

An Efficient Set of Parts of Speech in Persian Information Retrieval Systems

Mohammad-Ali Yaghoub-Zadeh-Fard^a, Saeed Rahmani^b, Omid Kashefi^c,
Behrouz Minaei-Bidgoli^d

^a *Iran University of Science and Technology.*

^b *Shiraz University.*

^c *Iran University of Science and Technology.*

^d *Iran University of Science and Technology.*

Abstract

Even though the ultimate aim of any information retrieval system is to fulfil its users' expectations, reducing index storage size and enhancing the system performance are sometimes infinitely preferable, especially for small-sized companies suffering from a lack of hardware resources. For such companies, it is of paramount importance to remove noninformative terms from their indices. Selecting a proper set of terms makes it possible to reduce the index storage size and consequently enhance the retrieval performance. In this paper, using parts of speech tagging, we show how to reduce the index storage size without losing precision. Through an experimental process and using Hamshahri corpora, we identify the most effective parts of speech in Persian language. Results demonstrate improvements in the response time and precision of the retrieval.

Keywords: Information Retrieval, Natural Language Processing, Part of Speech, Index storage reduction, Term selection, Stop-word detection

Introduction:

By the advent of vast information resources, users need to have faster and more accurate access to a greater variety of information that is dispersed all over the world. Now it is an undisputable fact that the amount of information is fast growing. Consequently, as time goes by, the need to develop more efficient information retrieval (IR) systems which are capable of handling a large amount of data is getting more and more indispensable than before [1]. Owing to the fact that the larger our index is, the less chance we have to accommodate it in the random access memory, the size of index has a direct influence on the response time. Therefore, by reducing of index size, the overall performance of IR systems will be improved [2]. However, pruning indices may eventuate in losing some information and as a result, decreasing accuracy.

One way to reduce the index storage size is to eliminate the terms which are not contributing to the information content of a document. At very least we can remove all stop-word. Stop-words occurs frequently in virtually in all documents and do not carry any information. Another approach to prune the index is using term selection methods. These methods have been used to select the appropriate terms for a document in various text processing systems. Moreover, some techniques have been borrowed from natural language processing (NLP) to choose set of terms for a document that show the content of it. Chunking, summarization, partial parsing, part of speech (POS) tagging are a small part of NLP techniques which have been used in IR widely [3-7]. Among these techniques, POS tagging plays a crucial role in almost all NLP tasks. POS tagging is the task of labeling each word in a sentence with its appropriate part of speech like as a noun, verb, adverb, adjective, conjunction, determiner, interjection, numeral, particle, preposition, and pronoun [1]. POS tagging has been used to improve IR performance in different languages such as English and Arabic[8, 9].

The role of any parts of speech is subjective [8, 9]. Not considering different senses of terms, nouns are meaningful by themselves; as a result, it is appropriate to build indexes with nouns. On the other hand, adjectives and adverbs are less useful than nouns because they work mainly as complements [9]. Nevertheless, in some areas such as biology or advertisement adjectives are more important. Therefore, with a proper set of terms, it is possible to reduce index storage size and consequently, enhance retrieval speed and system performance.

Determination of the effective set of POS tags could be useful in many fields such as IR and text classification. This paper shows how to reduce the index size without losing the IR precision and recall using POS information. Through an experimental process we identify the most effective set of tags for Persian IR.

The remainder of this paper is organized as follows: Section 3 is an overview of the related works. Section 3 introduces Persian language, and describes the most important parts of speech in Persian. Section 4 describes methods for examining the efficiency of particular POS and gives the efficient set of parts of speech in Persian IR. Section 5 evaluates our methodology using the Hamshahri corpus. Finally, Section 6 concludes the paper.

