USING PART OF SPEECH TAGGING IN PERSIAN INFORMATION RETRIEVAL

Reza Karimpour[1], Amineh Ghorbani[2], Azadeh Pishdad[3]
Mitra Mohtarami[4], Abolfazl AleAhmad[5], Hadi Amiri[6]
Farhad Oroumchian[7]

Abstract

With the emergence of vast resources of information, it is necessary to develop methods that retrieve most relevant information according to the users needs. These retrieval methods may benefit from natural language constructs to boost their results by achieving higher precision/recall rates. In this attempt, we have used part of speech attributes of terms as extra information about document and query terms and have evaluated the impact of such information on the performance of the 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: Persian information retrieval; Natural language; Part of speech

ACM Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

[1] Database Research Group, University of Tehran (rezaka@gmail.com)

[2] Database Research Group, University of Tehran (a.ghorbany@ece.ut.ac.ir)

[3] Database Research Group, University of Tehran (a.pishdad@ece.ut.ac.ir)

[4] Database Research Group, University of Tehran (m.mohtarami@yahoo.com)

[5] Database Research Group, University of Tehran (a.aleahmad@ece.ut.ac.ir)

[6] Database Research Group, University of Tehran (h.amiri@ece.ut.ac.ir)

[7] Database Research Group, University of Tehran (foroumchian@acm.org)

1. Introduction

Exploiting meta-information of the terms in the retrieval process can result in precision and recall improvements. Part of Speech tagging techniques can be used for this purpose by clarifying the role of the terms in the queries and documents and consequently making it possible to assign different priorities to different query terms based on their part of speech. In addition stemming can collapse many surface words in languages like 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 statements and the document collection in the Persian language and to observe the consequences of such incorporation in document retrieval. This general objective is also complemented by the investigating the effect of stemming in such environment.

2. Literature Review

Different algorithms and methods have been developed to make more accurate retrieval engines (Witten, Moffat, & Bell, Nov 1995), (Singhal, Buckley, & Mitra, 1996) (Strohman, Metzler, Turtle, & Croft, 2005). In addition Document retrieval has been an interesting topic for those working in natural language processing (NLP) (Allen, 1995), (LEWIS & Jones, 1996) but not much work has been done on the use of these techniques for Persian document retrieval.

There has been much work on Persian Information Retrieval but none of these approaches have used part of speech tagging which has been tested on other language leading to good results (Amiri, Aleahmad, Oroumchian, Lucas, & Rahgozar, 2007). On the other hand studies in Persian POS tagging have reported accuracy rates of up to 95% using methods such as TnT and MLE taggers (Amiri, Hojjat, & Oroumchian, 2007), (Raja, Amiri, Tasharofi, Sarmadi, Hojjat, & Oroumchian, 2007), (Mohtarami, Amiri, & Oroumchian, 2006), (Ooumchian, Tasharofi, Amiri, Hojjat, & Raja, 2006). Therefore it seemed reasonable to use these taggers in creating a new generation of retrieval engines for Persian language.

In this research we utilize POS tagging methods to assign weight to the terms in documents and queries and reduce the priority of those types of the words that don't have a big impact in retrieval in order to come up with more relevant results.

3. Part of Speech Tagging

Part of speech tagging selects the most likely sequence of syntactic categories for the words in a sentence. It determines 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 models used for various Information retrieval tasks, treat all the content words as the same and without considering their individual significance in the language. Paying attention to the special role of the word in the context tells us a lot about a word and the other words around it. Also it should be noticed that the role of each word in a document is subjective and it depends on what the user means by the words in the query (Allen, 1995), (Shah & Bhattacharyya, 2002)

In different languages and tagging systems, the number of tags vary from a dozen to several hundred depending on the specificity the information provided by the tag. For example a tag set may just categories nouns as singular and plural while another tag set may provide more detail such as *name of a location* or *person*. Obviously, not all of these tags have the same impact on retrieval of relevant documents from a corpus (Carlberger & Kann, 1999). That means calculating a proper tag set with the right size for a particular collection of a language is an issue worthy of studying.

