UniNE at CLEF 2009: Persian Ad Hoc Retrieval and IP

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

Abstract

This paper describes the participation of the University of Neuchâtel to the CLEF 2008 evaluation campaign. In the Persian *ad hoc* task, we suggest using a light suffix-stripping algorithm for the Farsi language and the evaluations demonstrated that such an approach performs better than a simple light stemmer, an approach ignoring the stemming stage or a language independent approach (n-gram). The use of a blind query expansion (e.g., Rocchio's model) may improve the retrieval effectiveness. Combining different indexing and search strategies may further enhance the corresponding MAP. In the Intellectual Property (IP) task, we try different strategies to select and weight pertinent words to be extracted from a patent description in order to form an effective query. We also evaluated different search models and found that probabilistic models tend to perform better than vector-space schemes.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Experimentation, Performance, Measurement, Algorithms.

Keywords

Intellectual Property, Persian Language, Stemming

1 Introduction

Our participation to the CLEF 2009 evaluation campaign was motivated by our objective to design and evaluate indexing and search strategies for other languages than English (e.g., Persian *ad hoc* retrieval track) on the one hand, and on the other, by developing effective domain-specific IR (patent retrieval in the current case, also called "Intellectual Property" or CLEF-IP).

If the English language was studied since 1960, other natural languages may reveal different linguistic constructions having an impact on the retrieval effectiveness. For some languages (e.g., Chinese, Japanese), word segmentation is not an easy task, while for others (e.g., German), the frequent use of different compound constructions to express the same object or idea may hurt the retrieval quality. The presence of numerous inflectional suffixes (e.g., Hungarian, Finnish), even for names (e.g., Czech, Russian) as well as numerous derivational suffixes must be taken

into account for an effective retrieval. In this context, the Persian language is member of the Indo-European family but it is written using Arabic letters. The underlying morphology is more complex than the English one but we cannot qualify it as hard compared to some languages such as Turkish or Finnish. In Persian (or Farsi) language, various suffixes are used to indicate the plural, the accusative or genitive cases as well as other suffixes (or prefixes) are employed to derive new words.

In the Intellectual Property *ad hoc* task, we face clearly a domain-specific IR problem. Based on a large set of patent descriptions written in part in three different languages (English, German and French), we need to retrieve patents that are similar to a submitted one. This task could be viewed as detecting conflict between patent claims or as claim validation. In such case and contrary to other search environments, the query formulation contains a large number of terms (e.g., full patent description) and the determination and extraction of the most useful terms for an effective search are a hard task. Due to the fact that a patent is composed of different parts with different relative importances from an IR point of view, a critical problem is to define the most pertinent passages to form the query. We think also that a language shift may occur between the language used in the submitted patent and the language used in other patents. For example, a patent proposal may not employ directly the term "pump" but may simply describe it to avoid direct and evident conflict with existing patents. As an additional problem, the submitted patent may concern only a subpart of a given object (e.g., injection in a pump system) and pertinent items must be related not on the general object (e.g., pump) but on the specific targeted element.

Of course we can automatically enlarge the query by extracting related terms provided by a general or specialized thesauri, by using commercial search engines or by using given web sites (e.g., Wikipedia). Using the citation information could be a way to improve the quality of the search (and such a search strategy may also cross easily the language barriers). The presence of drawings could also be used to enhance the retrieval effectiveness in some cases and under the assumption that an effective image-based search function is available (which is not the case for our participation).

The rest of this paper is organized as follows. Section 2 describes the main characteristics of the test-collections while Section 3 presents briefly the various IR models used in our evaluation. The evaluation of the different indexing and search models with our test-collections are described and analyzed in Section 4. Section 5 reports our official results for both the Persian *ad hoc* and the CLEF-IP track. Our main findings are regrouped in Section 6.