Related Works

The role of POS tagging have been investigated in various languages [2, 8, 9, 11]. To begin with, experiments on English and Arabic indicate that nouns more meaningful than other terms [2, 9]. Chowdhury and et al. in [8] indexed using only nouns and reduced the term dictionary by 60%, since only

40% of their terms were nouns. By doing so, they saved 28% of the index storage size. In the same way, Ghassan Kanaan and his colleagues removed 45% of terms in the index, only by keeping nouns [9]. Both for Arabic and English IR systems, by using only nouns, they faced only 1 percent reduction in mean precision. Moreover, noun phrases and proper nouns are usually noticed as key indexing terms in some studies [12, 13]. Barr et al. [2] leveraged POS tagging towards improved web queries. They investigated the applicability of POS tagging to typical English web search-engine queries and the potential value of these tags for improving search results. They found that proper-nouns constitute 40% of query terms, and over 70% of query terms are nouns or proper nouns. They also showed that the majority of queries are noun phrases. The role of verbs is also investigated by Klavans [14]. The linguists believe that verbs complete the nouns and adverbs complete the verbs [14].

Moreover, POS tagging has been used to weight terms [15, 16]. According to this approach, a weight is assigned to each term based on its POS tag, and the most common way to weight POS tags is to use mutual information [15] which is borrowed from Shannon's information theory [17]. The mutual information of two random variables is a quantity that measures the mutual dependence of the two variables.

Another use of POS tagging in IR is to disambiguate query terms. For instance, Allan and Raghavan [18] worked on POS patterns to reduce query ambiguity. POS patterns represent sequences of POS tags that occurred around the query terms. They showed that their method improves the mean average precision (MAP) by 41% for standard topics of TREC-5 queries and 25% for web queries collected from Yahoo search engine.

Hamshahri Corpus

Persian is one of the dominant languages in the Middle East, which is spoken in some countries such as Iran, Afghanistan, and Tajikistan. It is an Indo-European language and contains 32 letters. English and Persian languages have some differences; they are different in sentence structure, alphabet, etc. Due to different nature of Persian to other languages, designing IR systems for Persian language requires special considerations [10].

The Hamshahri corpus [10] is one of the most credible IR corpora in Persian language. It is a standard reliable Persian text collection that was used at Cross Language Evaluation Forum (CLEF) during years 2008 and 2009 for evaluation of Persian IR systems. Documents of this collection are prepared by crawling, preprocessing, and tagging of the Hamshahri newspaper website. The Hamshahri corpus is released in two versions; the first version, which is about 700MB, includes 100 standard queries, and the last version owns 50 queries with their relevant judgment. The last version is over 1400MB.

As it is mentioned before, various parts of speech play different roles in different systems. Therefore, in order to conduct a perfect experiment, it is of paramount importance to use a comprehensive corpus that includes documents from different subjects. We used Hamshahri corpus that include various document categories to evaluate general role of each POS in Persian IR. Table 1 shows the most important parts of speech for Persian language.

Figure 1 reflects the breakout of terms by their parts of speech in the Hamshahri corpus. It shows that nouns and adjectives constitute 41% and 13% of all terms in this corpus respectively.

Efficient Set of Persian POS Tags

In this section, we propose our methodology and experiments for determining the most informative POS tags in Persian IR.

Proposed Approach

Our goal is to select an efficient set of POS tags for IR tasks, and according to our definition the most efficient POS set is a set of POS tags which are used to represent documents while satisfying two goals: first reduces the index storage size as much as possible, and second does not cause any reduction in IR precision. In other words, we want to reduce the index size but preserve precision.

To determine an efficient set of POS tags for Persian IR, we consider different configurations using POS tags described in Table 1. In each of those configurations, a certain POS tag is used to construct the index; for example, one of the configurations only uses *nouns* to represent documents and ignores other terms. In this table, each configuration reveals the discrimination value of the corresponding POS tag. In other words, it shows how presenting documents just by words tagged by a special POS can discriminate between relevant and irrelevant documents.

After calculating the MAP for each configuration, the most valuable POS tag is selected; then other effective POS tags based on their IR precisions are added to the experiment. This process continues until IR precision stops increasing. Finally, we have an efficient set of POS tags which representative of all information content of all documents.

