Published in Lecture Notes in Computer Science 6241, 102-109, 2010
which should be used for any reference to this work 

# Ad Hoc Retrieval with the Persian Language

Ljiljana Dolamic and Jacques Savoy

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

**Abstract.** This paper describes our participation to the Persian *ad hoc* search during the CLEF 2009 evaluation campaign. In this task, we suggest using a light suffix-stripping algorithm for the Farsi (or Persian) language. The evaluations based on different probabilistic models demonstrated that our stemming approach performs better than a stemmer removing only the plural suffixes, or statistically better than an approach ignoring the stemming stage (around +4.5%) or a $n$-gram approach (around +4.7%). The use of a blind query expansion may significantly improve the retrieval effectiveness (between +7% to +11%). Combining different indexing and search strategies may further enhance the MAP (around +4.4%).

## 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 studied since 1960. In fact, other natural languages may reveal different linguistic constructions having an impact on the retrieval effectiveness. For some languages (e.g., Chinese, Japanese [1]), word segmentation is not an easy task, while for others (e.g., German), the use of different compound constructions to express the same concept or idea may hurt the retrieval quality [2]. The presence of numerous inflectional suffixes (e.g., Hungarian [3], Finnish), even for names (e.g., Czech [4], Russian [5]) 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 written using Arabic letters. The underlying morphology [6] is slightly more complex than the English one but we cannot qualify it as hard compared to some languages such as Turkish or Finnish.

The rest of this paper is organized as follows. The next section describes the main characteristics of the Persian morphology and presents an overview of the test-collection. Section 3 exposes briefly the various IR models used in our evaluation. The evaluation of the different indexing and search models are described and analyzed in Section 4 follows by the description of our official results. Our main findings are regrouped in the last section.

## 2 Farsi (Persian) Language and Test-Collection

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. The morphology of this language is based on various suffixes used to indicate the plural, the accusative or genitive cases as well as other suffixes (or prefixes) are employed to derive new words. 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). Moreover, even though 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 suggested "light" stemmer[1] removes the above mentioned suffixes with addition of certain number of possessive and comparative suffixes. It is clearly less aggressive than, for example, the Porter's stemmer [7] used in the English language. The second stemmer we proposed, denoted "plural", detects and removes only the plural suffixes from Persian nouns together with any suffix that might follow them. This stemmer is similar to the English S-stemmer [8]. As a stemming strategy we may also consider using a morphological analysis [9]. Recent research demonstrates however that using a morphological analysis, a light or a more aggressive stemming approaches tend to produce statistically similar performance, at least for the English language [10].

To evaluate these various stemming approaches we will use the Persian test-collection composed of newspaper articles extracted from *Hamshahri* (covering years 1996 to 2002). This corpus is the same one made available during the CLEF 2008 campaign containing 166,477 documents. In mean, we can find 202 terms per document (after stopword removal). The available documents do not have any logical structure and are composed of a few paragraphs. During the indexing process, we have found 324,028 distinct terms.

The collection contains 50 new topics (numbered from Topic #600 to Topic #650) having total of 4,464 relevant items, with mean of 89.28 relevant items per query (median 81.5, standard deviation 55.63). The Topic #610 ("Benefits of Copyright Laws") 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 ("Khatami Government Oil Crisis").

## 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

[1] Freely available at `http://www.unine.ch/info/clef/`

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 $D_i$ (or $tf_{ij}$) as well as its inverse document frequency ($idf_j = \ln(\frac{n}{df_j})$ with $n$ the number of documents in the corpus, and $df_j$ the number of documents in which $t_j$ occurs). Furthermore we normalized each indexing weight using the cosine normalization.

To define the similarity between a document surrogate and the query, we compute the inner product as given by Equation 1.

$$score(D_i, Q) = \sum_{t_j \in Q} w_{ij} \cdot w_{Qj} \tag{1}$$

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 $Q$.

