UniNE at CLEF 2008: TEL, Persian and Robust IR

Ljiljana Dolamic, Claire Fautsch, Jacques Savoy

Computer Science Department

University of Neuchatel, Switzerland

{Ljiljana.Dolamic, Claire.Fautsch, Jacques.Savoy}@unine.ch

Abstract

In participating in this evaluation campaign, our first objective is to analyze the retrieval effectiveness when using TEL (The European Library) corpora composed of very short descriptions (library catalogue records) and to evaluate the retrieval effectiveness of several IR models. As a second objective we want to design and evaluate a stopword list and a light stemming strategy for the Persian language, a language belonging to the Indo-European family and having a relatively simple morphology. Finally, we participated in the robust track in an attempt to understand the difficulty involved in retrieving pertinent documents, even when the query and document representations share many common terms. Moreover, we made use of word sense disambiguation (WSD) information to order to reduce problems related to polysemy when matching topic and document representation.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods, Linguistic processing. I.2.7 [**Natural Language Processing**]: Language models. H.3.3 [**Information Storage and Retrieval**]: Retrieval models. H.3.4 [**Systems and Software**]: Performance evaluation.

General Terms

Experimentation, Performance, Measurement, Algorithms.

Additional Keywords and Phrases

Natural Language Processing, Stemmer, Digital Libraries, Persian Language (Farsi), Robust Retrieval.

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 and popular Asian languages (namely, Chinese, Japanese, and Korean). Our main objective in this context is to promote effective monolingual IR in those 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 different IR models used with TEL collections together with the evaluation of our official runs and certain related experiments. Section 4 presents the principal features of the Persian (Farsi) language, presents the stopword list and stemming strategy we developed for this language and describes our official runs and results for this task. Our participation and results concerning the robust task are outlined in Section 5, and Section 6 presents our main conclusions.

2 Overview of TEL Corpus

In a certain sense, this first ad hoc task takes us back to our research roots, because we need to look for relevant items among the catalog cards for a library collection. The European Library (TEL) available at `www.TheEuropeanLibrary.org` was used in our experiments that can be compared to previous work done with a French scientific bibliographic collection (Savoy, 2005). It includes three sub-collections, one in the English language (extracted from *British Library* (BL)), the second in German (coming from the *Austrian National Library* (ONB)), and the third in French (*Bibliothèque nationale de France* (BnF)). In this case the real

challenge was to retrieve pertinent records composed of a very short description of the referred information item. The only information contained in many records consists of only a title and author, and manually assigned subject headings.

Typical documents are shown in the tables below. Table 1a (*British Library*), Table 1b (*Austrian National Library*), and Table 1c (*Bibliothèque nationale de France*) shown the descriptions that appear in different languages. Table 1a shows a record with a title (tag <dc:title>) in German from a BL record and the subject in English (tag <dc:subject>). Table 1c illustrates another example with the title (tag <dc:title>) and a part of the description (tag <dc:description>) written in Latin.

```
<record>  <set>  TEL_BL_opac  </set>
  <header>  <id>  010624878  </id>  </header>
  <document format="index">  <index>  <topic>  BL_opac  </topic>  </index>  </document>
  <document format="dcx">  <oai_dc:dc>
    <dc:title>  Fehlprägungen und Fälschungen von Schweizer Münzen ab 1850 : mit Preisangaben.  </dc:title>
    <dc:contributor>  Richter, Jürg.  </dc:contributor>
    <dc:publisher>  Zürich : Helvetische Münzenzzeitung, [1988]  </dc:publisher>
    <dcterms:issued>  [1988]  </dcterms:issued>
    <dcterms:extent>  132p. : ill.  </dcterms:extent>
    <dc:language xsi:type="ISO639-2">  ger  </dc:language>
    <dc:subject>  Swiss coins to date Catalogues  </dc:subject>
    <dc:type>  text  </dc:type>
    <dc:identifier xlink:href="http://catalogue.bl.uk/F/-?func=direct-doc-
set&amp;amp;l_base=BLL01&amp;from=TELgateway&amp;doc_number=010624878">010624878</dc:identif
ier>
    <dc:identifier >  <dc:identifier xsi:type="dcterms:URI">http://catalogue.bl.uk/F/-?func=direct-doc-
set&amp;amp;l_base=BLL01&amp;from=TELgateway&amp;doc_number=010624878</dc:identifier>
    <mods:location>  British Library HMNTS YA.1992.b.771  </mods:location>
  </oai_dc:dc>  </document>  </record>
```

Table 1a: Example of a British Library (BL) record

```
<record>  <set>  TEL_ONB_onb01  </set>
  <id>  oai:aleph.onb.ac.at:ONB01-000000086  </id>
  <document format="index">  <index>  <topic>  ONB_onb01  </topic>  </index>  </document>
  <document format="dcx">  <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/dc/terms/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:identifier xsi:type="onb:ACCRecordId">  AC00454800  </dc:identifier>
  <dcterms:spatial xsi:type="dcterms:ISO3166">  DE  </dcterms:spatial>
  <dc:language xsi:type="dcterms:ISO639-2">  ger  </dc:language>
  <dc:creator>  Butor, Michel  </dc:creator>
  <dc:title>  &lt;&lt;Die&gt;&gt; Wörter in der Malerei  </dc:title>
  <dcterms:alternative>  Essay  </dcterms:alternative>
  <dcterms:edition>  1. Aufl  </dcterms:edition>
  <dc:publisher xsi:type="onb:PlaceofPublisher">  Frankfurt am Main  </dc:publisher>
  <dc:publisher xsi:type="onb:NameofPublisher">  Suhrkamp  </dc:publisher>
  <dcterms:issued>  1992  </dcterms:issued>
  <dcterms:isPartOf>  Bibliothek Suhrkamp ; 1093  </dcterms:isPartOf>
  <dc:identifier xsi:type="onb:ISBN">  3-518-22093-4  </dc:identifier>
  <dc:subject>  Malerei  </dc:subject>
  <dc:subject>  Legende &lt;Kunst&gt;  </dc:subject>
  <dc:identifier xsi:type="onb:CallNumber">  812861-B.1093  </dc:identifier>
  <dc:identifier xsi:type="onb:Location">  MAG  </dc:identifier>
  <dc:identifier xsi:type="onb:Collection">  ZNEU  </dc:identifier>
  <dc:type xsi:type="onb:ONBType">  BOOK  </dc:type>
  <dc:subject>  Malerei - Legende &lt;Kunst&gt;  </dc:subject>
  </oai_dc:dc>  </document>  </record>
```