Table 1. Persian major parts of speech

POS	Description	Example
<i>Nouns</i>	Persian words that indicate people, beings, things, places, phenomena, qualities or ideas.	محمد، کسوف
<i>Verbs</i>	Persian words that indicate actions, occurrences or states.	رفتن، پخت
<i>Adverbs</i>	Persian words that modify clauses, sentences and phrases directly.	هیچ وقت، هنوز
<i>Adjectives</i>	Persian words that give attributes to nouns, extending their definitions.	خوش، خوب
<i>Conjunctions</i>	Persian words that connect words, phrases or clauses together.	اما، اگرچه
<i>Determiners</i>	Persian words that reference nouns, expressing their contexts directly.	آن، این
<i>Interjections</i>	Persian words that express emotions as exclamations.	ای، به به
<i>Numerals</i>	Persian words that indicate quantities of nouns.	هشت، دوازده
<i>Prepositions</i>	Persian words that connect nouns or pronouns that follow them to form adjectival or adverbial phrases.	بر، با
<i>Pronouns</i>	Persian words that refer to and substitute nouns.	آن، کدام یک

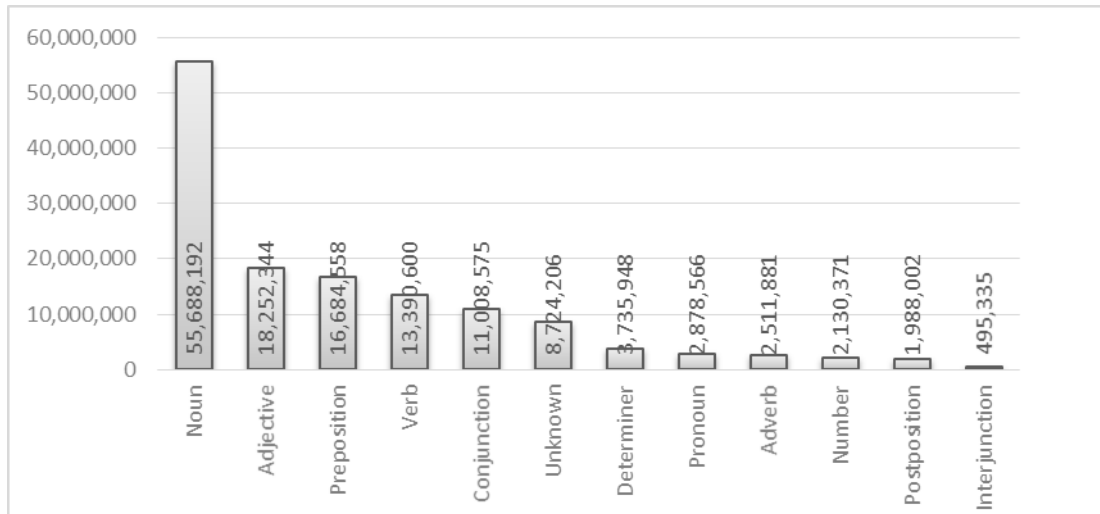


Fig. 1. Hamshahri part of speech breakout

Case Study

To determine and evaluate the proposed approach, we partitioned the Hamshahri corpus to five equal parts to perform 5-fold cross evaluation; then we repeated the algorithm for each partition. Moreover, we use Okapi BM25 [19] which is a probabilistic IR model.

Different configurations and the amount of effectiveness of each configuration in one of the iterations of our 5-fold evaluation process is shown in Table 2, where *Indexing Tag* column illustrates the POS tag of words that construct the index, and MAP is the mean average precision. For example, the configuration 1 only uses nouns to represent the documents and omits other terms. To evaluate the exact role of each POS in IR, any text preprocessing techniques such as stop-word removal or stemming are not employed in this experiment. However, we will finally evaluate our method using a standard stop-word list and applying lemmatization.

According to Table 2 nouns, adjectives, and unknown-words, which include proper names, words with spelling errors, or words from other languages, are more discriminative. This ranking of effectiveness of POS tags is for one of the iterations. However, our experiments show that it is quite the same story for other iterations.

