A Persian Part-Of-Speech Tagger Based on Morphological Analysis

Mahdi Mohseni, Behrouz Minaei-bidgoli
Iran University of Science and Technology
E-mail: mohseni@comp.iust.ac.ir, minaeibi@cse.msu.edu

Abstract

This paper describes a method based on morphological analysis of words for a Persian Part-Of-Speech (POS) tagging system. This is a main part of a process for expanding a large Persian corpus called Peyekare (or Textual Corpus of Persian Language). Peykare is arranged into two parts: annotated and unannotated parts. We use the annotated part in order to create an automatic morphological analyzer, a main segment of the system. Morphosyntactic features of Persian words cause two problems: the number of tags is increased in the corpus (586 tags) and the form of the words is changed. This high number of tags debilitates any taggers to work efficiently. From other side the change of word forms reduces the frequency of words with the same lemma; and the number of words belonging to a specific tag reduces as well. This problem also has a bad effect on statistical taggers. The morphological analyzer by removing the problems helps the tagger to cover a large number of tags in the corpus. Using a Markov tagger the method is evaluated on the corpus. The experiments show the efficiency of the method in Persian POS tagging.

1. Introduction

Part-Of-Speech (POS) tagging can be defined as assigning lexical tags to words and symbols constructing a text, in such a way that the tags indicate syntactic roles of words and symbols in a sentence. High percentage of words is often ambiguous in terms of the POS, so the task of POS tagging is offered in order to disambiguate the POSs according to their context. POS taggers and annotated corpora with POS tags are used in many other areas of Natural Language Processing (NLP) such as spell checker, text-to-speech, automatic speech recognition systems and machine translation, among others. Therefore creation of these corpora has been under consideration in different languages from many years ago, contemporary to progress of NLP methods.

So far, many corpora have been developed in other languages and base on them several different models and methods have been applied for POS tagging. The models and methods can be divided into two main approaches: the first one obeys a statistical approach which utilizes annotated corpora and the second one is the rule-based non-statistical approach which is based on machine learning and human knowledge. Some reported methods are as follow: hidden Markov model (Kupiec, 1992; Charniak et al., 1993), maximum entropy system (Ratnaparkhi, 1996), transformation-based tagger (Brill, 1995), memory-based system (Daelemans et al., 1996).

In Persian language there are two well-known corpora: Farsi Linguistic Database (Assi, 1997) and Peykare or Textual Corpus of the Persian Language [1] (In Persian:"پیکره متنی زبان فارسی") (Bijankhan, 2002; Mohseni, 2008). The former is an annotated corpus which is tagged base on a method proposed by (Schuetze, 1995) (see section 2 for more details). The latter is a corpus which is arranged into two parts: annotated and unannotated parts. The annotated part which constitutes about 10% of the corpus is tagged manually. Our goal is to tag the unannotated part (about 90% of the corpus) to have a corpus with 100 million tagged words.

In (Mohseni, 2008) we have done a vast research to discover different problems in Persian POS tagging and to design a comprehensive plan including all aspects of the issue. A main segment of the plan is about morphological analyzing and its effect on tagging. In this paper we offer a method in which a morphological analyzer is designed and used in Persian POS tagging. The results show that the method is completely fruitful in developing a system for tagging the unannotated part of the corpus. Applying the method large number of tags in the corpus is covered and simultaneously the accuracy of tagging still remains high.

The rest of the paper is organized as follows: In section 2 previous works in Persian POS tagging are reviewed. In section 3 describing the Corpus namely Peykare, we explain our method to deal with caused problems by the high number of tags in the tagset. Afterward, Section 4 describes used Markov taggers, trigram tagger which we used for experiments. In section 5 the results of experiments are represented and analyzed in detail. As we said the method is applicable for known words tagging; how we deal with unknown words in the taggers is also expressed in this section. The paper is finished by conclusion and future works in section 6.

