2009

Improving Persian Information Retrieval Systems Using Stemming and Part of Speech Tagging

Reza Karimpour
University of Tehran

Amineh Ghorbani
University of Tehran

Azadeh Pishdad
University of Tehran

Mitra Mohtarami
University of Tehran

Abolfazl Ale Ahmad
University of Tehran

See next page for additional authors

Publication Details

This conference paper was originally published as Karimpour, R., Ghorbani, A., Pishdad, A., Mohtarami, M., Aleahmad, A., Amiri, H. & Oroumchian, F. 2009, 'Improving Persian information retrieval system using stemming and part of speech tagging', in Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, 17-19 Sept.2008, Aarhus, Denmark, Lecture Notes in Computer Science, vol. 5706, pp. 89-96.

Authors
Reza Karimpour, Amineh Ghorbani, Azadeh Pishdad, Mitra Mohtarami, Abolfazl Ale Ahmad, Hadi Amiri, and Farhad Oroumchian

Improving Persian Information Retrieval Systems Using Stemming and Part of Speech Tagging

Reza Karimpour[1], Amineh Ghorbani[1], Azadeh Pishdad[1], Mitra Mohtarami[1],
Abolfazl AleAhmad[1], Hadi Amiri[1], Farhad Oroumchian [2]

[1] Electerical and Computer Engineering Faculty, University of Tehran
[2] University of Wollongong in Dubai

{r.karimpour, a.ghorbani, a.pishdad, m.mohtarami, a.aleahmad, h.amiri}@ece.ut.ac.ir,
farhadoroumchian@uowdubai.ac.ae

Abstract. With the emergence of vast resources of information, it is necessary to develop methods that retrieve the most relevant information according to needs. These retrieval methods may benefit from natural language constructs to boost their results by achieving higher precision and recall rates. In this study, we have used part of speech properties of terms as extra source of information about document and query terms and have evaluated the impact of such data on the performance of the Persian retrieval algorithms. Also the effect of stemming has been experimented as a complement to this research. Our findings indicate that part of speech tags may have small influence on effectiveness of the retrieved results. However, when this information is combined with stemming it improves the accuracy of the outcomes considerably.

Keywords: Natural language; Persian information retrieval; Part of speech;

1 Introduction

Exploiting meta-information of the terms in the retrieval process may result in precision and recall improvements. Part of speech information clarifies the role of each term in queries and documents. It may also help in assigning different priorities to different query terms. In addition stemming can collapse many surface words in languages such as Arabic and Persian into a single representation and improve the recall of the system.

The general objective of the present study is to further investigate the potential benefits of incorporating part of speech information into both query and document processing and to observe the consequences of such incorporation in Persian information retrieval. Another objective is to investigate the interaction of stemming and part of speech tagging in such environment.

Improving the performance of retrieval engines has been a major concern for years leading to development of many efficient and effective algorithms and systems [1, 2, 3]. However, the retrieval effectiveness of some European languages such as English have been studied in more depth than Middle Eastern languages such as Persian (Farsi). In addition document retrieval has been an interesting topic for those working in natural language processing (NLP) [4, 5] but not much work has been done on the use of these techniques for Persian document retrieval.

In recent years, there has been some interest on Persian information retrieval but none of those approaches have used part of speech tagging, although POS has been applied successfully to information retrieval in other language [6]. On the other hand studies in Persian POS tagging have reported accuracy rates of up to 95% using statistical methods such as TnT or with post-processing with MLE taggers [7, 8, 9, 10]. Therefore it seems reasonable to use these taggers in the development of a new generation of retrieval engines for Persian language.

In this research we utilize POS tagging methods to preferentially match the specified types of terms in documents and queries. We also try to control the impact of certain types of words that seems not to have a major contribution to the overall results.

2 Part of Speech Tagging

Part of speech tagging selects the most likely sequence of syntactic categories for the words in a sentence. It determines the tags that best represent the grammatical characteristics of the words, such as part of speech, grammatical number, gender, person, etc. This task is not trivial since many words are ambiguous. Most of the retrieval models ignore the role of the content words in the sentences and treat them uniformly. Although a lot could be realized from the role a word plays in a sentence and its surrounding words. Besides this, the role of each word depends on what the user means by the words in the query [11].