2 Overview of Test-Collections

2.1 Persian

The Persian test-collection used in this year's evaluation consists of newspaper articles extracted from *Hamshahri* (covering years 1996 to 2002). This corpus is the same one made available during the CLEF 2008 evaluation campaign and contains 611 MB of data with exactly 166,477 documents. In mean, we can find 202 terms per document (after stopword removal). All documents contain only <TEXT> information without further division. The last column of the Table 1 gives basic statistics for this collection. The collection also contains 50 new topics (e.g., Topic #600 - Topic #650) having total of 4,464 relevant items, with mean of 89.28 relevant items per query (median 81.5 and standard deviation 55.63). The Topic #610 has the smallest number of relevant items (e.g., 8) while the largest number of relevant items (e.g., 266) was found for the Topic #649.

2.2 Intellectual Property (IP)

For the CLEF-IP track a data collection of more than 1 million patent documents was made available. The patents are derived from sources from the European Patent Office (EPO) and cover English, German and French patents, with at least 100,000 documents in each language. The patent can be divided into four main parts, namely "front page" (biblio and abstract), state

	IP			Ad hoc
	English	German	French	Persian
# of documents	1,943,641	1,754,471	1,754,505	166,774
# of distinct terms	401,056,191	218,276,027	92,276,265	324,028
Number of distinct indexing terms per document				
Mean	204.73	111.43	47.11	119.26
Standard deviation	285.99	217.71	100.32	118.1
Median	44	9	6	80
Maximum	15,165	6,666	2,610	2,755
Minimum	0	0	0	0
Number of terms per document				
Mean	1,092.48	510.79	179.33	202.13
Standard deviation	2,060.91	1,155.09	482.83	228.14
Median	78	12	6	123
Maximum	100,001	69,841	44,257	12,548
Minimum	0	0	0	0

Table 1: Various Test-Collection Statistics

of the art, claims (actual protection) together with drawings and embedded examples and finally citations. As addition element, we may find search reports. Each patent is stored as XML file representing the logical structure of a patent description (filled filed between 1985 and 2000). In total we have 1,958,955 patent documents representing 13.8 GB of compressed data. The patent documents follow the "Alexandria XML" DTD[1], a XML DTD for standardized patent data.

Table 1 shows collection statistics. For the IP collection, the number of documents indicates the number of patents containing at least one entry in the given language. We observe that while 99.22% of the documents contain some information in English, only 89.51% contain German information and 89.56% French information.

Among the various information available in each patent document, we decided to keep only following information: revised international patent classification number (tag <CLASSIFICATION-IPCR>), abstract (<ABSTRACT>), patent description (<DESCRIPTION>), claims (<CLAIMS>) and the invention title (<INVENTION-TITLE>) for our experiments. Not each patent contains necessarily all of these fields and each document might be written using more than one language. We also kept the language information for each field, in order to apply language specific indexing strategies such as stemming, stopword removal or decomposition.

For this collection three different sets of topics were available, a small set containing 500 queries (denoted S bundle), a medium set (1,000 or M bundle) and a large set (10,000 or XL bundle). Each topic consist of an entire patent document stored in the same format as described previously, but not included in the corpus.

Since the the whole document could not be used as query, one of the main challenges of this task was to formulate an appropriate query out of a patent document. To generate the query we applied following procedure. For each term t_j contained in the abstract, description, claim or invention title of the patent, we computed a weight $w(t_j)$ based on Equation 1. The m terms with the highest weights are then chosen as query terms.

$$w(t_j) = \frac{tf_j \cdot idf_j}{\sqrt{\sum_k (tf_k \cdot idf_k)^2}} \tag{1}$$

where tf_j is the frequency of the term t_j in the patent and idf_j the inverse document frequency of t_j in the document collection. For our experiments we fixed $m = 100$. We reference to this query formulation as "Q". For some runs we added the classification numbers contained in the patent. In such cases, the query formulation will be referenced as "QC".

[1] http://www.ir-facility.org/pdf/clef/patent-document.dtd

Relevance assessments for this collection provide two levels of relevancy. For each topic documents considered relevant (level 1) and documents highly relevant (2) are given. If we consider both relevancy levels, we have an average of 6.22 relevant items per query, with a maximum of 56 and a minimum of 3. If however we consider only highly relevant patents, we have an average of 3.46 relevant items per query, with a maximum of 18 and a minimum of 1. All evaluations in this paper are done considering documents of level 1 and 2 as relevant. We ignore evaluation done only on the highly relevant items.

