Persian Language, is Stemming Efficient?

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

***Abstract*—The main goal of this paper is to describe and evaluate different indexing and stemming strategies for the Farsi (Persian) language. For this Indo-European language we have suggested a stopword list and a light stemmer. We have compared this stemmer to indexing strategy in which the stemming procedure was omitted, with or without stopword list removal, another publicaly available stemmer for this language as well as language independent n-gram indexing strategy. To evaluate the suggested solutions we used various IR models, including Okapi, *Divergence from Randomness* (DFR), a statistical language model (LM) as well as two vector space models, the classical *tf idf* and *Lnu-ltc* model. We have found that the *Divergence from Randomness* paradigm tends to propose better retrieval effectiveness than the Okapi, LM or vector-space models, the performance differences were however statistically significant only with the last two IR approaches. Ignoring the stemming ameliorates the MAP by more than 7%, giving the differences that are most of the time statistically significant. Finally, not removing the stoplist words for this language deprecates the MAP performance by 3%.**

***Keywords*-Farsi language; stemmer; natural language processing;**

I. Introduction

Persian language, also known as Farsi belongs to Indo-European language family. As such, Persian is distantly related to the majority of European languages, including English and German, but unlike them is written in modified version of the Arabic script. With more than 50 million native speakers, this language is official language in Iran, Afghanistan and Tajikistan.

Persian has affixitive morphology, meaning that the suffixes and prefixes are concatenated to the words to change their meaning. The CLEF 2008 campaign has created test-collection for the Persian language, and based on this collection, the main objective of this paper is to describe the main morphological difficulties when working with this language. We also proposed and evaluated a suitable stemmer for Persian IR and compared it to another publicly available stemmer. In IR it is assumed that applying a stemmer in order to conflate several word variants into the same stem will improve the pertinent matching between query and document surrogates. For example, when a query contains the word "horse," it seems reasonable to also retrieve documents containing the related word "horses." Moreover, stemming procedures will also reduce the size of inverted files. When designing a stemmer, we may create a "light" suffix-stripping procedure by removing only the morphological inflections to conflate the singular and plural word forms (e.g., "door" and "doors") to the same stem. More sophisticated approaches will remove derivational suffixes (e.g., "enhance" and "enhancement") usually used to generate a new part-of-speech word from a given stem. Even though different stemming procedures have been suggested for various European languages (e.g., Snowball project, CLEF, TREC and NTCIR campaigns [1], [2]), no stemming algorithm with its evaluation is available for the Persian language.

In this paper, we present related work in next section while Section III describes main aspects of the Persian language. Section IV depicts the main characteristics of the test collection, while Section V describes the IR models used in our experiments. Section VI evaluates different indexing and search strategies. The main findings of this paper are summarized in Section VII.

II. Related Work

Most stemming approaches are based on the target language's morphological rules (e.g., the Porter stemmer for the English language [3]) where suffix removal is also controlled by quantitative restrictions (e.g., ' ing' is removed when the resulting stem has more than three letters as in "jumping," but not in "king") or qualitative restrictions (e.g., '-ize' is removed if the resulting stem does not end with '-e' as in "seize"). Certain ad hoc spelling correction rules can also be applied to improve conflation accuracy (e.g., "running" gives "run" and not "runn"), particularly when phonetic rules are applied to facilitate easier pronunciation.

Another approach consults an online dictionary to obtain better conflation results [4], while Xu & Croft [5] suggest a corpus-based approach that more closely reflects the language use rather than all its grammatical rules. Few stemming procedures[1] have been suggested for languages other than English. The proposed stemmers usually pertain to the most popular languages [1], [6] and some of them, like the Finnish language, seem to require a deeper morphological analysis [7] to achieve good retrieval performance.

[1]Freely available at the Web site http://snowball.tartarus.org/ or http://www.unine.ch/info/clef/

Algorithmic stemmer ignores word meanings and tends to make errors, usually due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "create" and "creation" do not conflate to the same root). Most of the studies so far have been involved in evaluating IR performance for the English language, while studies on the stemmer performance for less popular languages are less frequent. For example, Tomlinson [6] evaluated the differences between Porter's stemmer strategy [3] and lexical stemmers (based on a dictionary of the corresponding language) for various European languages. For the Finnish and the German language, lexical stemmer tends to produce statistically better results, while for seven other languages performance differences were insignificant.