In different languages and tagging systems, the number of tags varies from a dozen to several hundred depending on the specificity of the information provided by the tag. For example a tag-set may just categorize nouns as singular and plural while another tag-set may provide more detail such as *name of location* or *person*. Obviously, not all of these tags have the same impact on the retrieval of the relevant documents [12]. Therefore the computation of a proper tag-set with the right size and granularity for a particular collection of a language is an issue worthy of studying.

In this study, we take advantage of the Bijankhan [13] corpus which is a manually tagged Persian text collection. In its original form it includes 550 different tags. This collection has been processed and prepared for machine learning applications. The new collection has over 2 million words and only 40 POS tags [7].

It has been reported that in some applications of IR, nouns are more important than the other tokens [14, 15]. However, sometimes even stop words can be useful [14]. The importance of various POS tags is very subjective. For example in some areas such as biology or advertisement that emphasize the differences among things and their characteristic, adjectives are more important. While in other applications such as music which are mostly adverb-rich, the role of adverbs become more important [11]. Some studies also have investigated the role of verbs in document analysis [16].

After analyzing the impact of these 40 different tags, eventually we find out that nouns, verbs, adjectives and adverbs are the most important POS Tags in Persian retrieval. In the result this section we will show the impact of using these tags on the performance.

In this study the TnT POS tagger[1] is used to determine the part of speech of Persian words. TnT is a very efficient statistical part-of-speech tagger that is trainable on different languages and virtually any tag-set. TnT requires a pre-tagged document collection for training phase. The system incorporates several methods of smoothing as well as handling unknown words. Employing the tagger to either a new language or new tag-set is a simple process [17].

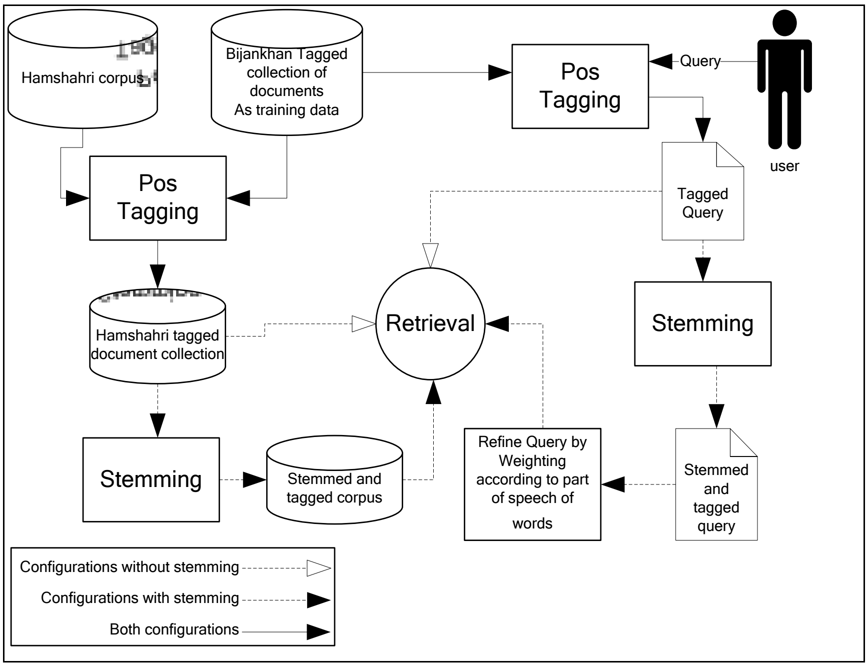
Hamshahri corpus
Bijankhan Tagged collection of documents As training data
Pos Tagging
Query
user
Pos Tagging
Tagged Query
Hamshahri tagged document collection
Retrieval
Stemming
Stemming
Stemmed and tagged corpus
Refine Query by Weighting according to part of speech of words
Stemmed and tagged query
Configurations without stemming
Configurations with stemming
Both configurations

Figure 1. The Framework of our approach.

3 Methodology & Implementation

This study and experiments have been conducted as part of the Persian track at CLEF 2008 [18]. As a retrieval engine, we have utilized the Indri retrieval system [3] which is provided as part of the Lemur project[2]. TNT POS tagger was trained on Bijankhan POS collection with 40 tags. Subsequently the Hamshahri corpus [19] and its CLEF queries were tagged using this tagger (see Figure 1).