In this study, we used Bijankhan (Bijankhan, 2004) corpus which is a manually tagged document set including 550 different tags. Out of theses tags, a subset of 40 -as shown in Table 1- were selected as the most important ones in practical Persian text based application (Amiri, Hojjat, & Oroumchian, 2007)

TABLE 1 TOP 40 MOST USED TAGS IN BIJANKHAN COLLECTION

Tag names			
PRO	MQUA	ADV_NI	ADJ
PS	MS	ADV_TIME	ADJ_CMPR
QUA	N_PL	AR	ADJ_INO
SPEC	N_SING	CON	ADJ_ORD
V_AUX	NN	DEFAULT	ADJ_SIM
V_IMP	NP	DELM	ADJ_SUP
V_PA	OH	DET	ADV
V_PRE	OHH	IF	ADV_EXM
V_PRS	P	INT	ADV_I
V_SUB	PP	MORP	ADV_NEGG

It has been reported that in some applications of IR, the nouns are more important than other tokens (Turney & Littman, 2002), (Paik, Liddy, Yu, & McKenna, 1993) .However, sometimes even stop words can be useful (Turney & Littman, 2002). 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 applications such as music that are mostly adverb-rich, the role of adverbs become more important (Shah & Bhattacharyya, 2002). Some studies also have investigated the role of verbs in document analysis (Klavans & Kan, 1998).

After analyzing the impact of these 40 different tags, eventually we find out that noun, verb, adjective and adverb are the most important POS Tags in Persian retrieval. In the result section we show the impact of using these tags on documents in the corpus and also queries.

The TnT POS tagger[8] is used in this study to determine the part of speech of Persian words (Brants, 2000). 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 (Brants, 2000).

4. Methodology

The experiments have been conducted using the indri retrieval system. Subsequently the Hamshahri corpus and the queries are tagged using the TNT POS tagger. According to the weighing procedure described in the previous section, queries are weighted using the Indri. The different weighting schemas as well as the omission of less important tags are explained in the next section.

After experimenting with different tagging schemas, the corpus and the queries are then stemmed to evaluate the consequences of stemming on the retrieval system. In addition the tagged corpus will be stemmed too in order to find out how these two approaches may complement each other.

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

Figure 1 shows the framework of our main approach which is the use of stemming on the POS tagged corpus.

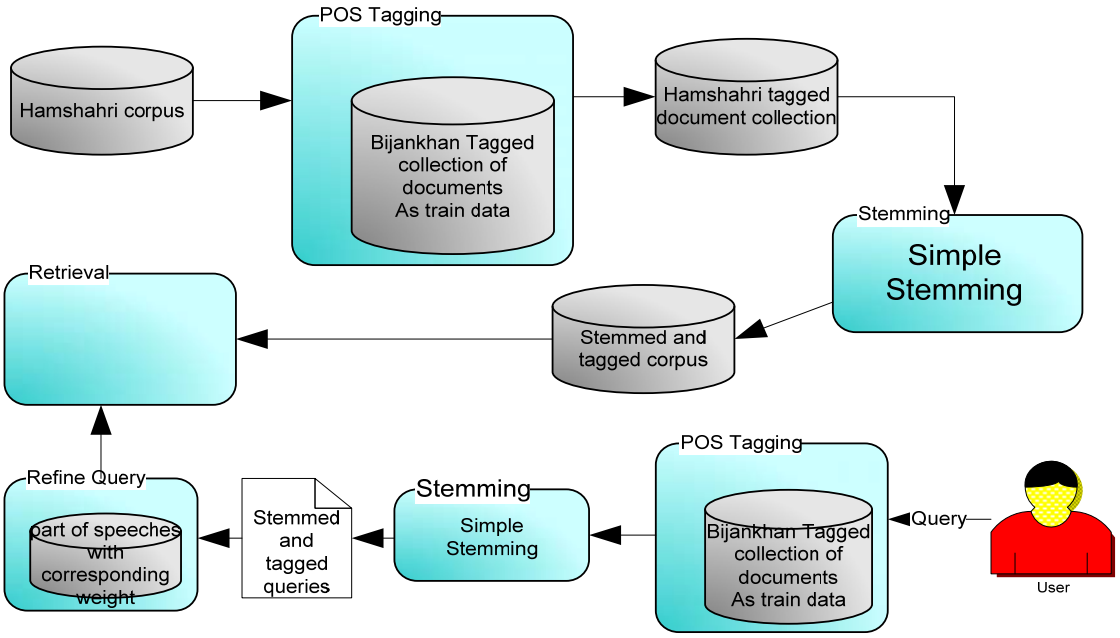

FIGURE 1 FRAMEWORK

5. Implementation

Using Parts of speech tagging metadata information as our main approach, we developed configurations from various combinations of term stemming beside POS tagging in order to observe its effectiveness in action. As stated in previous section tagged corpus was obtained using statistical techniques. Bijankhan tagged collection was used as the training data to estimate the role of each word in the Hamshahri collection. Stemming was performed by employing simple grammatical rules using PERSTEM Persian stemmer[9]. Consequently we prepared 4 different collections as shown in Table 2.