As other IR model, we implemented several probabilistic approaches. As a first probabilistic approach, we implemented the Okapi model (BM25) [11] evaluating the document score by applying following formula:

$$score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot \log\left[\frac{n - df_j}{df_j}\right] \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}} \tag{2}$$

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 $D_i$. The average document length is given by $avdl$ while $b$ (=0.75) and $k_1$ (=1.2) are constants.

As second probabilistic approach, we implemented several models issued from the *Divergence from Randomness* (DFR) paradigm [12]. In this framework, two information measures are combined to compute the weight $w_{ij}$ attached to the term $t_j$ in the document $D_i$ as shown in Equation 3.

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2(Prob^1_{ij}(tf_{ij})) \cdot (1 - Prob^2_{ij}(tf_{ij})) \tag{3}$$

As a first model, we implemented the DFR-PL2 scheme, defined by Equation 4 and 5.

$$Prob^1_{ij} = \frac{e^{-\lambda_j} \cdot \lambda_j^{tfn_{ij}}}{tfn_{ij}!} \tag{4}$$

$$Prob^2_{ij} = \frac{tfn_{ij}}{tfn_{ij} + 1} \tag{5}$$

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.

As second DFR model, we implemented the DFR-I$n_e$C2 model defined by following equations, with $n_e = n \cdot (1 - (\frac{n-1}{n})^{tc_j})$.

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2\left[\frac{n + 1}{n_e + 0.5}\right] \tag{6}$$

$$Prob^2_{ij} = 1 - \frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)} \tag{7}$$

Finally we also used a non-parametric probabilistic model based on a statistical language model. In this study we adopted a model proposed by Hiemstra [13] combining an estimate based on document ($P(t_j|D_i)$) and on corpus ($P(t_j|C)$) and defined by following equation:

$$P(D_i|Q) = P(D_i) \cdot \prod_{t_j \in Q} [\lambda_j \cdot P(t_j|D_i) + (1 - \lambda_j) \cdot P(t_j|C)] \quad (8)$$

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 $\lambda_j$ is a smoothing factor, and $lc$ an estimate of the size of the corpus $C$. In our experiments $\lambda_j$ is constant (fixed at 0.35) for all indexing terms $t_j$.

## 4 Evaluation

To measure retrieval performance we used the mean average precision (MAP) obtained from 50 queries. The best performance obtained under a given condition is shown in bold type in the following tables. In order to statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [14] based on two-sided non-parametric test ($\alpha = 5\%$). In all experiments presented in this paper our stoplist[2] for the Persian language containing 884 terms has been used.

**Table 1.** MAP of Various Indexing Strategies and IR models (T query)

| Query T | Mean Average Precision (MAP) | | | | | |
|---|---|---|---|---|---|---|
| Stemmer | none | plural | light | perstem | 5-gram | trunc-4 |
| Okapi | 0.3687† | 0.3746† | 0.3894† | 0.3788† | 0.3712† | 0.3954† |
| DFR-PL2 | **0.3765** | **0.3838** | **0.3983** | 0.3879 | 0.3682† | **0.4054** |
| DFR-I$n_e$C2 | 0.3762 | 0.3830 | 0.3952 | **0.3886** | **0.3842** | 0.4016 |
| LM | 0.3403† | 0.3464† | 0.3559† | 0.3471† | 0.3404† | 0.3546† |
| *tf idf* | 0.2521† | 0.2632† | 0.2521† | 0.2575† | 0.2441† | 0.2555† |
| Mean | 0.3428 | 0.3502 | 0.3582 | 0.3520 | 0.3416 | 0.3625 |
| % over "none" | | +2.17% | +4.50% | +2.69% | -0.33% | +5.76% |

Table 1 shows the MAP achieved by five IR models as well as different indexing strategies with the short query formulation. The second column in Table 1 (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[3] are given. This stemmer is based on numerous regular expressions to remove the corresponding suffixes. Finally the last two columns

[2] Freely available at `http://www.unine.ch/info/clef/`

[3] `http://sourceforge.net/projects/perstem/`

depict the performance of two language independent indexing strategies, namely 5-gram and trunc-4 [15]. With the $n$-gram approach, each word is represented with overlapping sequences of $n$ characters (e.g., from "computer", we obtain "compu", "omput", "mpute", and "puter"). With trunc-4, we retain only the first $n$ letter and, for example, the word "computer" will produce "comp". The values of 5 and 4 are selected to obtain the best possible performance.