Table 1b: Example of an Austrian National Library (ONB) record

<record> <set> **TEL_BnF**_opac </set>
<id>oai:bnf.fr:catalogue/ark:/12148/cb30000394c/description</id>
<document format="index"> <index> <topic>BnF_opac</topic> </index> </document>
<document format="dcx"> <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:dc="http://purl.org/dc/elements/1.1/">
<dc:identifier>http://catalogue.bnf.fr/ark:/12148/cb30000394c/description</dc:identifier>
<dc:title> Codex canonum vetus ecclesiae romanae a Francisco Pithoeo restitutus..</dc:title>
<dc:date> 1687 </dc:date>
<dc:description> Comprend : Apologeticus et epistolae </dc:description>
<dc:language> lat </dc:language>
<dc:type xml:lang="fre"> texte imprimé </dc:type>
<dc:type xml:lang="eng"> printed text </dc:type>
<dc:type xml:lang="eng"> text </dc:type>
<dc:rights xml:lang="fre"> Catalogue en ligne de la Bibliothèque nationale de France </dc:rights>
<dc:rights xml:lang="eng"> French National Library online Catalog </dc:rights>
</oai_dc:dc> </document> </record>
...
<record> <set> **TEL_BnF**_opac </set>
<id>oai:bnf.fr:catalogue/ark:/12148/cb319212546/description</id>
<document format="index"> <index> <topic>BnF_opac</topic> </index> </document>
<document format="dcx"> <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:dc="http://purl.org/dc/elements/1.1/">
<dc:identifier>http://catalogue.bnf.fr/ark:/12148/cb319212546/description</dc:identifier>
<dc:title> Ingénieux Hidalgo Don Quichotte de la Manche. Traduction nouvelle précédée d'une introduction par Jean Babelon </dc:title>
<dc:creator> Cervantes Saavedra, Miguel de (1547-1616) </dc:creator>
<dc:date> 1929 </dc:date>
<dc:description> Comprend : T. I. - Paris, A la Cité des Livres, 27, rue Saint-Sulpice. 1929. (16 mars.) In-8, XXIX-...55 p. [5224] ; T. 3. - 1929, 422 p. ; T. 4. - 1929, 423 p. </dc:description>
<dc:language> fre </dc:language>
<dc:type xml:lang="fre"> texte imprimé </dc:type>
<dc:type xml:lang="eng"> printed text </dc:type>
<dc:type xml:lang="eng"> text </dc:type>
<dc:rights xml:lang="fre"> Catalogue en ligne de la Bibliothèque nationale de France </dc:rights>
<dc:rights xml:lang="eng"> French National Library online Catalog </dc:rights>
</oai_dc:dc> </document>
</record>

Table 1c: Two examples of French records

TEL collections statistics are shown below in Table 2. The average size of each descriptor is relatively short (between 10 and 16), and similar across all three languages (perhaps a bit longer for the French corpus). During the indexing process we retained only the following logical sections from the original documents: <dc:title>, <dc:description>, <dc:subject>, and <dcterms:alternative>. From the topic descriptions we automatically removed certain phrases such as “Relevant document report …” or “Relevante Dokumente berichten …”, etc. All our runs were fully automatic.

As shown in Appendix 2, the available topics cover various subjects (e.g., Topic #452: “Celtic Art,” Topic #500: “Gauguin and Tahiti,” Topic #470: “Car Industry in Europe,” or Topic #498: “World War I Aviation”). We were surprised to see that the topic descriptions do not contain many proper names (creators and their works or geographical names). We found two topics with personal names (“Henry VIII” and “Gauguin”) but 23 with geographical names (e.g., “Europe,” “Eastern,” “Bordeaux” or “Greek”). The expression used to refer to a given location is not standardized, with various expressions being used to refer to a similar location (e.g., “USA,” “North America,” or “America”). Also, time periods are infrequently used (7 topics) and many include expressions having rather broad (e.g., “Modern,” “Ancient,” or “Roman”) or more precise (“World War I”) interpretations.

	English	French	German
Size (in MB)	1.2 GB	1.3 GB	1.3 GB
# of documents	1,000,100	1,000,100	869,353
# of distinct terms	9,087,132	15,189,862	10,629,539
Number of distinct indexing terms per document			
Mean	10	16	13
Standard deviation	6	11	9
Median	8	13	11
Maximum	168	618	222
Minimum	0	0	0
Number of indexing terms per document			
Mean	12	19	22
Standard deviation	8	17	17
Median	9	15	19
Maximum	330	1004	555
Minimum	0	0	0
Number of queries	50	50	50
Number rel. items	2,533	1,339	1,637
Mean rel./ request	50.66	26.78	32.74
Standard deviation	44.85	33.77	22.11
Median	32	16.5	28.5
Maximum	190 (T #472)	224 (T #465)	84 (T #477)
Minimum	7 (T #473)	3 (T #451)	2 (T #453)

Table 2: TEL test-collection statistics

3 IR models and Evaluation

3.1 Indexing Approaches

In defining our indexing strategies, we used a stopword list to denote very frequent forms having no important impact on sense-matching between topic and document representatives (e.g., "the," "in," "or," "has," etc.). In our experiments, the stopword list contains 589 English, 484 French and 578 German terms. The diacritics were replaced by their corresponding non-accented equivalent. We reused the light stemmers we developed for the French and German languages, because removing the inflectional suffixes attached only to nouns and adjectives tends to result in better retrieval effectiveness than more aggressive stemmers that also remove derivational suffixes (Savoy, 2006). These stemmers and stopword lists are freely available at the Web site `www.unine.ch/info/clef`. For the English languages we tried both a light stemming (S-stemmer proposed by Harman (1991) that removes only the plural form '-s') and a more aggressive one (Porter, 1980) based on a list of around 60 suffixes.

In the German language, compound words are widely used. For example, a life insurance company employee would be "Lebensversicherungsgesellschaftsangestellter" ("Leben" + 's' + "Versicherung" + 's' + "Gesellschaft" + 's' + "Angestellter" for life + insurance + company + employee). The augment (i.e. the letter 's' in our previous example) is not always present (e.g., "Bankangestelltenlohn" combines "Bank" + "Angestellten" + "Lohn" (salary)). Since compound construction is so widely used and written in many different forms, it is almost impossible to compile a dictionary providing quasi-total coverage of the German language. Thus an effective IR system including an automatic decompounding procedure for German had to be developed (Braschler & Ripplinger, 2004). In our experiments, we used our own automatic decompounding procedure (Savoy, 2004) leaving both the compounds and their composite parts in the topic and document representatives.