After experimenting with different tagging schemas, the corpus and the queries were stemmed in order to evaluate the effect of stemming and its interaction with

[1] TnT can be found at http://www.coli.uni-saarland.de/~thorsten/tnt/

[2] The Lemur Project. 2001-2008. University of Massachusetts and Carnegie Mellon University. [www.lemurproject.org]

POS tagging in retrieval context. Stemming was performed by employing simple grammatical rules using PERSTEM Persian stemmer [20]. Consequently we prepared 4 different variations of the Hamshahri document collection which included, normal (neither stemmed nor tagged), stemmed, tagged (terms tagged with related parts of speech), and both stemmed and tagged.

We conducted two types of experiments. In one set of experiments all terms were treated equally. That is, there were no preferences among the term types except for their statistical weight calculated by the Indri system. In the second set of experiments, we defined preferences among the term types based on their POS tags. For example in one experiment, nouns could have received a weight of 3 while verbs might have received a weight of 1, which means that the nouns were given three times more importance than the verbs. Experiments also differed based on what sections of the queries were used. Some experiments used only the title section of the queries and some others used both the title and the description sections of the CLEF queries. Table 1 lists the configurations used in our experiments.

Table 1. Different configurations

Config.	*Corpus*	*Query*
1	Normal	Title (Neither stemmed nor tagged)
2	Tagged	Title with equal weighting for all POS tags
3	Tagged	Title plus description with equal weighting for all POS tags
4	Stemmed	Stemmed title without POS tagging
5	Stemmed	Stemmed Title plus description
6	Stemmed (stop words removed)	Stemmed Title plus description (stop words removed)
7	Stemmed and tagged	Stemmed title with equal weighting for all POS tags
8	Tagged	Title with various weighting schemes for different POS tags

4 Results

Before discussing the results, it should be noted that since the Hamshahri collection has tagged automatically as described above we do not have any measurement of the accuracy of the tagging yet, however basic observations and sampling has shown reasonable accuracy.

Table 2 summarizes the outcomes of our experiments. The result of the base line system without employing tagging or stemming has an average precision of 27% and R-precision at 36%. By matching the tagged corpus with the tagged title of the queries the average precision climbs to 35% and the R-Precision increases by 1%. This is an interesting result since no part of speech preferences has been implemented in this run. This search is based on matching similar terms with similar roles in documents and queries. When the description field of the queries is added to the model, the performance of the system experiences a minor setback with the average precision at 29% which is still higher than the normal corpus and the R-Precision declins to 34% which again is a little higher than that of the normal retrieval performance. Generally we observed that adding descriptions in all configurations would degrade the performance of the system. The reason for this reduction is the negative effect of the extra terms in the query description that misleads the retrieval. We concluded that these misleading terms add more ambiguity than those POS tags can clarify.

Table 2. Main Results

Config.	*Average precision*	*R-Precision*
Normal corpus	0.2716	0.3627
Tagged (title)	0.3505	0.3784
Tagged (title + description)	0.2989	0.3497
Stemmed title	0.3625	0.4102
Stemmed (title + description)	0.1723	0.2157
Stemmed(title + description + stop words)	0.1672	0.2106
Stemmed and tagged (title)	0.3944	0.4151
Different weightings (average)	0.2263	0.2655

The results we obtained indicate that Persian retrieval benefits from stemming. Stemming the documents and queries alone returned one of the best results of our experiments with the average precision at 36% and R-Precision at 41%. This is in contrast with experiments conducted before by other groups in University of Tehran on the same corpus. However, when the title and description were used as query, the performance fell sharply. This configuration had one of the worst performances, even lower than the base line system. The reason of this poor outcome again was the extra text in the description which seemed to be too general and ambiguous. In this case stemming made the situation worse because it collapsed many surface words into a single representation and added to the ambiguity. In general, the effect of the stemming in Persian retrieval is still a research question. More experiments need to be performed with different types of stemmers as well as further scrutinizing the stemming techniques and their effect on Persian text retrieval. At the moment our conclusion is that the aggressive stemming is not useful and the simple stemming is sufficient.

Stop word removal is normally a very powerful tool in improving the precision. However, when stop word removal was applied to stemming of the title and the description of the queries, it did not improve the precision.