III. Persian Morphologie and Stemming Strategies

While creating stemming procedures for the Persian language we adopted the same strategy as for the European languages for which we have created stemmers during the past years. We believe that effective stemming should focus mainly on nouns and adjectives (sustaining most of the meaning of a document), thus ignoring numerous verb forms (tending to generate more stemming errors when taken into account).

The Persian language belongs to Indo-European languages and is written using modified Arabic script containing 28 Arabic letters to which four new characters were added (پ چ ژ گ) - to express sounds not present in Classical Arabic. These 32 letters are written from right to left and have for the most part different forms according as they are initial, medial or final, and connected or unconnected with the letter that precedes or follows them. This language does not contain the definite article in the strict sense (کتاب - means "book" or "the book" according to the context), however the particle را which follows a definitive noun in accusative case (آب را بیار - bring the water) can be said to perform the function of the article, so do the relative suffix ی - (کتابی که - the book which) and ending ه - (پسره - the son, informal writing). As for indefinite article it is expressed by suffix ی - (which can also be added to plural nouns کتابهایی), placing a ء (hamze) over final ه - or single by means of numeral یک "one".

There is no gender in Persian. Natural gender can be expressed by use of different words (مرد - man, زن - woman), by means of the adjectives نر - "male" and ماده - "female" and in the case of Arabic words the use of Arabic feminine ending ه .

Even though there is no inflection strictly speaking the cases are expressed in various ways, accusative by the particle را , genitive by means of coupling nouns using particle " " known as ezafe ("پسرِ مرد" -the man's son)while other relations are expressed by means of prepositions.

Plural of nouns in Persian language is formed by adding suffixes ان for animate beings (پدران - fathers) and ها for inanimate objects (گلها - flowers). In some cases the plural ending ها can be written separately from the noun (e.g.,خانه‌ها or خانها - houses, first alternative being preferred) [8]. Arabic nouns take plural according to Arabic grammar by adding ات or ین for "sound" plurals or by alternating the vowel-pattern of the singular for "broken" plurals (e.g. قلب - heart vs. قلوب - hearts).

Suffixes predominate Persian morphology even though there is a small number of prefixes. Derivational Persian morphology is accomplished by means of prefixation and suffixation of a stem, a usual construction with the Indo-European languages. Usually, the part-of-speech of the stem changes after adding a suffix (e.g., '-ness' in "good" and "goodness") while the prefix changes the original meaning of the stem (e.g., "prehistory" vs. "historic" from the stem "history"). In Persian language this phenomenon occurs also with suffix adding. We can take as an example suffix ی - added to nouns to from relative adjectives, resulting in ایران - ایرانی (Iran - Iranian), but also آب - آبی (water - blue).

Suffixes described in this section are removed by our light stemmer in an effort to conflate different forms of nouns and adjectives to the same stem. The light stemmer we propose does not deal with verbal suffixes, which are, like in other languages numerous in the Persian as well. On the other hand, the publicly available stemmer and morphological analyzer for the Persian language "perstem" , which is also evaluated in this paper, besides removing suffixes from nouns and adjectives, removes also certain number of verbal suffixes.

Finally, to define pertinent matches between search keywords and documents, we removed very frequently occurring terms having no important significance (e.g. in, but, some). Unlike stoplists for other languages the stopword lists for this language contains also large number of suffixes already separated from the word stem. Suffixes in Persian can be written together with the word they alter or separated from it, and in the later case this should be performed by using so called short space. However, this short space is sometimes replaced by an ordinary space, this being the case in the collection used for this evaluation, causing these suffixes to be treated as words (e.g. plural suffix ها appears as a word). Both our light stemmer and the suggested stopword list for the Persian language are freely available at *www.unine.ch/info/clef/*.

IV. Test-collection

The test collection for Persian language used in this papers evaluation consists of newspaper articles extracted from *Hamshahri*(years 1996 to 2002) made available during the CLEF 2008 evaluation campaign. This corpus contains 166,477 documents with an average of 202 terms per document. This collection contains 100 topics covering various

subjects (e.g., "Gardening handbooks," "Human cloning," "Mad cow disease") including both regional ("Water shortage in Tehran") and more international coverage ("Global warming"). Topics #574 ("Champion team Iran first league") owns the smallest number of pertinent articles (7) while Topic #552 ("Tehran's stock market") has the greatest number of correct answers (255). Based on the TREC model, each topic was structured into three logical sections, comprising a brief title (T), a one-sentence description (D), and a narrative part (N) specifying the relevance assessment criteria. In our experiments, we used only the title part of the topic formulation in order to reflect more closely queries sent to commercial search engines. Using only the title section, queries had a mean size of 2.79 search terms.