3.2 IR Models

In order to obtain high MAP values, we considered adopting different weighting schemes for the terms included in documents or queries. This would allow us to account for term occurrence frequencies (denoted tf_{ij} for indexing term t_j in document D_i), as well as their inverse document frequency (denoted idf_j). Moreover, we considered normalizing each indexing weight using the cosine to obtain the classical *tf·idf* formulation.

In addition to this classical vector-space approach, we also considered probabilistic models such as the Okapi (or BM25) (Robertson *et al.* 2000) that also take document length into account. As a second probabilistic

approach, we implemented three variants of the DFR (*Divergence from Randomness*) family of models suggested by Amati & van Rijsbergen (2002). In this framework, the indexing weight w_{ij} attached to term t_j in document D_i combines two information measures as follows

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}(tf)] \cdot (1 - Prob^2_{ij}(tf))$$

As a first model, we implemented the PB2 scheme, defined by the following equations:

$$Prob^1_{ij} = (e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}! \quad \text{with } \lambda_j = tc_j / n \tag{1}$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)] \tag{2}$$

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* the average document length, *n* the number of documents in the corpus, and *c* a constant (the corresponding values are given in the Appendix 1).

For the second model called GL2, the implementation of $Prob^1_{ij}$ is given by Equation 3, and $Prob^2_{ij}$ is given by Equation 4, as follows:

$$Prob^1_{ij} = [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \tag{3}$$

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \tag{4}$$

where λ_j and tfn_{ij} were defined previously.

For the third model called I(n_e)B2, the implementation was applied using the following two equations:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0{,}5)] \quad \text{with } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \tag{5}$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)] \tag{6}$$

where *n*, tc_j and tfn_{ij} were defined previously, and df_j indicates the number of documents in which the term t_j occurs.

Finally, we also considered an approach based on a statistical language model (LM) (Hiemstra, 2000; 2002), known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus not be based on any known distribution (e.g., as in Equation 1 or 3), but rather be directly estimated based on the term occurrence frequencies in document D_i or corpus C. Within this language model paradigm, various implementations and smoothing methods might be considered, although in this study we adopted a model proposed by Hiemstra (2002), as described in Equation 7, combining an estimate based on document ($P[t_j | D_i]$) and on corpus ($P[t_j | C]$) corresponding to the Jelinek-Mercer smoothing approach.

$$P[D_i | Q] = P[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | D_i] + (1-\lambda_j) \cdot P[t_j | C]]$$
$$\text{with } P[t_j | D_i] = tf_{ij}/l_i \quad \text{and } P[t_j | C] = df_j/lc \quad \text{with } lc = \sum_k df_k \tag{7}$$

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

3.3 Overall Evaluation

To measure retrieval performance, we adopted MAP values computed on the basis of 1,000 retrieved items per request as calculated with the TREC_EVAL program. Using this evaluation tool, some evaluation differences may occur in the values computed according to the official measure (the latter always takes 50 queries into account while in our presentation we do not account for queries having no relevant items). In the following tables, the best performance under the given conditions (with the same indexing scheme and the same collection) is listed in bold type.

Table 3 shows the MAP achieved by various probabilistic models using the English collection with two different query formulations (T or TD) and two stemmers. The last two columns show the MAP achieved by the French corpus and using our light stemmer. An analysis of this data shows that the best performing IR model would be usually the DFR I(n_e)B2, for all stemming approaches or query sizes. For the English corpus with Porter stemmer and TD query formulation, the LM model produces however a slightly better performance (0.3701 vs. 0.3643, a relative difference of 1.6%).

In the last lines we reported the MAP average over these 5 IR models together with percentage variations derived from comparing the short (T) query formulation to the performance achieved using Porter stemmer and T query (last line). As depicted in the last lines, increasing the query size improves the MAP (around +12.4% to +14.7%). According to the average performance, the best indexing approach seemed to be the stemming

approach using Porter's approach. In this case, the MAP with TD query formulation was 0.3559 on average, versus 0.3416 for the S-stemmer, a relative difference of 4.2%.

	Mean average precision					
Query Stemmer Model \ # of queries	English T S-stemmer 50 queries	English TD S-stemmer 50 queries	English T Porter 50 queries	English TD Porter 50 queries	French T 50 queries	French TD 50 queries
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	0.3072	**0.3541**	0.3258	0.3643	**0.2825**	**0.3291**
LM (λ=0.35)	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
Average over the 5 best IR	0.2981	0.3416	0.3166	0.3559	0.2740	0.3142
% change over T		+14.57%		+12.43%		+14.68%
% change over S-stemmer			+6.19%	+4.20%		

Table 3: MAP of various IR models and query formulations (English & French collection)

In Table 4 we reported the MAP achieved by probabilistic models using the German collection with two query formulations (T or TD) and comparing the performance with and without our automatic decompounding approach. The best IR model seemed to be the DFR PB2 (without decompounding) or the LM model when applying our decompounding scheme. By adding terms to the topic descriptions, we were also able to improve retrieval performance (between 17.4% to 31.0%). From comparing the average performances, it can be seen that applying an automatic decompounding approach improves retrieval effectiveness (see last line of Table 4, with an average improvement of 46.8% for short query formulations, or +31.5% when considering TD queries).

	Mean average precision			
Query Decompounding? Model \ # of queries	German T 50 queries	German TD 50 queries	German T + decompounding 50 queries	German TD + decompounding 50 queries
Okapi	0.1433	0.1872	0.2145	0.2522
DFR PB2	**0.1603**	**0.2097**	0.2150	0.2555
DFR GL2	0.1439	0.1878	0.2264	0.2615
DFR I(n_e)B2	0.1574	0.2071	0.2204	0.2615
LM (λ = 0.35)	0.1499	0.1972	**0.2315**	**0.2697**
tf idf	0.1084	0.1382	0.1286	0.1598
Average	0.1510	0.1978	0.2216	0.2601
% change over T		+31.03%		+17.39%
% change			+46.77%	+31.49%

Table 4: MAP of various IR models and query formulations (German collection)

An analysis showed that pseudo-relevance feedback (whether PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach (denoted "Roc" in the following tables) (Buckley *et al.*, 1996) 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 our previous experiments we learned that this type of blind query expansion strategy does not always work well. More particularly, we believe that including terms occurring frequently in the corpus (because they also appear in the top-ranked documents) may introduce more noise, and thus be an ineffective means of discriminating between relevant and non-relevant items (Peat & Willett, 1991). Consequently we also chose to apply our *idf*-based query expansion model (denoted "idf" in following tables) (Abdou & Savoy, 2008).