The noteworthy result is the relatively small effect of adverbs and specially verbs in Persian IR. However, it has been reported that verbs and adverbs in English IR play an significant role [8, 14]. This must be due to the different nature of the verbs in Persian. Verbs usually consist of one or more non-verb parts; usually nouns followed by a verb part. Information that non-verb part(s) of compound verbs carries is more effective than verb part. However, in our experiments, just verb-parts of compound verbs is considered as verb.

Table 2. Effectiveness of different POS tags in one of the folds

Configuration	Indexing Tag	MAP
1	<i>Noun</i>	0.2248
2	<i>Adjective</i>	0.0874
3	<i>Unknown</i>	0.0350
4	<i>Number</i>	0.0030
5	<i>Verb</i>	0.0007
6	<i>Preposition</i>	0.0006
7	<i>Pronoun</i>	0.0006
8	<i>Conjunction</i>	0.0002
9	<i>Adverb</i>	0.0002
10	<i>Determiner</i>	0.0001
11	<i>Classifier</i>	0.0000

Table 3. MAP of retrieval using different POS sets in one of the folds

Indexing Terms	MAP
<i>All words</i>	0.2153
<i>Nouns</i>	0.2248
<i>Nouns and adjectives</i>	0.2371
<i>Nouns, adjectives, and unknown words</i>	0.2499
<i>Nouns, adjectives, unknown words, and numbers</i>	0.2563
<i>Nouns, adjectives, unknown words, numbers, verbs</i>	0.2550

The performance of different POS sets is shown in Table 3. The results show that using nouns, adjectives and unknown-words is as effective as using all terms, and the best precision is achieved by the set of {Nouns, Adjectives, Proper-nouns, Number}, interestingly, even better results in comparison with indexing all of the terms. Table 3 reveals that indexing only nouns preserve remarkable IR performance, the same conclusion reported for Arabic and English IR systems [8, 9].

Evaluation

In this section, we are going to determine the efficient set of POS tags and evaluate it. To begin with, let's look at Table 4. This table shows the retrieval performance for all the iterations we mentioned in the previous section. The best precision is achieved by the set of {Nouns, Adjectives, Proper-nouns, Number}, even better results in comparison with using all of the terms. This set improves response time by 119%, reduces index size by 26%, and even enhances MAP by 19% in comparison with the traditional indexing. It indicates that using the efficient set of POS tags enhance the performance of Persian IR significantly.

Figure 2 also shows precision at different thresholds for the five folds using the efficient set in comparison with indexing all words. It also indicates that using the efficient set have a better influence on top rank documents.

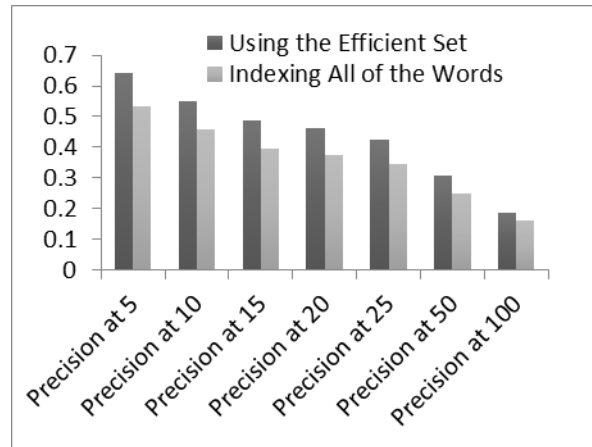


Figure 2. Mean precision at different thresholds for the five folds

Table 4. System performance for the five Hamshahri folds

Fold	Indexing All of the Terms			Using the Efficient Set				
	MAP	R-Precision	Recall	MAP	R-Precision	Recall	Speed-Up	Storage Reduction
1	0.2071	0.3359	0.7695	0.2518	0.4179	0.8497	113%	26.6%
2	0.2233	0.3401	0.7411	0.2650	0.4124	0.8376	124%	27.2%
3	0.2153	0.3641	0.7757	0.2563	0.4385	0.8882	117%	26.9%
4	0.2329	0.3491	0.7365	0.2789	0.4360	0.8296	125%	26.8%
5	0.2230	0.3685	0.7533	0.2617	0.4383	0.8483	118%	25.9%
Average	0.2203	0.3515	0.7552	0.2627	0.4286	0.8507	119%	26.7%