V. IR Models

To evaluate our proposed two stemming approaches with respect to various IR models, first we used the classical tf idf model wherein the weight attached to each indexing term was the product of its term occurrence frequency tf_{ij} (for indexing term t_j in document d_i) and the logarithm of its inverse document frequency (idf_j). To measure similarities between documents and the request, we computed the inner product after normalizing (cosine) the indexing weights [9].

For the vector-space model better weighting schemes have been suggested, especially in cases where the occurrence of a term in a document is viewed as a rare event. A term presence in a shorter document might also provide stronger evidence than it would in a longer document. In order to take document length into account, we could make use of more complex IR models, the "*Lnu-ltc*" IR model suggested by [10]. In this case Equation 1 calculates the indexing weight assigned to document term (*Lnu*) and Equation 2 the indexing weight assigned to query term (*ltc*).

$$w_{ij} = [\log(tf_{ij}) + 1] \cdot norm_i \tag{1}$$

with

$norm_i = \frac{1}{(1+\log(\frac{\sum tf_{ij}}{nt_i}))\cdot((1-slope)\cdot pivot+(slope\cdot nt_i))}$

$$w_{qj} = (\log(tf_{qj}) + 1) \cdot idf_j \cdot norm_q \tag{2}$$

with

$norm_q = \frac{1}{\sqrt{\sum_k (tf_{qk}\cdot idf_j)^2}}$

To complement these vector-space models, we have implemented probabilistic models, such as the Okapi (or BM25) approach [11], and one model derived from *Divergence from Randomness* (DFR) paradigm [12] wherein two information measures formulated below are combined:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2\left[Prob^1_{ij}(tf)\right] \cdot (1 - Prob^2_{ij}) \tag{3}$$

where $Prob^1_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document. On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given tf_{ij} occurrences of this term had already been found. To model these two probabilities, we used the PL2 model based on the following estimates:

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

$$Prob^2_{ij} = 1 - \frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)} \tag{5}$$

with

$\lambda_j = \frac{tc_j}{n}$ and $tfn_{ij} = tf_{ij} \cdot \log(1 + \frac{c \cdot meandl}{l_i})$

where tc_j is the number of occurrences of term t_j in the collection, df_j indicates the number of documents in with the term t_j occurs, n the number of documents in the corpus, l_i the length of document d_i, mean dl ($= 202$), the average document length, and c a constant (fixed empirically at 1.5).

Finally, we also used an approach based on a language model (LM) [13], known as a non-parametric probabilistic model. Various implementations and smoothing methods might also be considered within this language model paradigm. In this paper we adopted a model proposed by Hiemstra [14] as described in Equation 4 using the Jelinek-Mercer smoothing, a combination of an estimate based on document ($P[t_j|d_i]$) and one based on the whole corpus ($P[t_j|C]$).

$$P(D_i|Q) = P(D_i) \prod_{t_j \in Q} (\lambda_j \cdot P(t_j|D_i) + (1 - \lambda_j) \cdot P(t_j|C)) \tag{6}$$

with

$P(t_j|D_i) = tf_{ij}/l_i$ and $P(t_j|C) = df_j/lc$ with $lc = \sum_k df_k$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and lc is a constant related to the size of the underlying corpus C.

VI. Evaluation

In order to measure retrieval performance, we have adopted the mean average precision (MAP) computed by TREC_EVAL based on maximum of 1,000 retrieved items. By using a mean to measure performance we give equal importance to all queries. To statistically determine whether or not a given search strategy is statistically better than another, we have applied the bootstrap methodology [15], with the null hypothesis H_0 stating that both retrieval schemes produce similar performance. In the experiments presented in this paper statistically significant differences were detected by a two-sided test (significance level $\alpha = 5\%$). Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected.