To evaluate these propositions, we applied certain probabilistic models and enlarged the query by adding the 20 to 150 terms retrieved from the 3 to 10 best-ranked articles contained in the English collection (Table 5), and both the French and German corpora (Table 6).

	Mean average precision			
Query TD PRF	English S-stemmer / idf	English S-stemmer / idf	English Porter / Roc	English Porter / Roc
IR Model / MAP	Okapi **0.3171**	DFR GL2 **0.3300**	Okapi **0.3329**	LM 0.3701
k doc. / *m* terms	5/10 0.2878 5/20 0.3076 5/50 0.3099 5/100 0.3100	10/10 0.2811 10/20 0.2983 10/50 0.3041 10/100 0.3053	3/10 0.3142 3/20 0.3178 3/50 0.3181 3/100 0.3181	5/10 0.3913 5/20 0.3991 5/50 0.4025 10/50 **0.4041**

Table 5: MAP using blind-query expansion (English collection)

	Mean average precision			
Query TD PRF	French idf	French Roc	German + decomp. / idf	German + decomp. / Roc
IR Model / MAP	Okapi 0.2998	DFR I(n_e)B2 0.3291	Okapi 0.2522	DFR I(n_e)B2 0.2615
k doc. / *m* terms	10/10 0.2838 10/20 0.2951 10/50 0.2953 5/50 **0.3062**	5/10 **0.3304** 10/10 0.3253 10/20 0.3239 10/50 0.3268	3/10 0.2444 5/10 0.2302 5/20 0.2414 5/50 **0.2543**	5/10 0.2654 5/20 0.2713 5/50 0.2757 10/50 **0.2851**

Table 6: MAP using blind-query expansion (French & German collection)

3.4 Data Fusion

It is usually assumed that combining different search models may improve retrieval effectiveness (Vogt & Cottrell, 1999), for three main reasons. First there is a skimming process in which only the *k* top-ranked retrieved items from each ranked list are considered. In this case, we would combine the best answers obtained from various document representations (which would retrieve various pertinent items). Second we would count on the chorus effect, by which different retrieval schemes would retrieve the same item, and as such provide stronger evidence that the corresponding document was indeed relevant. Third, an opposite or dark horse effect may also play a role, whereby a given retrieval model may provide unusually high (low) and accurate estimates regarding a document's relevance. Thus, a combined system could possibly return more pertinent items by accounting for documents having a relatively high (low) score, or when a relatively short (long) result lists occurs. Such a data fusion approach however requires more storage space and processing time. In the trade-off between the advantages and drawbacks, it is unclear whether such approaches might be of any real commercial interest.

In this current study we combined three probabilistic models representing both the parametric (Okapi and DFR) and non-parametric (language model or LM) approaches. To produce such a combination we evaluated various fusion operators (see Table 7 for a detailed list of their descriptions). The "Sum RSV" operator for example indicates that the combined document score (or the final retrieval status value) is simply the sum of the retrieval status value (RSV_k) of the corresponding document D_k computed by each single indexing scheme (Fox & Shaw, 1994). Table 7 thus illustrates how both the "Norm Max" and "Norm RSV" apply a normalization procedure when combining document scores. When combining the retrieval status value (RSV_k) for various indexing schemes and in order to favor certain more efficient retrieval schemes, we could multiply the document score by a constant α_i (usually equal to 1), reflecting the differences in retrieval performance.

Sum RSV	$\text{SUM}\ (\alpha_i \cdot RSV_k)$
Norm Max	$\text{SUM}\ (\alpha_i \cdot (RSV_k / Max^i))$
Norm RSV	$\text{SUM}\ [\alpha_i \cdot ((RSV_k - Min^i) / (Max^i - Min^i))]$
Z-Score	$\alpha_i \,.\, [((RSV_k - Mean^i) / Stdev^i) + \delta^i]$ with $\delta^i = [(Mean^i - Min^i) / Stdev^i]$

Table 7: Data fusion combination operators used in this study

In addition to using these data fusion operators, we also considered the round-robin approach, wherein we took one document in turn from each individual list and removed any duplicates, retaining only the highest ranking occurrence. Finally we suggested merging the retrieved documents according to the Z-Score, computed for each result list. More details can be found in Savoy & Berger (2005). In Table 7, Min^i (Max^i) lists the minimal (maximal) RSV value in the *i*th result list. Of course, we might also weight the relative contribution of each retrieval scheme by assigning a different α_i value to each retrieval model (fixed to 1 in all our experiments).

	Mean average precision (% of change)			
Language / Query Model	English TD 50 queries	French TD 50 queries	German TD 50 queries	German T 50 queries
Okapi & PRF doc/term	idf 10/20 0.3190	idf 10/20 0.2951	idf 5/10 0.2302	idf 5/10 0.2568
DFR GL2	idf 10/50 0.3041	idf 10/50 0.3070	Roc 5/20 0.2356	Roc 5/20 0.1967
DFR I(n_e)B2	Roc 10/10 **0.3745**	Roc 10/10 **0.3253**	Roc 5/50 **0.2757**	Roc 5/50 **0.2838**
Official run name	UniNEen1	UniNEfr1	UniNEde1	UniNEde4
Round-robin	0.3187 (-14.9%)	0.2950 (-9.3%)	0.2045 (-26.0%)	0.2316 (-18.4%)
Sum RSV	0.3510 (-6.3%)	0.3282 (+0.9%)	0.2917 (+5.8%)	**0.2840** (+0.1%)
Norm Max	0.3542 (-5.4%)	0.3284 (+1.0%)	0.2912 (+5.6%)	0.2730 (-3.8%)
Norm RSV	0.3534 (-5.6%)	0.3274 (+0.6%)	0.2945 (+6.8%)	0.2777 (-2.1%)
Z-Score	0.3543 (-5.4%)	**0.3284** (+1.0%)	**0.3013** (+9.3%)	0.2838 (0.0%)

Table 8: Mean average precision using different combination operators (with blind-query expansion)