3 IR Models

In order to analyze the retrieval effectiveness under different conditions, we adopted various retrieval models for weighting the terms included in queries and documents. To be able to compare the different models and analyze their relative merit, we first used a classical *tf idf* model. We would thus take into account the occurrence frequency of the term t_j in the document D_i (tf_{ij}) as well as the inverse document frequency of term t_j in the collection ($idf_j = \ln(\frac{n}{df_j})$ with n the number of documents in the corpus and df_j the number of documents in which the term t_j occurs). Furthermore we normalized each indexing weight using the cosine normalization. Additionally to this classical vector-space model, we used other models issued from both the vector-space and probabilistic model families.

We implemented the "doc=Lnu, query=ltc" (or Lnu-ltc) and "doc=Lnc, query=ltc" (Lnc-ltc) weighting schemes proposed by Buckley *et al.* [1] issued from the class of vector-space models. For these models the document score for document D_j for the query Q is calculated by applying following formula:

$$score(D_i, Q) = \sum_{t_j \in Q} w_{ij} \cdot w_{Qj} \tag{2}$$

where w_{ij} represents the weight assigned to the term t_j in the document D_i and w_{Qj} the weight assigned to t_j in the query Q. For the Lnu-ltc model the weight assigned to the document term ("doc=Lnu") is defined by Equation 3 while Equation 4 gives the weight assigned to the query term ("query=ltc").

$$w_{ij} = \frac{\frac{\ln(tf_{ij})+1}{\ln(mean\_tf)+1}}{(1 - slope) \cdot pivot + slope \cdot nt_i} \tag{3}$$

$$w_{qj} = \frac{\ln(tf_{qj} + 1) \cdot idf_j}{\sqrt{\sum_{t_k \in Q}((\ln(tf_{qk}) + 1) \cdot idf_k)^2}} \tag{4}$$

where nt_i is the number of distinct indexing terms in the document D_i, *pivot* and *slope* are constants, and $mean\_tf$ is the mean term frequency in the document D_i (values are given in Table 1).

For the "doc=Lnc, query=ltc" model, the weight for the query is calculated by Formula 4 while for the document ("doc=Lnc") it is calculated using following equation:

$$w_{ij} = \frac{\ln(tf_{ij} + 1)}{\sqrt{\sum_{t_k \in D_i}((\ln(tf_{ik}) + 1) \cdot idf_k)^2}} \tag{5}$$

To complete the vector-space models, we implemented several probabilistic approaches. As a first probabilistic approach, we implemented the Okapi model (BM25) as proposed by Robertson *et al.* [2] evaluating the document score by applying following formula:

$$score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot \log\left[\frac{n - df_j}{df_j}\right] \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}} \tag{6}$$

with $K = k_1 \cdot ((1 - b) + b \cdot \frac{l_i}{avdl})$ where qtf_j denotes the frequency of term t_j in the query Q, and l_i the length of the document D_i. Average document length is represented by $avdl$ ((values are given in Table 1) while b and k_1 are constants.

As second probabilistic approach, we implemented several models issued from the *Divergence of Randomness* (DFR) paradigm as proposed by Amati *et al.* [3]. In this framework, two information measures are combined to compute the weight w_{ij} attached to the term t_j in the document D_i. The weight is then calculated using following formula:

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

As a first model, we implemented the DFR-PL2 scheme, defined by the following equations:

$$Prob^1_{ij} = \frac{e^{-\lambda_j} \cdot \lambda_j^{tfn_{ij}}}{tfn_{ij}!} \tag{7}$$

$$Prob^2_{ij} = \frac{tfn_{ij}}{tfn_{ij} + 1} \tag{8}$$

with $\lambda_j = \frac{tc_j}{n}$ and $tfn_{ij} = tf_{ij} \cdot \log_2(1 + \frac{c \cdot mean\_dl}{l_i})$ where tc_j represents the number of occurrences of term t_j in the collection. The constants c and $mean\_dl$ (average document length) are fixed according to the underlying collection (see Table 1).