Table I
MAP OF VARIOUS IR MODELS AND DIFFERENT STEMMERS

	none	light	perstem	5-gram
tf idf	0.2966	0.2625*	0.2799*	0.2814
Lnu-ltc	0.4533	0.4277*	0.4369	0.4107*
Okapi	0.4811	0.4535*	0.4610*	0.4511
DFR-PL2	**0.4939**	**0.4693***	**0.4750***	**0.4423***
LM	0.4348	0.4000*	0.4113*	0.3892*

A. IR Models Evaluation

In the Table I are given the MAP using three different stemming approaches with five IR models. In the last column we have also included a language-independent indexing approach based on 5-gram [16]. Under this indexing scheme, words are decomposed by overlapping sequences of 5 letters. For example, the sequence "prime minister" generates the following 5-grams {"prime," "minis," "inist," ... and "ister"}.

The best performing IR model in Table I, is given in bold, and will be used as the baseline for statistical testing. Our experiments show that the IR model derived from *Divergence from randomness* (DFR) paradigm tend to produce best results under all indexing and stemming strategies. Compared to this best performance the differences are always statistically significant.

B. Stemming Evaluation

Even though Persian language uses affixation extensively in the process of word forming, the effect of their automatic removal must be evaluated.

As it can be seen from Table I we have first evaluated indexing strategy in which the stemming procedure was ignored, under the label "none" in this table. Then we give the results obtained by our light stemming approach (labeled "light"). Finally, in the column "perstem" are given results obtained by using stemmer and morphological analyzer for the Persian language. After applying different stemming strategies average number of distinct indexing terms per document changes from 127 with "none" to 119 with "light" and 118 with "perstem". If we use retrieval performance for which the stemming procedure was omitted, marked "none" in Table I as a baseline, we can see that both stemming strategies, "light" and "perstem", result in somewhat poorer results than that of the baseline. If we apply statistical testing using the same baseline we can see that these differences are mostly statistically significant (significant differences are marked with "*" after the MAP values). If we average the performance over five given models, we find a decrease of 4.28% when "perstem" is used and 6.67% with light stemmer. We have also tested a basic stemmer which removes only two most frequent suffixes that is ی and ه . Even this restrained stemming procedure results in decrease of performance of 3.08% compared to no stemming procedure. Explanation for this phenomenon can be partially found in a fact that large numbers of suffixes were already separated from the word they alter, due to the fact that they can be written together or separated, thus resulting in over stemming(e.g. stemmer is applying suffix stripping rules on the stem since the suffix is already separated from the root). Another explanation for such behavior can be found in the fact that in this language the same suffixes are used for different purposes. We can take as an example Topic #525 "Seasonal Diseases", the root فصل (season, term, article) of the adjective فصلی (seasonal) is used in compound constructions. These constructions in the Persian language, unlike in English (e.g., handgun, weekend) are written separately (e.g.,فصل انگورچینی - vintage, فصل کم کاري - layoff) causing non relevant documents to be retrieved in high positions (e.g., the MAP goes from 0.5941 with "none" to 0.4171 with "light"). This is the case with both indexing strategies "light" and "perstem". We have also encountered certain cases of over stemming. In the Topic #522 "Teheran metro project" the term متروي (the metro) is stemmed into متر by the light stemmer (which is also stem for the word "meter") resulting the MAP to go from 0.7787 with "none" to 0.2385 with "light".

Denoted as "5-gram" in Table I are shown retrieval performances of the given IR models when language independent 5-gram indexing strategy (without applying a stemming procedure). The performance difference between 5-gram indexing strategy and word-based no stemming indexing is -7.87% and is always statistically significant. Apart from 5-gram deprecated in Table I we have also tested different length n-gram, thus obtaining the relative difference, over five models, of -26.42% when comparing no stemming to 3-gram and -15.29% compared to 4-gram. This brings us to a conclusion that by augmenting the length of n-gram, thus approaching no stemming strategy the retrieval performance for this language is being increased.It is interesting to know that we have found only two topics, with no stemming indexing strategy and DFR-PL2 retrieval model that did not have any relevant documents among first ten retrieved. Those are Topic #574 "Champion team Iran first league'" retrieves first relevant document 14 position, the reason for this poor performance could be found in the fact that Persian version of this topic does not contain "Iran", resulting in retrieving documents talking about different sport leagues before the relevant ones. Second is the Topic #599 "2nd of Khordad election" having the first relevant score at 11th position. Number of topics not having any relevant documents among first ten retrieved stays stable across different indexing schemes.