TABLE 2 VARIATIONS OF HAMSHAHRI DOCUMENT COLLECTIONS USED IN RETRIEVALS

Corpus
Normal (Neither stemmed nor tagged)
Stemmed
Terms tagged with related parts of speech
Stemmed and tagged

[9] http://sourceforge.net/projects/perstem

As mentioned earlier the retrieval system used in experiments is the indri system (Strohman, Metzler, Turtle, & Croft, 2005) which is provided as part of Lemur project[10].

In the configurations that include weighting, weights are set based on the POS of the word. For example in a typical configuration all the terms which are tagged as verb may have 0.3 weight compared with 1.0 for nouns. In our first configuration we used the tagged collection along with equal weights for all the terms (i.e. 1.0). Generally queries compromise title, description and narrative. In these experiments the effect of using both title alone and along with description have also been studied. Table 3 depicts the configurations used in these experiments.

TABLE 3 DIFFERENT CONFIGURATIONS

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

6. Results

Before going through the results, it's worth mentioning that since the Hamshahri collection has never been tagged before, we do not have any measurement of the accuracy of the POS tags however basic observations and sampling shows reasonable accuracy.

Table 4 summarizes the outcomes of our experiments, the results of the retrieval system without employing tagging or stemming have an average precision of 27% and R-precision at 36%. With the tagged corpus and title of the queries the average precision climbs to 35% and the R-Precision increases by 1%. This is an interesting result since no weighing has been done yet and there are no priorities but the algorithm searches for terms in documents with the same role as in the queries. On the other hand when the description of the queries are 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 declines to 34% which again is a little more than that of our normal retrieval performance. The reason for this reduction is the negative effect of the extra terms in the query description that misleads the search.

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

TABLE 4 MAIN RESULTS

	normal corpus	tagged (title)	tagged (title and description)	stemmed title	stemmed (title and description)	stemmed(title plus description and stop words)	stemmed and tagged (title)	tags weighting average
Average precision	0.2716	0.3505	0.2989	0.3625	0.1723	0.1672	0.3944	0.2263
R-Precision	0.3627	0.3784	0.3497	0.4102	0.2157	0.2106	0.4151	0.2655

The results we obtained indicate that the 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 by other groups in University of Tehran on the same corpus. Therefore further experiments with different types of stemmers and stemming techniques are required in order to clarify the role of stemming in Persian text processing.

Even though stemming increases the performance of the system, there is a sudden fall when the description is added to the queries, this time even worse than our base line system. The reason for this again goes back to the expansion of the queries by words which might be unnecessary for our results, but this time since many words have turned into a single representation, the effect of misleading terms have increased dramatically. This undesired performance does not get any better even with the omission of stop words that seem to play an important role.

The best result of our experiment was achieved by using the stemming schema on the tagged corpus and title of the queries. This time the average precision reached its peak at 39% and the R-precision stayed at 41% like the stemmed corpus without tags which could mean that the effect of stemming on our retrieval system is much higher than part of speech tagging.

The main focus of our experiments was the weighting of the different terms based on their tags. Many different combinations of weights for different tags have been tried but not only had none of the weighing schemes improved the performance, some of them demonstrated the worse performances. Table 5 depicts the results of some of these experiments. Giving a weight of zero to a tag is the same as omitting that tag from the corpus and the queries. In some experiments as much as 20 least significant tags were omitted from the queries but the precision and recall were reduced. In general the average precision for all the tag weighting schemas was 22% and the average R-Precision was 26% as seen in table 4.

TABLE 5 WEIGHTING SCHEMAS

	20 less used tags omitted, others equal weight	Noun=3, Verb=2, Adjective=1, Adverb=1	Noun=3, Verb=0, Adjective=3, Adverb = 0	Noun=0, Verb=2, Adjective=0, Adverb=0	Noun=0, Verb=0, Adjective=1, Adverb=0	Noun=0, Verb=0, Adjective=0, Adverb=1
Average precision	0.2745	0.2635	0.2597	0.1108	0.1198	0.0977
R-Precision	0.3097	0.3104	0.2888	0.1256	0.1186	0.1111

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 by omitting them the performance of the system declined.