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). To verify whether this retrieval performance is significantly better than the other IR models, we have applied our statistical test. In Table 1, we have added the symbol "†" after MAP values showing a statistically significant differences with the best performance. Clearly the best IR model is always significantly better than the classical *tf idf* vector-space model or than the LM approach. If the Okapi model performs always at a lower performance level, the differences are usually not statistically significant.

When analyzing the various stemming approaches, the best performing indexing strategy seems to be the "light" stemming approach. An an exception we can mention the *tf idf* IR model for which the best performance was obtained by "plural" indexing approach (0.2632). From data shown in Table 1, even if the "light" stemmer is the best approach, the performance differences are usually significant only when compared to an approach ignoring the stemming stage. Finally, the performance differences between both language-independent approaches ($n$-gram and trunc-$n$) and our "light" stemming are never statistically significant.

We performed a query-by-query comparison to understand the effect of stemming concentrating on DFR-PL2, the best performing IR model. Analyzing Topic #630 ("Iranian Traditional Celebrations") we come across almost full range of reasons for better performance of the light stemming resulting in MAP 0.3808 compared to 0.1458 achieved by "none" or 0.2042 by trunc-4. While the topic title contains the adjectives ایرانی (Iranian) and سنتی (traditional), the relevant documents contain also ایران (Iran), سنت (tradition), ستهای (traditions) being conflated into the same respective indexing term by our "light" stemmer, but not when ignoring the stemming stage. Topic contains also the plural form of the noun جشنهای (the celebrations) while جشن (celebration) and جشنها (celebrations) are also found in the relevant documents. With the trunc-4 scheme, the resulting indexing term is composed of three letters (the stem "celebration") and one letter of the suffix. Thus it is not possible to conflate the two forms "celebration" (3 letters) and "celebrations" (5 letters) under the same entry.

Table 2 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 the longest form or TDN). It can be seen that augmenting the query size improves the MAP over T query formulation by +8% in average for TD queries and +15% for TDN queries. Moreover, the performance difference that are statistically significant over the T query formulation are shown with the symbol "‡".

**Table 2.** MAP of Various IR Models and Query Formulations

| | 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-I$n_e$C2 | 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 | | +8.0% | +16.2% | | +7.2% | +14.6% |

Upon inspection of obtained results, we have found that the pseudo-relevance feedback can be useful to enhance retrieval effectiveness. Table 3 depicts MAP obtained by using Rocchio's approach (denoted "Roc") [16] whereby the system was allowed to add $m$ terms ($m$ varies from 20 to 150) extracted from the $k$ best ranked documents (for $k = 5$ to 10) from the original query results. The MAP enhancement spans from +2.4% (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 [17] in our official runs (see Table 4).

**Table 3.** MAP using Rocchio's Blind-Query Expansion

| | 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 |

## 5 Official Results

Table 4 gives description and results of the four official runs submitted to the CLEF 2009 Persian *ad hoc* track. Each run is based on a data fusion scheme combining several single runs using different IR models (DFR, Okapi, and language model (LM)), indexing strategies (word with and without stemming or 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 our Z-score operator [18]. In all cases we can see that combining different models, indexing and search strategies using Z-score approach improves clearly the retrieval effectiveness. For example, using the short query formulation (T), the best single

IR model achieved a MAP value of 0.4197, while after applying our data fusion operator, we obtained a MAP value of 0.4380, a relative improvement of +4.3%. In these different combinations, we however did not use our "light" stemmer showing a relatively hight retrieval effectiveness as depicted in Table 1.