C. Stopword List Evaluation

Finally, we have compared the retrieval effectiveness of the IR model with and without the stopword list. Average document length after stopword list removal for this language decreases from 381 to 202 terms per document. Apart

Table II
MAP OF VARIOUS IR MODELS USING TWO DIFFERENT STEMMERS WITH OR WITHOUT STOPLIST

	none		perstem	
	stoplist	no stoplist	stoplist	no stoplist
tf idf	0.2966	0.2449*	0.2799	0.2341*
Lnu-ltc	0.4533	0.4519	0.4369	0.4363
Okapi	0.4811	0.4806	0.4610	0.4591
DFR-PL2	**0.4939**	**0.4888**	**0.4750**	**0.4747**
LM	0.4348	0.4186*	0.4113	0.4034

from inverted file being reduced as well as query processing time, we can see from the Table II that, for this language, this procedure does not have significant impact on MAP performance. The difference is 2.74% when no stemming approach is in question and 3.47% when comparing "perstem" performance with and without stop list. If we use the approach in which the stopword list is not removed as a basline for statistical testing and compare it to the corresponding approach in which this step was performed the statistically significant differnces are marked with "*" after the MAP value in the Table II.

VII. CONCLUSION

In this paper, we present the main aspects of the Persian language morphology and we suggested stemmer for this language, which removes most frequent suffixes used for denoting article, plural or in certain cases for deriving different word types. We also suggested stopword list containing 881 terms. These linguistic tools are freely available on the Internet. Using the most effective current IR models, we have evaluated different stemming approaches and found that the best performing IR model is derived from *Divergence from Randomness* (DFR) paradigm. This approach performs statistically better when compared to any other retrieval model used in this study.

Our various experiments clearly show that a stemming procedure decreases retrieval effectiveness when applied to the Persian language. From a statistical point of view, the differences are always significant when no stemming approach is compared to an approach that incorporates stemming procedure.

From comparing different stemming strategies, with or without stopword list removal it seems that including this procedure into indexing process produces slightly better MAP than does the same indexing strategy with it being omitted. The differences between corresponding cases are rarely statistically significant.

Acknowledgments: This research was supported in part by the Swiss NSF under Grant #200021-113273.

REFERENCES

[1] C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. Oard, A. Peñas, and D. Santos, Eds., *Advances in Multilingual and Multimodal Information Retrieval*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2008, vol. 5152.

[2] D. Harman, "Beyond english," in *TREC Experiment and Evaluation in Information Retrieval*, E. Voorhees and D. Harman, Eds. Cambridge , MA: The MIT Press, 2005, pp. 153–182.

[3] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[4] J. Savoy, "Stemming of french words based on grammatical category," *Journal of the American Society for Information Science*, vol. 44, no. 1, pp. 1–9, 1993.

[5] J. Xu and B. Croft, "Corpus-based stemming using cooccurrence of word variants," *ACM-Transactions on Information Systems*, vol. 16, no. 1, pp. 61–81, 1998.

[6] S. Tomlinson, "Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003," in *Comparative Evaluation of Multilingual Information Access Systems*, ser. Lecture Notes in Computer Science, vol. 3237. Berlin: Springer-Verlag, 2004, pp. 286–300.

[7] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the ACM-CIKM*. Washington ,DC: The ACM Press, 2004, pp. 625–633.

[8] L. Elwell-Sutton, *Elementary Persian Grammar*. Cambridge,UK: Cambridge University Press, 1963.

[9] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge,UK: Cambridge University Press, 2008.

[10] A. Singhal, "AT&T at TREC-6," in *25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, 2002, pp. 35–41.

[11] S. E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at TREC," *Information Processing & Management*, vol. 36, pp. 95–108, 2000.

[12] 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, pp. 357–389, 2002.

[13] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM-Transactions on Information Systems*, vol. 22, no. 2, pp. 179–214, 2004.

[14] D. Hiemstra, "Using language models for information retrieval," Ph.D. dissertation, CTIT, 2000.

[15] J. Savoy, "Statistical inference inretrieval effectiveness evaluation," *Information Processing & Management*, vol. 33, 1997.

[16] P. McNamee and J. Mayfield, "Character *n*-gram tokenization for European language text retrieval," *IR Journal*, vol. 7, no. 1-2, pp. 73–97, 2004.