As second model issued from the DFR framework, we implemented the DFR-InL2 model where $Prob^2_{ij}$ is defined as in Equation 8 and Inf^1 is defined as follows

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2 \left[\frac{n+1}{df_j + 0.5} \right] \tag{9}$$

where df_j is the number of documents in which the term t_j appears and tfn_{ij} defined as before.

As third and last DFR model, we implemented the DFR-In_eC2 model defined by following equations:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2 \left[\frac{n+1}{n_e + 0.5} \right] \tag{10}$$

$$Prob^2_{ij} = 1 - \frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)} \tag{11}$$

with $n_e = n \cdot (1 - (\frac{n-1}{n})^{tc_j})$ and $tfn_{ij} = tf_{ij} \cdot \ln(1 + \frac{c \cdot mean\_dl}{l_i})$.

Finally we also used a non-parametric probabilistic model based on a statistical language model. Both Okapi and DFR are viewed as parametric probabilistic models. In this study we adopted a model proposed by Hiemstra [4] combining an estimate based on document ($P(t_j|D_i)$) and on corpus ($P(t_j|C)$) and defined by following equation

$$P(D_i|Q) = P(D_i) \cdot \prod_{t_j \in Q} [\lambda_j \cdot P(t_j|D_i) + (1 - \lambda_j) \cdot P(t_j|C)] \tag{12}$$

with $P(t_j|D_i) = \frac{tf_{ij}}{l_i}$ and $P(t_j|C) = \frac{df_j}{lc}$ with $lc = \sum_k df_k$ where λ_j is a smoothing factor, l_i the length of document D_i, and lc an estimate of the size of the corpus C. In our experiments λ_j is constant for all indexing terms t_j.

4 Evaluation

To measure retrieval performance of our different runs, we adopted MAP values computed using the TREC_EVAL program. Using this tool, the MAP values are computed on the basis of 1,000 retrieved documents per query. In the following tables, the best performance under the given conditions are listed in bold.

Query T	Mean Average Precision (MAP)				
Stemmer	none	plural	light	perstem	5-gram
Okapi	0.3687	0.3746	0.3894	0.3788	0.3712
DFR-PL2	**0.3765**	**0.3838**	**0.3983**	0.3879	0.3682
DFR-In_eC2	0.3762	0.3830	0.3952	**0.3886**	**0.3842**
LM	0.3403	0.3464	0.3559	0.3471	0.3404
tf idf	0.2521	0.2632	0.2521	0.2575	0.2441
Mean	0.3428	0.3502	0.3582	0.3520	0.3416
% over "none"		+2.17%	+4.50%	+2.69%	-0.33%

Table 2: MAP of Various Indexing Strategies and IR models (Persian Collection)

4.1 Persian

The Persian language, belonging to the Indo-Aryan language family is written using 28 Arabic letters, with additional 4 letters (پ چ ژ گ) being added to express sounds not present in classical Arabic. Suffixes predominate Persian morphology. Even thought this language does not have the definite article in the strict sense, it can be said that the relative suffix ی (کتابی که, the book which) and suffix ه (پسره, the son, informal writing) perform this function. The plurals in the Persian are formed by means of two suffixes, namely ان for animate (پدر, father, پدران, fathers) and ها for inanimate (گل, flower, گلها, flowers) nouns, while the plural of Arabic nouns in this language is formed according to Arabic grammar rules (e.g., ات or ین for "sound" plurals). The "light" stemmer we propose for this language removes the above mentioned suffixes with addition of certain number of possessive and comparative suffixes, while the second stemmer we proposed, named "plural", detects and removes only the plural suffixes from Persian nouns together with any suffix that might follow them.

Table 2 shows the MAP achieved by five IR models as well as different indexing strategies with the short query formulation. Second column in Table 2 (marked "none") depicts the performance obtained by the word based indexing strategy without stemming, followed by the MAP achieved by our two stemmers, namely "plural" and "light". In the column marked "perstem" the results obtained using publicly available stemmer and morphological analyzer for the Persian language[2] are given. This stemmer is based on numerous regular expressions in order to remove the corresponding suffixes. Finally the last column of the table depicts the performance of the language independent 5-gram indexing strategy. It can be seen from this table that the best performing models for all indexing strategies are the models derived from the DFR paradigm (marked bold in the table). The best performing indexing strategy proves to be the "light" stemming approach with the exception of the *tf idf* IR model for which the best performance was obtained by "plural" indexing approach. In all experiments presented in this paper the stoplist[3] for the Persian language containing 884 terms has been used.