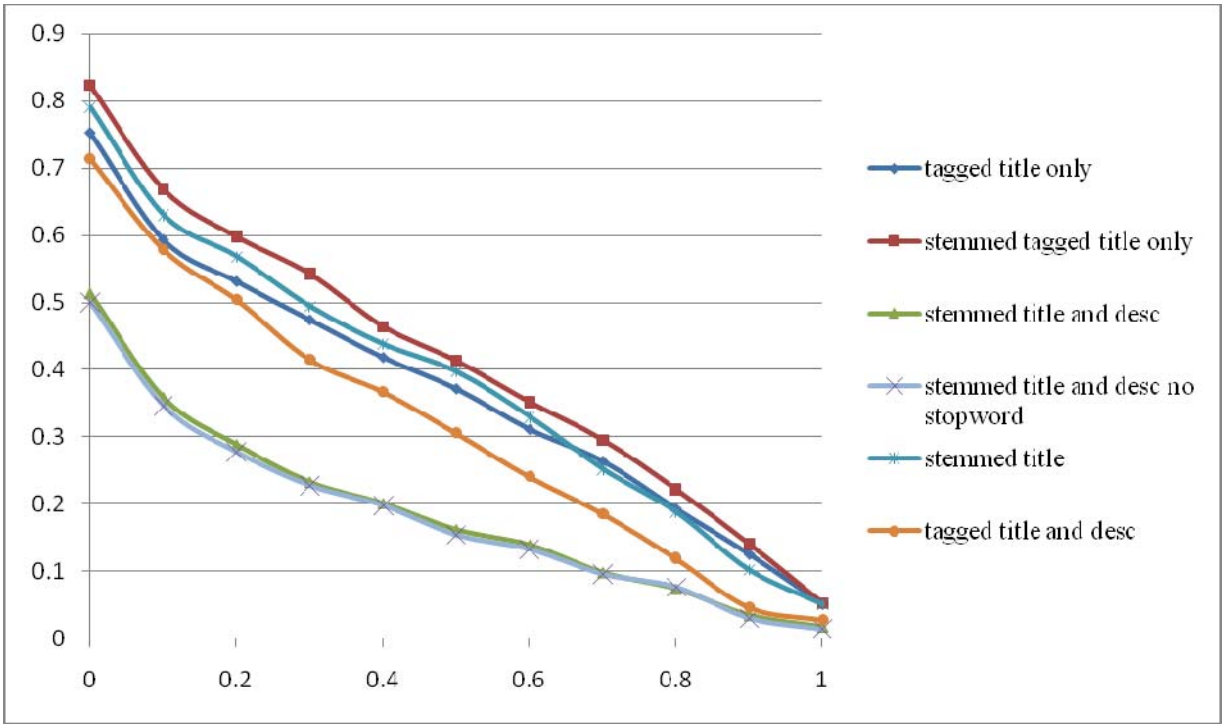

FIGURE 2 R-PRECISION OF THE DIFFERENT CONFIGURATIONS

7. CLEF 2008 Results

All the preceding results were based on training data that is a part of Hamshahri Corpus. In order to further benchmark our system we decided to participate in CELF 2008 monolingual Persian ad hoc track. Our system was named Tehran-NLP and its results were not promising and ranked very low among the other participants. There are some explanations as:

- Our model was developed on the basis of including meta-information about the part of speech of words in information retrieval and since this was our first participation in CLEF 2008, our system was not fine tuned for CLEF.
- Not knowing the rules correctly, we submitted the top 100 retrieved documents instead of top 1000 as all other University of Tehran participants did. This could have an undesirable effect on our runs.

As a result our system was not that successful in CLEF 2008 and we could not make it to the top five.

8. Conclusion and Future Work

This study attempted to measure the effectiveness of part of speech tagging and stemming on Persian information retrieval. Different configurations were used as weighting schemas and the first set of results on training data showed that tagging does improve the performance of the system while weighting based on POS tags reduces it.

The second part of our studies demonstrated that stemming when used along with tagging has a significant effect on the performance.