The best result of our experiments was achieved by stemming the tagged corpus and the title of the queries. This configuration produced an average precision of 39% which was the best. The R-precision in this case stays at 41%. In other words, combining simple stemming and part of speech tagging improves the average precision but does not change the R-precision. This shows that the stemming is more powerful than the part of speech tagging when it comes to precision.

Table 3. Weighting schemes

Noun	*Verb*	*Adjective*	*Adverb*	*Average Precision*	*R-Precision*
3	2	1	1	0.2635	0.3097
3	0	3	0	0.2597	0.2888
0	2	0	0	0.1108	0.1256
0	0	1	0	0.1198	0.1186
0	0	0	1	0.0977	0.1111
20 less used tags omitted, others equal weight				0.2745	0.3097

We explored the idea of POS tag preferences and their effect on precision. In these experiments, a weight of zero to three was given to each POS tag which then was multiplied with the actual weight of the term itself. So, we could emphasis or de-emphasis the contribution of the terms with certain part of speech tags. We explored many different combinations of preferences for different tags but in general we did not find any meaningful improvement in these experiments. Yet, on the contrary we found that many combinations have strong negative effect on precision. Table 3 depicts the results of some of these experiments. Assigning a weight of zero to a tag is the same as omitting the terms with that tag from the corpus and the queries. For example, (Noun=3, Verb=2, Adjective=1, Adverb=1) means the terms that are noun have been weighted three times the terms that are adjective or adverb. Similarly, the terms that are verbs have been weighted twice those of adjectives or adverbs. We also carried out experiments on the contribution of each tag to the overall performance of the retrieval. In some experiments as much as 20 least significant tags were omitted from the queries but it negatively affected the precision and recall. In general the average precision for all the tag weighting schemes was 0.22 and the average R-Precision was 0.26. The best run achieved a precision of 0.26 with R-precision of 0.31 which is much lower than one can achieve by simple stemming. The reason for such behavior can be explained by the importance of different tags in the Persian language. Despite our original study that led us to the omission of the 20 least important tags, they actually played a role in the retrieval. Thus omitting them or down playing their contribution declines the performance of the system.

5 Conclusion and Future Work

This study attempted to measure the effectiveness of part of speech tags and stemming on Persian information retrieval. Different configurations were tested and the results demonstrated that retrieving documents by matching the terms and their part of speech in documents with the terms and their part of speech in queries improves the performance. However, it was evident that while some parts of speech are more important than others, eliminating the least important ones or reducing their overall impact on the query processing degrades the performance of the system. The best results were achieved by giving equal importance to all POS tags.

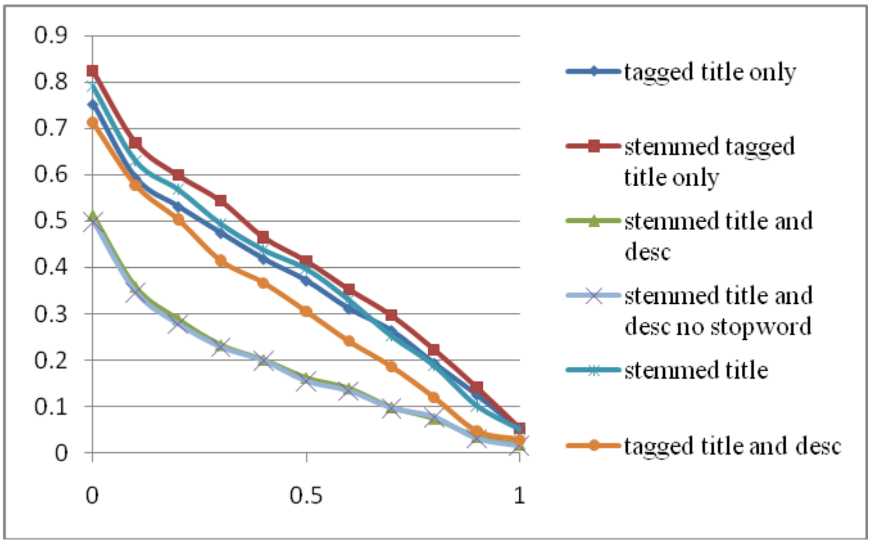

Figure 2. R-Precision of the different configurations

The effect of stemming was also studied and it became clear that simple stemming in these experiments greatly improves precision. This study also observed that combining simple stemming and POS matching yields the best performance.