Table 3 shows the MAP obtained using two different indexing strategies, namely "none" and "light" over five IR models with three query formulations (short or T, medium or TD and long or TDN). It can be seen from Table 3 that augmenting the query size ameliorates the MAP over T query formulation by 8% in average for TD queries and 15% for TDN queries.

Upon inspection of obtained results, we have found that the pseudo-relevance feedback (PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness for this language. Table 4 depicts MAP obtained by using Rocchio's approach (denoted "Roc") [1] whereby the system was allowed to add m terms extracted from the k best ranked documents from the original query results. The MAP enhancement spans from +2.35% (light, Okapi, 0.4169 vs. 0.4267) to +11.1% (light, DFR-PL2, 0.4247 vs. 0.4718). We have also applied another *idf*-based query expansion model [5] in our official runs (see Table 6).

[2]http://sourceforge.net/projects/perstem/
[3]http://www.unine.ch/info/clef/

	Mean Average Precision					
Query	T	TD	TDN	T	TD	TDN
Stemmer	none	none	none	light	light	light
Okapi	0.3687	0.3960	0.4233	0.3894	0.4169	0.4395
DFR-PL2	0.3765	0.4057	0.4326	0.3983	0.4247	0.4521
DFR-In_eC2	0.3762	0.4051	0.4284	0.4226	0.4226	0.4417
LM	0.3403	0.3727	0.4078	0.3559	0.3867	0.4268
$tf\ idf$	0.2521	0.2721	0.2990	0.2521	0.2687	0.2928
mean	0.3428	0.3703	0.3982	0.3582	0.3839	0.4106
% over T		+8%	16.17%		+7.2%	14.62%

Table 3: MAP of Various IR Models and Query Formulations (Persian Collection)

	Mean Average Precision			
Query	TD	TD	TD	TD
Index	light	light	5-gram	5-gram
IR Model/MAP	Okapi 0.4169	DFR-PL2 0.4247	Okapi 0.3968	DFR-PL2 0.3961
PRF Rocchio	5/20 0.4306	5/20 0.4621	5/50 0.4164	5/50 0.4164
k doc./m terms	5/70 **0.4480**	5/70 0.4620	5/150 0.4238	5/150 **0.4238**
	10/20 0.4267	10/20 **0.4718**	10/50 0.4173	10/50 0.4173
	10/70 0.4441	10/70 0.4700	10/150 **0.4273**	10/150 0.4169

Table 4: MAP using Blind-Query Expansion (Persian Collection)

4.2 Intellectual Property

For indexing the patent documents, we applied different strategies depending on the language in order to improve retrieval effectiveness. To eliminate very frequent terms having no impact on sense-matching between topic and document we used a language specific stopword list for each language. Furthermore for each language the diacritics were replaced by their corresponding non-accented equivalent. We also applied a language specific stemming strategy. For the English language we used the S-stemmer as proposed by Harmann [6] and a stopword list containing 571 terms, while for the German language, we applied our light stemmer[4], a stopword list containing 603 words and a decompounding algorithm [7]. Finally for the French language we also used our light stemmer and a stopword list containing 484 words.

Table 5 shows the MAP achieved by the seven different models used on the IP collection as well as the different indexing strategies and query formulations. The evaluations were done using the small topic set. The last line indicates the MAP average computed for all IR models. For previous art search in the patents, we tried three different techniques. First we weighted search terms for each field (abstract, description, title, ...) separately and then added the results to obtain the final score for the given document. For example if one query term appears once in the abstract and once in the title, the term would have tf one for each field and *idf* related to the corresponding field. The weight is then calculated for each field and added up. We reference to this strategy as "Separated Fields". Second we weighted the search terms considering the whole document, i.e., if a term t occurs once in two different fields it has tf of two and to compute *idf* we consider the whole patent document. To this strategy we reference as "Single Field". The third and last strategy consist in searching only the description of the patent ("Description"). Furthermore we applied two query formulations either taking into account classification numbers ("QC") or not ("Q").