**Table 4.** Description and MAP of Official Persian Runs

| Run name | Query | Index | Model | Query exp. | MAP | Comb.MAP |
|---|---|---|---|---|---|---|
| | T | word | PL2 | none | 0.3765 | |
| UniNEpe1 | T | 5-gram | LM | idf 10 docs/50 terms | 0.3726 | **0.4380** |
| | T | plural | Okapi | Roc 10 docs/70 terms | 0.4197 | |
| | TD | 5-gram | I$n_e$C2 | none | 0.4113 | |
| UniNEpe2 | TD | word | PL2 | none | 0.4057 | 0.4593 |
| | TD | plural | Okapi | Roc 5 docs/70 terms | 0.4311 | |
| | TD | word | PL2 | idf 10 docs/50 terms | 0.4466 | |
| | TD | word | Okapi | Roc 5 docs/50 terms | 0.4228 | |
| UniNEpe3 | TD | plural | Okapi | Roc 5 docs/70 terms | 0.4311 | **0.4663** |
| | TD | perstem | PB2 | idf 10 docs/50 terms | 0.4462 | |
| | TDN | word | LM | Roc 10 docs/50 terms | 0.4709 | |
| UniNEpe4 | TDN | plural | Okapi | Roc 5 docs/70 terms | 0.4432 | **0.4937** |
| | TDN | perstem | PL2 | Roc 10 docs/20 terms | 0.4769 | |

## 6 Conclusion

From our past experiences in various evaluation campaigns, the results achieved in this track confirm the retrieval effectiveness of the *Divergence from Randomness* probabilistic model family. In particular the DFR-PL2 or the DFR-I$n_e$C2 implementation tends to produce high MAP when facing different test-collections written in different languages. We can also confirm that using our Z-score operator to combine different indexing and search strategies tends to improve the resulting retrieval effectiveness.

In this Persian *ad hoc* task, we notice three main differences between results achieved last year and those obtained this year. First, using 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 2). Second, unlike last year, the use of our stemmers was effective this year, and particularly the "light" stemming approach (see Table 1). As language-independent approach, we can mention that the trunc-$n$ indexing scheme is also effective for the Persian language. Third, applying a pseudo-relevance feedback enhance the retrieval effectiveness of the proposed ranked list (see Table 3). 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 the $n$-gram scheme.

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

## References

1. Savoy, J.: Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. ACM Transactions on Asian Languages Information Processing 4, 163–189 (2005)
2. Braschler, M., Ripplinger, B.: How Effective is Stemming and Decompounding for German Text Retrieval? IR Journal 7, 291–316 (2004)
3. Savoy, J.: Searching Strategies for the Hungarian Language. Information Processing & Management 44, 310–324 (2008)
4. Dolamic, L., Savoy, J.: Indexing and Stemming Approaches for the Czech Language. Information Processing & Management 45, 714–720 (2009)
5. Dolamic, L., Savoy, J.: Indexing and Searching Strategies for the Russian Language. Journal of the American Society for Information Sciences and Technology 60, 2540–2547 (2009)
6. Elwell-Sutton, L.P.: Elementary Persian Grammar. Cambridge University Press, Cambridge (1999)
7. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 130–137 (1980)
8. Harman, D.K.: How Effective is Suffixing? Journal of the American Society for Information Science 42, 7–15 (1991)
9. Miangah, T.M.: Automatic Lemmatization of Persian Words. Journal of Quantitative Linguistics 13, 1–15 (2006)
10. Fautsch, C., Savoy, J.: Algorithmic Stemmers or Morphological Analysis: An Evaluation. Journal of the American Society for Information Sciences and Technology 60, 1616–1624 (2009)
11. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management 36, 95–108 (2002)
12. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20, 357–389 (2002)
13. Hiemstra, D.: Using Language Models for IR. Ph.D. Thesis (2000)
14. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management 33, 495–512 (1997)
15. McNamee, P., Nicholas, C., Mayfield, J.: Addressing Morphological Variation in Alphabetic Languages. In: Proceedings ACM - SIGIR, 75–82 (2009)
16. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings TREC-4, Gaithersburg, pp. 25–48 (1996)
17. Abdou, S., Savoy, J.: Searching in Medline: Stemming, Query Expansion, and Manual Indexing Evaluation. Information Processing & Management 44, 781–789 (2008)
18. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal 7, 121–148 (2004)