different selected models while applying our light stemmers or alternatively skipping the stemming phase. In some cases we applied a pseudo-relevance feedback strategy [11] to evaluate its impact on the system's performance. We also tried to merge different models into a single ranked list using the Z-score scheme [16] in order to improve the retrieval effectiveness.

Table 8. Description & MAP of our monolingual & bilingual official runs

Run Name	Priority	Language (topic-corpus)	Stemming	Model	Query Expansion	MAP
UnineENEN1	2	EN-EN	SStemmer	DFR-I(n_e)C2		**0.4486**
UnineENEN2	4	EN-EN	NoStem	Okapi (b=0.9)	5docs/10terms	0.3764
UnineENEN3	3	EN-EN	NoStem SStemmer SStemmer	Okapi (b=0.9) DFR-I(n_e)B2 DFR-PL2	5docs/30terms 10docs/30terms 10docs/30terms	0.3826
UnineENEN4	1	EN-EN	NoStem SStemmer	Okapi (b=0.9) DFR-I(n_e)C2	5docs/30terms 5docs/30terms	0.3689
UnineFRFR1	1	FR-FR	NoStem LStemmer	Okapi (b=0.9) DFR-I(n_e)C2	5docs/10terms	0.3572
UnineFRFR2	2	FR-FR	LStemmer	DFR-I(n_e)B2		0.3365
UnineFRFR3	3	FR-FR	LStemmer LStemmer NoStem	DFR-I(n_e)B2 DFR-I(n_e)C2 Okapi (b=0.9)	5docs/10terms 5docs/10terms	**0.3792**
UnineFRFR4	4	FR-FR	LStemmer NoStem LStemmer	DFR-I(n_e)B2 DFR-PL2 DFR-I(n_e)C2	10docs/10terms 10docs/30terms	0.3540
UnineFREN1	1	FR-EN	NoStem SStemmer	Okapi (b=0.9) DFR-I(n_e)C2	5docs/30terms 5docs/30terms	0.3456
UnineFREN2	2	FR-EN	SStemmer	DFR-I(n_e)B2		**0.4346**
UnineDEEN3	3	DE-EN	SStemmer	dtu-dtn		**0.4460**
UnineDEEN2	4	DE-EN	SStemmer	DFR-I(n_e)B2	5docs/10terms	0.4225

5   Conclusion

The results obtained in CLEF 2012 CHiC lab, state that the models derived from the *Divergence from Randomness* (DFR) family, yield the best retrieval effectiveness regardless the underlying language and test-collection. Applying DFR-I(n_e)B2 and DFR-PL2 for both the French and English corpora produced a high MAP compared to other tested models. Our results reveal that the Okapi model (with b=0.5) tends also to be an effective model. The resulting question is to define the best values for the underlying constants.

Our experiment shows that applying a light stemmer (removing only the plural '-s') for English, helps to achieve better results than when the stemming phase is skipped. On the contrary, when using our light stemmer for French (removing plural and feminine suffixes) does not seem to enhance the retrieval performance. A simpler stemmer for the French language may produce a better effectiveness than the applied light stemmer.

Considering the results, we can also conclude that when dealing with relatively short documents, blind-query expansion is not a useful expansion method in order to

improve the retrieval effectiveness. In such cases, it seems difficult to select the most appropriate terms to be included in the expanded query.

Finally, our results from the bilingual search confirm the effectiveness of DFR-I(n_e)B2 model and the S-stemmer (used for English). Furthermore, they show that a combined translation strategy leads to perform better results than a single one. Even though in our experiment, having very short topics (and mostly name entities), the difference between the various translation methods is not remarkable.

References

1. Harman, D.K.: How effective is suffixing? JASIS. 42(1), 7-15 (1991)
2. Savoy, J.: A stemming procedure and stopword list for general French corpora. JASIS. 50, 944-952 (1999)
3. Savoy, J.: Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages. Proceedings ACM-SAC, 1031-1035. The ACM Press, (2006)
4. Fautsch C., Savoy J.: Algorithmic Stemmers or Morphological Analysis: An Evaluation. JASIST. 60, 1616-1624 (2009)
5. Savoy J., Rasolofo Y.: Report on the TREC 11 Experiment: Arabic, Named Page and Topic Distillation Searches. In: Proceedings of the eleventh text retrieval conference TREC-2002, pp. 765–774. NIST Special Publication (2003)
6. Singhal, A.: AT & T at TREC-6. ACM Conference on Research and Development in Information Retrieval, pp. 35-41. ACM/SIGIR (2002)
7. Robertson, S.E., Walker, S. & Beaulieu, M.: Experimentation as a way of life: Okapi at TREC. Information Processing & Management. 36(1), 95-108 (2000)
8. Amati, G., & van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems. 20(4), 357-389 (2002)
9. Akasereh, M., Savoy, J.: Ad Hoc Retrieval with Marathi Language. Working notes, Forum for Information Retrieval Evaluation (2011)
10. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. Proceedings TREC-4, 25-48. NIST Publication #500-236, Gaithersburg, (1996).
11. Abdou, S., Savoy, J.: Searching in Medline: Stemming, Query Expansion, and Manual Indexing Evaluation. Information Processing & Management. 44, 781-789, (2008).
12. Peat, H.J., Willett, P.: The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. JASIS. 42, 378-383, (1991).
13. Vogt, C.C., & Cottrell, G.W.: Fusion via a linear combination of scores. IR Journal. 1(3), 151-173 (1999)
14. Savoy J.: Data Fusion for Effective European Monolingual Information Retrieval. CLEF 2004. LNCS, vol. 3491, pp. 233-244. Springer, Heidelberg (2005)
15. Dolamic, L., Fautsch, C., Savoy, J. UniNE at CLEF 2008: TEL, and Persian IR. CLEF 2008. LNCS, vol. 5706, pp. 178-185. Springer, Heidelberg (2009)
16. Savoy, J., Berger, P.: Selection and Merging Strategies for Multilingual Information. CLEF 2004. LNCS, vol. 3491, pp. 27-37. Springer, Heidelberg (2005)
17. Dolamic, L., Savoy J.: How effective is Google's translation service in search? Commun. ACM. 52, 139-143 (2009)