Run name	Query	lang.	Index	Model	Query expansion	Single MAP	Comb MAP
UniNEen1	TD	EN	Porter	Okapi	idf 10 docs / 20 terms	0.3190	Z-score
	TD	EN	S-stem	GL2	idf 10 docs / 50 terms	0.3041	0.3543
	TD	EN	Porter	I(n_e)B2	Roc 10 docs / 10 terms	0.3745	
UniNEen2	TD	EN	Porter	PB2	Roc 5 docs / 50 terms	0.3850	Z-Score
	TD	EN	S-stem	Okapi	idf 5 docs / 50 terms	0.3099	0.3706
UniNEen3	TD	EN	Porter	Okapi		0.3329	Z-score
	TD	EN	S-stem	I(n_e)B2		0.3541	**0.3754**
	TD	EN	Porter	LM	Roc 5 docs / 10 terms	0.3913	
UniNEen4	T	EN	Porter	Okapi	idf 10 docs / 20 terms	0.3135	Z-score
	T	EN	S-stem	GL2	idf 10 docs / 50 terms	0.3541	0.3446
	T	EN	Porter	I(n_e)B2	Roc 10 docs / 10 terms	0.3913	
UniNEfr1	TD	FR	stem	Okapi	idf 10 docs / 20 terms	0.2951	Z-score
	TD	FR	stem	GL2	idf 10 docs / 50 terms	0.3070	0.3284
	TD	FR	stem	I(n_e)B2	Roc 10 docs / 10 terms	0.3253	
UniNEfr2	TD	FR	stem	PB2	Roc 5 docs / 50 terms	0.3052	Z-Score
	TD	FR	stem	Okapi	idf 5 docs / 50 terms	0.3262	0.3254
UniNEfr3	TD	FR	stem	Okapi		0.2998	Z-score
	TD	FR	stem	I(n_e)B2		0.3291	**0.3327**
	TD	FR	stem	LM	Roc 5 docs / 10 terms	0.315	
UniNEfr4	T	FR	stem	Okapi	idf 10 docs / 20 terms	0.2741	Z-score
	T	FR	stem	GL2	idf 10 docs / 50 terms	0.2856	0.2898
	T	FR	stem	I(n_e)B2	Roc 10 docs / 10 terms	0.2798	
UniNEde1	TD	DE	decomp.	Okapi	idf 5 docs / 10 terms	0.2302	Z-score
	TD	DE		GL2	Roc 5 docs / 20 terms	0.2356	**0.3013**
	TD	DE	decomp.	I(n_e)B2	Roc 5 docs / 50 terms	0.2757	
UniNEde2	TD	DE	decomp	Okapi	Roc 5 docs / 20 terms	0.2521	Z-score
	TD	DE	decomp	PB2	idf 5 docs / 50 terms	0.2779	0.2786
UniNEde3	TD	DE	decomp	I(n_e)B2	idf 5 docs / 50 terms	0.2726	Z-score
	TD	DE		Okapi		0.1872	0.2797
	TD	DE	decomp	LM	idf 5 docs / 10 terms	0.2378	
UniNEde4	T	DE	decomp.	Okapi	idf 5 docs / 10 terms	0.2568	Z-score
	T	DE		GL2	Roc 5 docs / 20 terms	0.1967	0.2838
	T	DE	decomp.	I(n_e)B2	Roc 5 docs / 50 terms	0.2586	

Table 9: Description and mean average precision (MAP) of our official TEL monolingual runs

Table 8 depicts the evaluation of various data fusion operators, comparing them to the best single approach using the Okapi and two DFR probabilistic models (GL2 or I(n_e)B2). From this data, we can see that combining three IR models might improve retrieval effectiveness, only slightly for the French or the German collection with short query formulations (T), moderately for the German with TD queries. When combining different retrieval models, the Z-Score scheme tended to perform the best, or at least it had one of the best performing MAP (e.g.,

for the German corpus with T queries). Finally, when compared to the best single search model, the performance achieved by the various data fusion approaches can not be improved with the English corpus.

3.5 Official Results

Table 9 shows the exact specifications of our 12 official monolingual runs for the TEL evaluation task, based mainly on the probabilistic models (Okapi, DFR and statistical language model (LM)). For all languages we submitted three runs with the TD query formulation and one with the T. All runs were fully automatic and in all cases the same data fusion approach (Z-score) was applied. For the German corpus however we sometimes applied our decompounding approach (denoted by "decomp." in the "Index" column), but we always applied our light stemmer.

4 IR with Persian language

The Persian (or Farsi) language is a member of the Indo-European family with relatively few morphological variations. This year we used a corpus extracted from the newspapers Hamshahri, made available thought the efforts of the University of Tehran (`http://ece.ut.ac.ir/dbrg/hamshahri/`). As usual in various evaluation campaigns, the corpus contains news articles (611 MB, for the years 1996 to 2002). This corpus contains exactly 166,774 documents on a variety of subjects (politic, literature, art, and economy, etc.) and includes about 448,100 different words. Hamshahri articles vary between 1 KB and 140 KB in size, comprising on average about 202 tokens (or 127 if we only count the number of word types). The corpus was coded in UTF-8 and written using the 28 Arabic letters plus an additional 4 letters for the Persian language.

Table 10 lists statistics on the test-collection. Of the three situations considered, there was no stemming approach used in the first, a light stemmer in the second, and 4-gram indexing approach for the third (McNamee & Mayfield, 2004).

	No stemmer	Light stemmer	4-gram
Size (in MB)	611 MB	611 MB	611 MB
# of documents	166,774	166,774	166,774
# of distinct terms	448,100	324,028	175,914
Number of distinct indexing terms (word type) per document			
Mean	127.23	119.26	258.26
Standard deviation	124.58	118.1	237
Median	83	80	178
Maximum	3,561	2,755	5,266
Minimum	0	0	0
Number of indexing terms (tokens) per document			
Mean	202.13	202.13	445.63
Standard deviation	228.14	228.14	494.26
Median	123	123	278
Maximum	12,548	12,548	25,139
Minimum	0	0	0
Number of queries	50	50	50
Number rel. items	5,161	5,161	5,161
Mean rel./ request	103.22	103.22	103.22
Standard deviation	67.88	67.88	67.88
Median	93	93	93
Maximum	255 (T #552)	255 (T #552)	255 (T #552)
Minimum	7 (T #574)	7 (T #574)	7 (T #574)

Table 10: Persian test-collection statistics

For the Persian language we first built a stopword list containing 884 terms. Unlike most other lists, this one contains words most frequently occurring in the collection (determinants, prepositions, conjunctions, pronouns or some auxiliary verb forms), plus a large number of suffixes already separated from word stems in the collection (see examples given below).