A future study would be utilizing retrieval models and systems other than Indri in order to make sure that the obtained results are not system dependent. However, given our previous experiences with different retrieval models on Persian language, we do not consider this as a major issue.

References

1. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images. Information Theory, IEEE Transactions on, **41**(6) (1995)
2. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In Proc. 19th ACM SIGIR, pp. 21-29. ACM New York, NY, USA (1996)
3. Strohman, T., Metzler, D., Turtle, H., Croft, W.: Indri: A Language-Model Based Search Engine for Complex Queries. Technical Report IR-407, CIIR, UMass Amherst (2005)
4. Liddy, E.D.: Automatic Document Retrieval. Encyclopedia of Language and Linguistics. Elsevier Press (2005)
5. Lewis, D., Jones, K.: Natural Language Processing for Information Retrieval. Communications of the ACM, **39**(1), 92-101 (1996)

6. Amiri, H., AleAhmad, A., Oroumchian, F., Lucas, C., Rahgozar, M.: Using OWA Fuzzy Operator to Merge Retrieval System Results. Computational Approaches to Arabic Script-based Languages, (2007)
7. Amiri, H., Hojjat, H., Oroumchian, F.: Investigation on a Feasible Corpus for Persian POS Tagging. In Proc. 12th International CSI Computer Conference (CSICC) (2007)
8. Raja, F., Amiri, H., Tasharofi, S., Sarmadi, M., Hojjat, H., Oroumchian, F.: Evaluation of Part of Speech Tagging on Persian Text. The Second Workshop on Computational Approaches to Arabic Script-Based Languages, Stanford University, U.S.A. (2007)
9. Mohtarami, M., Amiri, H., Oroumchian, F.: Using Heuristic Rules to Improve Persian Part of speech Tagging Accuracy. In Proc. 6th International Conference on Informatics and Systems, INFOS2006 (2006)
10. Oroumchian, F., Tasharofi, S., Amiri, H., Hojjat, H., Raja, F.: Creating a Feasible Corpus for Persian POS Tagging. Technical Report, No. TR3/06, University of Wollongong (Dubai Campus) (2006)
11. Shah, C., Bombay, I.I.T., Mumbai, P., Maharashtra, I., Bhattacharyya, P.: A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR). In Proc. International Conference on Universal Knowledge and Languages (ICUKL) (2002)
12. Carlberger, J., Kann, V.: Implementing an Efficient Part-Of-Speech Tagger. Software Practice and Experience, **29**(9), 815-832 (1999)
13. BijanKhan, M.: The Role of the Corpus in Writing a Grammar: An Introduction to a Software. Iranian Journal of Linguistics, **19**(2) (2004)
14. Turney, P., Littman, M.: Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. National Research Council of Canada (2002)
15. Paik, W., Liddy, E., Yu, E., McKenna, M.: Interpretation of Proper Nouns for Information Retrieval. In Proc. Workshop on Human Language Technology, pp. 309-313. Association for Computational Linguistics Morristown, NJ, USA (1993)
16. Klavans, J.L., Kan, M.Y.: The Role of Verbs in Document Analysis. In Proc. Coling-ACL, Vol. 36, pp. 680-686. Association for Computational Linguistics (1998)
17. Brants, T.: TnT–a Statistical Part-of-Speech Tagger. In Proc. 6th Conference on Applied Natural Language Processing (ANLP-2000), pp. 224-231, Seattle, WA (2000)
18. Agirre, E., Nunzio, G.M.D., Ferro, N., Mandl, T., Peters, C.: Multilingual Textual Document Retrieval (Ad Hoc), in Evaluating Systems for Multilingual and Multimodal Information Access. In Proc. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus (2008)
19. Aleahmad, A., Hakimian, P., Mahdikhani, F., Oroumchian, F.: N-gram and Local Context Analysis for Persian Text Retrieval. In Proc. IEEE International Symposium on Signal Processing and its Applications, pp. 1-4, Sharjah, UAE (2007)
20. Dehdari, J., Lonsdale, D.: A Link Grammar Parser for Persian. Aspects of Iranian Linguistics, Vol. 1. Cambridge Scholars Press (2008)