The result of this study has several implications for future research, first all POS tags are important. Omitting words with specific parts of speech tags in queries or even in documents can have negative effect on Precision and Recall.

We used the simplified tag set with only 40 tags and simplified version of Bijankhan Tagged corpus as our training data. After training we tagged the entire Hamshahri collection automatically. For future research, one might consider a more complex set of part of speech tags. For example, it seems reasonable to assume distinguishing location names from general nouns could be important for query processing and might have a positive effect on effectiveness.

In this research a simple grammatical based algorithm was used to stem Hamshahri collection. Other research groups in DBRG lab of University of Tehran have tried other stemming algorithms and methods and have come up with different results. So, in future, one might want to conduct experiments with all the available stemming methods and algorithms and come up with an explanation for these different behaviors. In addition, other retrieval models should be studied in these experimental configurations in order to ensure that the results obtained are general and not dependent on a particular retrieval model such as Indri.

References

Allen, J. (1995). *Natural Language Understanding, Second Edition.* Benjain/Cummings Publishing Company.

Amiri, H., Aleahmad, A., Oroumchian, F., Lucas, C., & Rahgozar, M. (2007). Using OWA Fuzzy Operator to Merge Retrieval System Results. Computational Approaches to Arabic Script based Languages, CAASL2007 .

Amiri, Hojjat, & Oroumchian. (2007). Investigation on a Feasible Corpus for Persian POS Tagging. *12th International CSI Computer Conference (CSICC) 2007* .

Bijankhan. (2004). The Role of the Corpus in Writing a Grammar: an Introduction to a Software. *Iranian Journal of Linguistics* , vol. 19, no. 2.

Brants, T. (2000). TnT, a statistical part-of-speech tagger. *In Proc. Sixth Conference on Applied Natural Language Processing (ANLP-2000), Seattle, WA* .

Carlberger, J., & Kann, V. (1999). Implementing an efficient part-of-speech tagger. *Software Practice & Experience* , 815-832.

Darrudi, E., Hejazi, M., & Oroumchian, F. (2004). Assessment of a Modern Persian Corpus. *The Second Workshop on Information Technology and its Disciplines,WITID2004* .

Klavans, j., & Kan, M.-y. (1998). Role of Verbs in Document Analysis. *Coling-ACL* , 680–686.

LEWIS, D., & Jones, K. (1996). Natural Language Processing for Information Retrieval. *Communication of the ACM, Volume 39, Issue 1* , 92-101.

Mohtarami, Amiri, & Oroumchian. (2006). Using Heuristic Rules to Improve Persian Part of speech Tagging Accuracy. *the 6th international conference on informatics and systems, INFOS 2006* .

Ooumchian, Tasharofi, Amiri, Hojjat, & Raja. (2006). Creating a Feasible Corpus for Persian POS Tagging. *Rechnical Report, NO. TR3/06, University of Wollongong (Dubai Campus)* .

Paik, W., Liddy, E., Yu, E., & McKenna, M. (1993). Interpretation of proper nouns for information retrieval. *Proceedings of the workshop on Human Language Technology* , 309-313.

Raja, Amiri, Tasharofi, Sarmadi, Hojjat, & Oroumchian. (2007). Evaluation of Part of Speech Tagging on Persian Text. *The Second Workshop on Computational Approaches to Arabic Script-Based Languages, LSA 2007 Linguistic Institute, Stanford University, usa* .

Shah, C., & Bhattacharyya, P. (2002). A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR). *Proceedings of International Conference on Universal Knowledge and Languages (ICUKL) 2002* .

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted Document Length Normalization. *In Proc. of the 19th ACM SIGIR Conference* .

Strohman, T., Metzler, D., Turtle, H., & Croft, W. (2005). Indri: a language-model based search engine for complex queries. *CIIR technical report* .

Tasharofi, S., Raja, F., Oroumchian, F., & Rahgozar, M. (2007). Evaluation of statistical part of speech tagging of persian text. *9th International Symposium on Signal Processing and Its Applications, 2007. ISSPA 2007*, 1-4.

Turney, P. D. (2002). Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data. *National Research Council of Canada* .

Turney, P. D., & Littman, M. L. (2002). Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. *National Research Council of Canada* .

Witten, I., Moffat, A., & Bell, T. (Nov 1995). Managing Gigabytes: Compressing and Indexing Documents and Images. *IEEE Transactions on Information Theory*.