As a stemming strategy, we can use a morphological analysis (Miangah, 2006) or our simple, fast and light stemming approach that attempts to remove only nouns and adjective inflections. In the Persian language, the general pattern for inflectional suffixes is as follows: <possessive> <plural> <other-suffix> <stem>. In our light

stemming strategy, we usually removed possessive, plural and some of the suffixes marked as others. The following examples of our light stemmer illustrate the relatively simple Persian morphology. From the plural form درختان ("trees"), we can obtain درخت ("tree"). For the possessive form, دستم ("my hand"), our stemmer will return دست ("hand"), and for the form ایرانیان ("Iranians") we obtain ایران ("Iran"). In this corpus we saw that in some circumstances the suffixes might be written together or separated from the word as in کشتی‌ها and ها کشتی ("boats"), or منزل‌ها and ها منزل ("houses"). The 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 تر , and ترین .

The Persian language uses few case markers (the accusative case and certain specific genitive cases), unlike the Latin, German or Hungarian languages. The accusative for the definite noun is followed by را which can be joined to the noun or written separately (e.g., مرد را for the noun "man"). The genitive case is expressed by means of coupling two nouns by means of the particle known as ezafe (e.g.پسر مرد "man's son"). As is usually done in the English language, other relations are expressed by means of prepositions (e.g., in, with, etc.). Both the stopword list and our light stemmer are freely available at `http://www.unine.ch/info/clef/`.

	Mean average precision					
Query Stemmer Model \ # of queries	T none 50 queries	TD none 50 queries	T light 50 queries	TD light 50 queries	T 4-gram 50 queries	TD 4-gram 50 queries
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 (λ=0.35)	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
Average (4 IR models)	0.3992	0.4183	0.4006	0.4214	0.3789	0.3961
% change over T		+4.78%		+5.21%		+4.55%
% change over "none"	baseline	baseline	+0.35%	+0.76%	-5.09%	-5.29%

Table 11: MAP of various IR models and query formulations (Persian collection)

Table 12 shows the exact specifications of our 4 official monolingual runs for the Persian IR evaluation task, based mainly on three probabilistic models (Okapi, DFR and statistical language model (LM)). We submitted runs with all three topic formulations (short or T, medium or TD, and long or TDN). All runs were fully automated and the same data fusion approach (Z-score) was applied in all cases. The combination strategy we followed attempted to combine different indexing units (words, stemmed words or 4-grams), based on various probabilistic and efficient IR models (Okapi or DFR) and using three different blind-query expansion techniques (Rocchio, *idf*-based or none).

Run name	Query	Index	Stem	Model	Query expansion	Single MAP	Comb MAP
UniNEpe1	T	word	none	PL2	none	0.4078	Z-score
	T	4-gram	none	LM	idf 10 docs / 100 terms	0.3783	0.4675
	T	word	none	Okapi	Roc 10 docs / 20 terms	0.4376	
UniNEpe2	TD	4-gram	none	I(n_e)C2	none	0.4235	Z-Score
	TD	word	none	PL2	none	0.4274	**0.4898**
	TD	word	light	PL2	Roc 10 docs / 20 terms	0.4513	
	TD	word	none	PL2	idf 10 docs / 20 terms	0.4311	
UniNEpe3	TD	4-gram	none	Okapi	Roc 5 docs / 100 terms	0.4335	Z-Score
	TD	word	none	LM	idf 10 docs / 70 terms	0.4141	0.4814
	TD	word	none	PL2	none	0.4274	
UniNEpe4	TDN	4-gram	none	LM	idf 10 docs / 100 terms	0.3738	Z-score
	TDN	word	none	LM	Roc 10 docs / 20 terms	0.4415	0.4807
	TDN	word	none	PL2	none	0.4425	

Table 12: Description and mean average precision (MAP) for our official Persian monolingual runs

5 Robust Retrieval

In the robust task (Voorhees, 2006), we were interested in learning why retrieving relevant items for a given topic could be hard, even if the query contains certain common terms found in the relevant documents. In order to evaluate various search techniques, we used a corpus created during recent CLEF evaluation campaigns. This

collection consists of articles published in 1994 in the newspaper *Los Angeles Times*, as well as articles extracted from the *Glasgow Herald* and published in 1995. This collection contains a total of 169,477 documents (or about 579 MB of data). On average each article contains about 250 (median: 191) content-bearing terms (not counting commonly occurring words such as "the," "of" or "in"). Typically, documents in this collection are represented by a short title plus one to four paragraphs of text, and both American and British English spellings can be found in the corpus. To compile the test set, we used the topics created during the CLEF 2003 campaign (Topics #141 - #200) as well as queries from the 2005 (Topics #251 - #300) and 2006 (Topics #301 - #350) evaluation campaign. In this test set we found 153 queries able to return at least one relevant item from the collection.

This year we were interested in verifying whether word-sense disambiguation (WSD) might improve retrieval effectiveness. For this reason the organizers provides us with a new version of both the document and topic descriptions containing the correct lemma (entry in the dictionary) and SYNSET number(s) of the corresponding entry in the WordNet thesaurus (version 1.6). Table 13 lists an example for the title of Topic #47. Under the attribute LEMA the corresponding English dictionary entry is shown (therefore a stemming procedure is no more needed) and under the tag SYNSET, we can find both the score and the SYNSET number. The surface form is indicated under the label <WF> and the Part-of-Speech (POS) tag is also available for each word.

```
<num> C047 </num>
<EN-title> Russian Intervention in Chechnya </EN-title>
...
<top>
<num> C047 </num>
<EN-title>
<TERM ID="C047-1" LEMA="russian" POS="JJ">
  <WF> Russian </WF>
  <SYNSET SCORE="1" CODE="02726367-a"/>  </TERM>
<TERM ID="C047-2" LEMA="intervention" POS="NN">
  <WF> Intervention </WF>
  <SYNSET SCORE="1" CODE="00805766-n"/>
  <SYNSET SCORE="0" CODE="04995117-n"/>  </TERM>
<TERM ID="C047-3" LEMA="in" POS="IN">
  <WF> in </WF>  </TERM>
<TERM ID="C047-4" LEMA="Chechnya" POS="NNP">
  <WF> Chechnya </WF>  </TERM>
</EN-title>
...
```

Table 13: Examples of a query (title-only) with and without WordNet thesaurus number, part of speech tag (POS) and lemma

Various possibilities have been put forward to explain why certain successful IR systems may fail for some queries (Buckley, 2004; Savoy, 2007). The organizers thought that the polysemy (already known as a problem in finding pertinent matches between query and document surrogates) could be partially resolved in an appropriate manner by using the SYNSET information.