Table 5. System performance for the five Hamshahri test folds after lemmatization

Fold	Indexing All of the Terms			Using the Efficient Set				
	MAP	R-Precision	Recall	MAP	R-Precision	Recall	Speed-Up	Storage Reduction
1	0.2122	0.3462	0.7781	0.2555	0.4379	0.8691	117%	13.8%
2	0.2104	0.3573	0.7535	0.2577	0.4165	0.8645	127%	14.3%
3	0.2299	0.3621	0.7662	0.2724	0.4385	0.8615	109%	12.5%
4	0.2103	0.3522	0.7534	0.2543	0.4240	0.8648	112%	12.6%
5	0.2187	0.3552	0.7722	0.2682	0.4593	0.8650	131%	15.0%
Average	0.2163	0.3546	0.76468	0.26162	0.43524	0.86898	119%	13.64%

Table 6. System performance for the first version of Hamshahri corpus applying stop-word elimination along with lemmatization

Method	Using the Efficient Set		
	MAP	R-Precision	Recall
<i>Traditional indexing using wtop-word elimination and lemmatization</i>	0.4354	0.4281	0.5584
<i>Using the efficient set along with wtop-word elimination and lemmatization</i>	0.4409	0.4346	0.5549
Improvement	1.3%	1.5%	-0.6%

Effects of Lemmatization

As mentioned in the previous sections, verbs and some other tags do not really contribute in presenting document content in Persian. Perhaps it is because of a large number of possible declensions and conjugations in Persian. In this part of the experimentation, we lemmatize the verbs along with other inflected words to find out if it affects the efficient set of POS tags for Persian IR systems. We lemmatize words to their dictionary forms. Performing 5-fold cross evaluation and employing lemmatization, we evaluated the proposed approach using the Hamshahri corpus. The results are shown in Table 5.

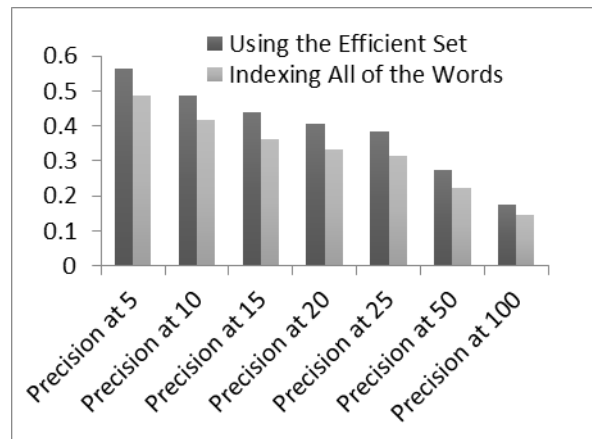


Fig. 2. Mean of precision at different thresholds in the five folds

Our experiments show that the best precision is almost achieved by the set of {Nouns, Adjectives, Unknown-words, Verbs, Numbers}. The only difference with the efficient set we built without using lemmatization is the presence of verbs in this new set. It means stemmed verbs contribute to the information contents of pages. This set improves the retrieval speed by 120%, reduces the indexing size by 14%, and even improves MAP by 21%. Figure 5 shows precision at different thresholds for the five folds using the efficient set in comparison with indexing all words.

Effects of Stop-word Removal

Let's back to the real world. It is wiser to evaluate our methodology applying the basic pre-processing method: stop-words elimination. To evaluate our methodology, this time, we used the first version of Hamshahri corpus and run the 100 standard CLEF queries. Applying Stop-word elimination along with lemmatization, Table 6 shows system performance for the first version of Hamshahri corpus using traditional indexing method in comparison with the efficient set of POS tags –{Nouns, Adjectives, Unknown-words, Numbers, Verbs}. According to our experiments, using the efficient POS set improves the IR mean average precision by 1.3%, although it decreases the recall by 0.6%. Moreover, the index storage size is reduced by 8.4% and the retrieval speed is improved for about 29.3%.