We observe that for all searching strategies except if we take only into consideration the

[4]http://www.unine.ch/info/clef/

	Mean Average Precision (MAP)			
Query Index Model / # of queries	Q Separated Fields 500	QC Separated Fields 500	Q Single Field 500	Q Description 500
Okapi	0.0832	0.0832	0.0843	0.0856
DFR-InL2	0.0849	0.092	0.0645	0.0872
DFR-PL2	0.083	0.0909	0.0515	0.0673
LM	**0.0886**	**0.0952**	**0.0891**	0.0787
Lnc-ltc	0.0735	0.0839	0.0554	**0.0884**
Lnu-ltc	0.0675	0.0782	0.0695	0.0589
tf idf	0.0423	0.0566	0.0380	0.0337
Mean	0.0748	0.0829	0.0464	0.0714

Table 5: MAP of Various IR Models and Query Formulations (CLEF-IP)

Run name	Query	Index	Model	Query exp.	MAP	Comb.MAP
UniNEpe1	T	word	DFR-PL2	none	0.3765	**0.4380**
	T	5-gram	LM	idf 10 docs/50 terms	0.3726	
	T	plural	Okapi	Roc 10 docs/70 terms	0.4197	
UniNEpe2	TD	5-gram	DFR-In_eC2	none	0.4113	0.4593
	TD	word	DFR-PL2	none	0.4057	
	TD	plural	Okapi	Roc 5 docs/70 terms	0.4311	
	TD	word	DFR-PL2	idf 10 docs/50 terms	0.4466	
UniNEpe3	TD	word	Okapi	Roc 5 docs/50 terms	0.4228	**0.4663**
	TD	plural	Okapi	Roc 5 docs/70 terms	0.4311	
	TD	perstem	DFR-PB2	idf 10 docs/50 terms	0.4462	
UniNEpe4	TDN	word	LM	Roc 10 docs/50 terms	0.4709	**0.4937**
	TDN	plural	Okapi	Roc 5 docs/70 terms	0.4432	
	TDN	perstem	DFR-PL2	Roc 10 docs/20 terms	0.4769	

Table 6: Description and MAP of Official Persian Runs

description part of the patent, the language modeling approach (LM) shows the best performance. We can also see that keeping the various fields separated (index "Separated Fields") shows slightly better performance then if we index everything together. Searching only in the the description field of the patent, we obtain similar performances as when searching in the whole patent document.

We furthermore observe that except if searching only in the description, vector-space models are generally outperformed by probabilistic models.

5 Official Results

5.1 Persian

Table 6 gives description and results of the four official runs submitted to the CLEF 2009 Persian *ad hoc* track. Each run is a fusion of several single runs using different IR models (DFR, Okapi, statistical language model(LM)), indexing strategies (word with and without stemming, 5-gram), query expansion strategies (Rocchio, *idf*-based or none) and query formulation (T, TD and TDN). The fusion was performed for all four runs using a Z-score operator [8]. In all cases we can see that combining different models, indexing and search strategies using Z-score approach improves clearly the retrieval effectiveness. In these different combinations, we however did not use our "light" stemmer showing a relatively hight retrieval effectiveness as depicted in Table 2.

Run name	Query	Index	#Queries	Model	MAP	Comb.MAP
UniNE_strat1	Q	Single	500	Lnu-ltc	0.0695	0.0695
UniNE_strat2	Q	Single	500	LM	0.0891	0.0891
UniNE_strat3	QC	Separated	500	DFR-InL2	0.092	**0.1024**
	Q	Description		Okapi	0.0856	
UniNE_strat4	Q	Separated	500	LM	0.0886	0.0961
	QC	Separated		Okapi	0.0832	
	Q	Single		LM	0.0891	
UniNE_strat5	Q	Description	500	Okapi	0.0856	0.0856
UniNE_strat6	QC	Separated	500	DFR-PL2	0.0909	0.0955
	Q	Single		Lnu-ltc	0.0554	
UniNE_strat7	QC	Separated	500	Lnc-ltc	0.0839	0.0839
UniNE_strat8	QC	Separated	10,000	Okapi	0.0994	0.0994