Based on past experiments (Dolamic & Savoy, 2008) with this corpus and using the TD queries and Porter's stemmer (Porter, 1980), we achieved a MAP of 0.2216 with $tf \cdot idf$ IR model to 0.4070 with Okapi model (Robertson *et al.*, 2000). With this last IR model, the set of hardest topics (defined as a query listing no relevant items in the top-20) were composed of seven topics, namely Topic #153 ("Olympic Games and Peace"), Topic #301 ("Nestlé Brands"), Topic #320 ("Energy Crises"), Topic #188 ("German Spelling Reform"), Topic #258 ("Brain-Drain Impact"), Topic #309 ("Hard Drugs"), and Topic #322 ("Atomic Energy").

Run name	Query	Index	Model	Query expansion	Single MAP	Comb MAP
UniNERobust1	TD		$I(n_e)C2$	idf 5 docs / 50 terms	0.4019	Z-score
	TD		Okapi	none	0.4086	0.4317
UniNERobust2	TD	WSD & POS	$I(n_e)C2$	idf 5 docs / 50 terms	0.3829	Z-Score
	TD	WSD & POS	Okapi	none	0.3896	0.4000
UniNERobust3	TD		LM	idf 5 docs / 200 terms	0.4345	Z-Score
	TD	POS	$I(n_e)C2$	win 5 docs / 200 terms	0.4000	0.4347
	TD	WSD	$I(n_e)C2$	win 5 docs / 200 terms	0.3966	
UniNERobust4	TD		$I(n_e)C2$	none	0.3990	Z-score
	TD		LM	win 5 docs / 200 terms	0.4331	**0.4515**
	TD		Okapi	win 5 docs / 200 terms	0.3783	
UniNERobust5	TD	WSD	$I(n_e)C2$	none	0.4033	Z-score
	TD	WSD	LM	win 5 docs / 200 terms	0.4386	0.4410
	TD	WSD	Okapi	win 5 docs / 200 terms	0.3888	
UniNERobust6	TD		Okapi	none	0.4086	Z-score
	TD	WSD	LM	win 5 docs / 200 terms	0.4294	**0.4499**
	TD		$I(n_e)C2$	idf 5 docs / 50 terms	0.4019	

Table 14: Description and mean average precision (MAP) for our official robust monolingual runs

In the current experiments, we generated six different runs using word-sense disambiguation information. As shown in Table 14 above, we followed our combination strategy, taking into account the various probabilistic models using different blind query expansion approaches. Our best results were achieved in the UniNERobust4 run with a MAP of 0.4515. Moreover, if we compare runs with or without word sense disambiguation (WSD) information (lemma, POS tags and SYNSET), we see no real and important differences (e.g., UniNERobust1 vs. UniNERobust2, and UniNERobust4 vs. UniNERobust3).

Table 15 below lists the set of hard topics for each of our official runs (hard topics here are defined as those providing no relevant items listed in the top-20). Included in the list covering all six runs (shown in italics in Table 15) were Topic #153 ("Olympic Games and Peace", 1 relevant item), followed by Topic #343 ("South African National Party", 1 relevant article), and Topic #313 ("Centenary Celebrations", 20 relevant documents).

Run name	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
UniNERobust1	*153*	169	316	*313*	*343*	266	318	151	314	
UniNERobust2	*153*	178	188	266	*313*	*343*	280	314	320	
UniNERobust3	*153*	*343*	318	320	286	*313*	314	280		
UniNERobust4	*153*	*343*	266	318	151	155	*313*	169		
UniNERobust5	*153*	*343*	318	*313*	169	266	188	286		
UniNERobust6	*153*	169	*343*	318	*313*	314	151			

Table 15: The hardest topics ranked according to the first relevant and retrieved items (and with rank > 20)

As shown in Table 14, in our official runs a hard topic was where the query resulted in low average precision. Using this definition, Table 16 lists the 10 topics having the lowest mean average precision. When all six runs are listed we obtain: Topic #153 ("Olympic Games and Peace"), followed by Topic #343 ("South African National Party"), Topic #313 ("Centenary Celebrations"), Topic #320 ("Energy Crises"), Topic #286 ("Football Injuries"). In an attempt to explain why a topic was difficult, we might mention that for Topics #343 and #153 only one relevant document was retrieved. Based on our best run (UniNERobust4), this item was ranked low on the retrieved list (44^{th} for Topics #343, and 382^{th} with Topics #153) even though they contained a large number of search terms.

Run name	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
UniNERobust1	*153*	169	*313*	*343*	*320*	*286*	266	280	314	336
UniNERobust2	*153*	178	188	336	*313*	266	*286*	*320*	280	*343*
UniNERobust3	*153*	*343*	336	*320*	*286*	280	*313*	169	266	314
UniNERobust4	*153*	336	*343*	155	*320*	*313*	*286*	169	266	280
UniNERobust5	*153*	*286*	*313*	169	*343*	*320*	322	188	266	314
UniNERobust6	*153*	169	*343*	*286*	*313*	*320*	280	322	314	151

Table 16: The ten hardest topics showing their mean average precision (MAP)

6 Conclusion

In this ninth CLEF campaign we evaluated various probabilistic IR models using two different test-collections, the first composed of short bibliographic notices extracted from the TEL corpora (written in English, German and French languages), and the second newspapers articles written in the Persian language. For the latter we also suggested a stopword list and a light stemmer 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 3, 4 and 11). The Okapi model used in our experiments usually results in retrieval performances inferior to those obtained with the DFR or LM approaches.

For the Persian language (Tables 11 and 12), our light stemmer tends to produce better MAP than does the 4-gram indexing scheme (relative difference of 5.5%). On the other hand, the performance difference with an approach ignoring a stemming stage is rather small.

Using the TEL corpora, the pseudo-relevance feedback (Rocchio's model) tends to hurt the retrieval effectiveness (see Tables 5 or 6). A data fusion strategy may enhance the retrieval performance for the French and German (Table 8) or Persian languages (Table 12), but not with the English corpus.

In the robust track, using the blind query expansion and data fusion approaches (combining three different probabilistic models), we are able to improve the MAP from 0.4086 (Okapi) to 0.4515. However, if we define hard topics as queries for which we cannot find any relevant items listed in the top-20, then these two runs produce the same number of hard topics (7 over 153). Finally the performance differences with and without word sense disambiguation (WSD) information are rather small.