Conclusion

In this paper, we investigated the effects of indexing using terms with certain POS tags in Persian IR. Through experimental evaluation we determined two sets of POS tags in Persian and built indices using these sets. We showed that the proposed approach preserves the same or even superior IR precision compared to the traditional indexing. In addition, indexing the terms belong to one of those sets led to remarkably decrease in index size and increase in IR speed, which both are critical IR challenges for small-sized companies.

The experiments show that nouns, adjectives, unknown-words (including proper names, words with spelling errors, or words from other languages), numbers, and verbs are the most effective POS tags in Persian IR. It also implies that in Persian, substantial amount of semantic information are embedded in just nouns, and in the next order adjectives carries the meaning of speech. It must be noted that compound verbs, which constitute most of Persian verbs, are divided to noun and verb parts, and noun parts were added to the noun category.

References:

- Abolfazl, A., et al., Hamshahri: A standard Persian text collection. *Knowledge-Based Systems*, 2009. 22(5): p. 382-387.
- Allan and Raghavan, *Using Part-of-speech Patterns to Reduce Query Ambiguity*. 2007.
- Allen, J., *Natural language understanding* (2nd ed.). 1995: Benjamin-Cummings Publishing Co., Inc. 654.
- Barr, C., R. Jones, and M. Regelson, The linguistic structure of English web-search queries, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2008, Association for Computational Linguistics: Honolulu, Hawaii. p. 1021-1030.
- Chowdhury, A. and M. McCabe, *Improving Information Retrieval Systems using Part of Speech Tagging*. 1998, ISR; TR 1998-48.
- Dinçer, B. and B. Karaoğlu, The Effect of Part-of-Speech Tagging on IR Performance for Turkish, in *Computer and Information Sciences - ISCIS 2004*, C. Aykanat, T. Dayar, and I. Körpeoglu, Editors. 2004, Springer Berlin / Heidelberg. p. 771-778.
- Harksoo Kim, K.k., Jungyun Seo and Gary Geunbae Lee, A Fast and Reliable Question-Answering System Based on Predictive Answer Indexing and Lexico-Syntactic Pattern Matching. *Int. J. Comput. Proc. Oriental Lang*, 2001. 14(4): p. 341-359.

- Kanaan, G., R. al-Shalabi, and M. Sawalha, Improving Arabic Information Retrieval Systems Using Part of Speech Tagging. *Information Technology Journal*, 2005. 4(1): p. 32-37.
- Kao, A. and S.R. Poteet, *Natural Language Processing and Text Mining*. 2006: Springer-Verlag New York, Inc.
- Karimpour, R., et al., Improving Persian information retrieval systems using stemming and part of speech tagging, in *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*. 2009, Springer-Verlag: Aarhus, Denmark. p. 89-96.
- Klavans, J.L. and M.-Y. Kan, The Role of Verbs in Document Analysis, in *36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*. 1998: Montreal, Quebec, Canada.
- Lewis, D.D. and K.S. Jones, Natural language processing for information retrieval. *Commun. ACM*, 1996. 39(1): p. 92-101.
- Manning, C., et al., *Foundations of Statistical Natural Language Processing*. 1999, Cambridge, MA: MIT Press.
- Robertson, S.E., et al., Okapi at TREC-3, in *3rd Text REtrieval Conference (TREC-3)*, D. Harman, Editor. 1995.
- Shah, C. and P. Bhattacharyya, A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR), in *International Conference on Universal Knowledge and Languages (ICUKL)*. 2002.
- Shannon, C.E. and W. Weaver, *A Mathematical Theory of Communication*. 1963: University of Illinois Press.
- Strzalkowski, T., *Natural Language Information Retrieval*. 1999: Kluwer Academic Publishers. 384.
- Strzalkowski, T., *Natural language information retrieval. Information Processing & Management*, 1995. 31(3): p. 397-417.
- Zhai, C., Fast Statistical Parsing of Noun Phrases for Document Indexing, in *5th conference on Applied natural language processing*. 1997: Washington, DC.