Table 7: Description and MAP of Official IP Runs

5.2 Intellectual Property

Table 7 shows our eight official runs submitted to the CLEF 2009 Intellectual Property task (IP). Each run is either one single run or a fusion of several single runs, created using a Z-score fusion operator as described in [8]. For the first seven strategies we used only the small topic set (500 queries or S bundle) while for the last strategy, we used all 10,000 available topics (XL bundle). We observe that combining various single runs with the Z-score method at best improves retrieval effectiveness. The best performing strategy (UniNE_strat3) is a combination of two probabilistic models, namely DFR-InL2 and Okapi and two different indexing strategies. The results for all runs lie relatively close together and present rather low MAP values. We do not consider expanding automatically query formulation due to the fact that the original topic expression was already unusually long compared to other *ad hoc* search done in past CLEF evaluation campaigns.

6 Conclusion

From our past experiences in various evaluation campaigns, the results achieved this year in CLEF confirm the retrieval effectiveness of the *Divergence from Randomness* probabilistic model family. In particular the DFR-PL2 or the DFR-In_eC2 implementation tends to produce high MAP when facing different test-collection. In both tracks, we found that using our Z-score operator to combine different indexing and search strategies tends to improve the resulting retrieval effectiveness.

For the Persian *ad hoc* task, we notice three main differences between results achieved last year and those obtained this year. First, using very short (title-only or T) query formulation, we achieved the best results in 2008. This is the contrary this year with results based on TDN topic formulation depicting the best MAP (see Table 3). Second, unlike last year, the use of our stemmers was effective this year (see Table 2), and particularly the "light" stemming approach. Third, applying a pseudo-relevance feedback enhance the retrieval effectiveness of the proposed ranked list (see Table 4). For the moment, we do not have found a pertinent explanation to such difference between the two years. However, during both evaluation campaigns we found that a word-based indexing scheme using our "light" stemmer tends to perform better than a n-gram scheme.

In the Intellectual Property task (CLEF-IP), we were not able to propose an effective procedure to extract the most useful search terms or passages able to discriminate between the relevant and non-relevant patents. In our case, we fixed an arbitrary and fixed limit of 100 search terms to be extracted from the submitted patent description based on their $tfidf$ weights. We experiment different indexing strategies and search models. It seems that building separate index for each field (title, abstract, description, ...) and then combining the resulting ranked list may improve the MAP. This task is particularly challenging knowing that only one participating group achieved a

MAP clearly higher than 0.1 demonstrating also that additional evaluation campaigns are needed in this domain-specific task.

Acknowledgments The authors would like to also thank the CLEF-2009 organizers for their efforts in developing test-collections. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

[1] C. Buckley, A. Singhal, and M. Mitra, "New retrieval approaches using SMART," in *Proceedings of TREC-4*, pp. 25–48, 1995.

[2] S. E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at TREC," *Information Processing & Management*, vol. 36, no. 1, pp. 95–108, 2000.

[3] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357–389, 2002.

[4] D. Hiemstra, "Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term," in *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, pp. 35–41, 2002.

[5] S. Abdou and J. Savoy, "Searching in Medline: Query expansion and manual indexing evaluation," *Information Processing & Management*, vol. 44, pp. 781–789, 2008.

[6] D. Harman, "How effective is suffixing," *Journal of the American Society for Information Science*, vol. 42, pp. 7–15, 1991.

[7] J. Savoy, "Report on CLEF-2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval," in *Comparative Evaluation of Multilingual Information Access Systems* (C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, eds.), vol. 3237 of *Lecture Notes in Computer Science*, pp. 322–336, Springer, 2004.

[8] J. Savoy and P.-Y. Berger, "Selection and merging strategies for multilingual information retrieval," in *Multilingual Information Access for Text, Speech and Images* (C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, eds.), vol. 3491 of *Lecture Notes in Computer Science*, pp. 27–37, Springer, 2005.