Acknowledgments

The authors would like to also thank the CLEF-2008 task organizers for their efforts in developing various European language test-collections. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

Abdou, S., & Savoy, J. (2008). Searching in Medline: Stemming, query expansion, and manual indexing evaluation. *Information Processing & Management*, 44(2), p. 781-789.

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), p. 357-389.

Braschler, M., & Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval? *IR Journal*, 7, p. 291-316.

Buckley, C. (2004). Why current IR engines fail. *Proceedings ACM-SIGIR'2004*, The ACM Press, p. 584-585.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART. In *Proceedings of TREC-4*, Gaithersburg: NIST Publication #500-236, 25-48.

Dolamic, L., & Savoy, J. (2008). Monolingual and Bilingual Searches: Evaluation, Challenges and Failure Analysis. Submitted.

Fox, E.A., & Shaw, J.A. (1994). Combination of multiple searches. In *Proceedings TREC-2*, Gaithersburg: NIST Publication #500-215, p. 243-249.

Harman, D.K. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), p. 7-15.

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. In *Proceedings of the ACM-SIGIR*, The ACM Press, p. 35-41.

McNamee, P. & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), 73-97.

Miangah, T.M. (2006). Automatic lemmatization of Persian words. *Journal of Quantitative Linguistics*, 13(1), p. 1-15.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5), p. 378-383.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, p. 130-137.

Robertson, S.E., Walker, S. & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

Savoy, J. (2004). Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7, p. 121-148.

Savoy, J., & Berger, P.-Y. (2005): Selection and merging strategies for multilingual information retrieval. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (Eds.): Multilingual Information Access for text, Speech and Images. Lecture Notes in Computer Science: Vol. 3491. Springer, Heidelberg, p. 27-37.

Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing & Management*, 41(4), 873-890.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. *Proceedings ACM-SAC*, The ACM Press, p. 1031-1035.

Savoy, J. (2007). Why do successful search systems fail for some topics? *Proceedings ACM-SAC*, The ACM Press, p. 872-877.

Vogt, C.C., & Cottrell, G.W. (1999). Fusion via a linear combination of scores. *IR Journal*, 1(3), 151-173.

Voorhees, E.M. (2006). The TREC 2005 robust track. *ACM SIGIR Forum*, 40, 2006, p. 41-48.

Appendix 1: Parameter Settings

	Okapi			DFR	
Language	*b*	k_1	*avdl*	*c*	*mean dl*
English TEL	0.55	1.2	12	2.0	12
French TEL	0.55	1.2	19	2.0	19
German TEL	0.55	1.2	22	2.0	22
Persian word	0.75	1.2	216	1.5	216
Persian 4-gram	0.75	1.2	445	1.5	445
English Robust	0.55	1.2	1984	4.5	1984

Table A.1: Parameter settings for the various test-collections

Appendix 2: Topic Titles

C451	Roman Military in *Britain*	C476	Contrastive Analysis of Electoral Systems
C452	Celtic Art	C477	Web Advertising
C453	Bombing of *Japanese* Cities	C478	Multilingual Upbringing
C454	The Inquisition in *Italy*	C479	Food Allergies
C455	*Irish* Emigration to *North America*	C480	Pilgrimage to *Santiago de Compostela*
C456	Women's Vote in the *USA*	C481	Famous Jazz Musicians
C457	Big Game Hunting in *Africa*	C482	Vegetarianism
C458	The Wives of **Henry VIII**	C483	Solar Energy
C459	Gardening for Children	C484	Soap-making
C460	Scary Movies	C485	Counterfeiting Money
C461	Ancient *Greek* Coins	C486	Pictures of Vintage Cars
C462	*Israeli* Secret Service	C487	Jousting in the Middle Ages
C463	Churches in *France*	C488	*African Americans* and the *American* Civil War
C464	Piano Lessons	C489	Graphics Programming
C465	Trade Unions	C490	*Bordeaux* Wine Guides
C466	Gay Fiction	C491	Salary Inequality between Sexes
C467	Formula One Drivers	C492	Homeopathic Cures for Children
C468	Modern *Japanese* Culture	C493	Recipes for Chocolate Desserts
C469	*Scottish* Music	C494	Youth Employment in *Europe*
C470	Car Industry in *Europe*	C495	Women in the *French* Revolution
C471	Watchmaking	C496	Gods in *Greek* Mythology
C472	Man in Space	C497	20th Century *S. American* Authors
C473	*British* Women Authors	C498	World War I Aviation
C474	Journeys to *Antarctica*	C499	Wonders of the Ancient World
C475	*Eastern* philosophy	C500	**Gauguin** and *Tahiti*

Table A.2: Query titles for CLEF-2008 TEL ad-hoc test-collections

C551	Wimbledon tennis cup	C576	Iran Khodro company
C552	Tehran's stock market	C577	Anti-Cancer Drugs
C553	2002 world cup	C578	Traffic Congestion in Tehran
C554	Stress and Health	C579	Tehran International book festival
C555	Road casualty statistics	C580	Iranian presidential election
C556	Nuclear energy regulations	C581	Plane crashes
C557	Iran football coaches	C582	Water shortage in Tehran
C558	Danger of solid oil	C583	Earthquake damages
C559	Best Fajr film	C584	Oil price changes
C560	Iran economic sanction	C585	Air pollution control
C561	Gardening handbooks	C586	European football champion league final
C562	Reconstruction of Kandovan tunnel	C587	Development of Iranian software industry
C563	Mad cow disease	C588	Chemical attacks
C564	Sport blood pressure	C589	Iranian carpet export
C565	Drought losses	C590	Merchandise smuggling
C566	Prevention detection kidney diseases	C591	Global warming
C567	Population growth control	C592	Widely used narcotics in Iran
C568	Cell phone expansion	C593	Masouleh (Masooleh) Province
C569	Cases of economic corruption	C594	Aircraft ticket prices
C570	Iran dam construction	C595	World cup South Korea Japan
C571	Global oil economy	C596	Iraqi weapons of mass destruction
C572	Shajarian Concert	C597	Tehran murders
C573	Gross amount film cinema	C598	Serial Killings
C574	Champion team Iran first league	C599	2nd of Khordad election
C575	PersPolis Club establishment date	C600	Inflation in Iran

Table A.3: Query titles for CLEF-2008 Persian ad-hoc test-collections