2. Literature Review

The first work for Persian POS tagging which is done by (Assi & Abdolhoseini, 2000) is based on a method proposed by (Schuetze, 1995). The system was designed as a part of the annotation procedure for a Persian corpus called the Farsi Linguistic Database (FLDB) (Assi, 1997). The idea is to gather all the neighbors of a word in two vectors called Left Context Vector and Right Context Vector. Words with low frequency are ignored, because it has been observed that rare words will have empty context vectors. Afterwards, the word types are categorized according to their distributional similarity (their similarity in terms of sharing the same neighbors),

[1] http://ece.ut.ac.ir/dbrg/Bijankhan/

and then each category can be manually tagged. Used tagset is made up of 45 tags. Reported accuracy is as follows: accuracy in numbers, different categories of verbs and nouns has been 69-83%, and in general, the accuracy of the automatic part of the system proved to be 57.5%. However, the authors confess that since some of the Persian tags refer to ambiguous words, their offered system is not able to disambiguate POSs of words, as well as tagging less frequent words of texts. Also the accuracy of the system is very low for some categories such as adjectives and adverbs.

Another research for Persian POS tagging is done by (Megerdoomian, 2004). This work does not report any experiments. The author outlines only some of the challenges that arise in the development of a Persian POS tagger, explaining some issues from linguistically viewpoint.

(Raja et al., 2007) presents evaluation of some tagging methods on texts in old version of Peykare (Textual Corpus of the Persian Language). By ignoring many morphosyntactic features of words, the number of tags in the tagset decreases to 40. Also they claim that using some simple heuristics and post-processing the accuracy of used methods is improved. The simple heuristics are actually a few morphological rules to improve the results of unknown words tagging.

In (Mohseni et al., 2008) a POS tagging system based on first order Markov model has applied on old version of Peykare. In the paper some aspects of Persian morphology and some issues in developing a tagging system are offered. The results of the system have been reported in major categories of Persian words.

In (Shamsfard & Fadaee, 2008) an algorithm is presented to tag Persian unknown words. Using 60 inflectional and derivational affixes and a set of 140 rules, they try to analysis words morphologically. The algorithm detects the probable affixes in the word, constructs and prunes the word's parse tree, calculates the truth probability of the remaining derivations and in the last step it assigns the most probable tags to the words. There are some ambiguities in this work. The number of tags and the tagset are not uttered in the paper. Also used corpus and its details are not described. The authors expressed that their algorithm tags words in only 65% cases. This result shows the algorithm is not so effective. According to discussed matters in the paper and the number of used affixes (60) and rules (140) one can find out that the author tries to detect the major categories of words. In Persian a large part of unknown words is classified as noun. For example according to our experience as table 2 shows 57% of words is noun (common noun and proper noun) in Peykare. In other words if we tag all unknown words as noun we have tagged 57% of words correctly. Therefore the accuracy 65% for tagging unknown words is not a remarkable result which is obtained in (Shamsfard & Fadaee, 2008).

In the next section described Peykare, another famous corpus in Persian, we explain our method for Persian POS tagging.

3. Morphological Analyzing and Tagging

3.1 Peykare

Peykare or Textual Corpus of the Persian Language (Bijankhan, 2002; Mohseni, 2008) is a well-known corpus in the Persian language. Peykare is arranged into two parts: annotated and unannotated parts. The annotated part consists of approximately 10 million words (about 10% of the corpus). The texts in this corpus can be divided into formal and colloquial forms. A large part of this corpus contains formal texts got from Persian newspapers, journals and books. Another part of the corpus includes colloquial texts which were selected from Persian story books, interviews and plays. The tagset of the corpus contains 90 single tags of which 16 tags are major categories like noun, adjective, adverb, verb, etc. The structure of words' tags in the corpus is hierarchical base on EAGLES model (Leech & Wilson, 1999). Using this hierarchical structure the tag of words can depict the major category, subtype, inflectional affixes, clitics and other features of words. Here is an example of one tagged word in the corpus:

N, COM,SING,1 کتابم (my book)

First single tag from left (N) represents the major category of the word, second one (COM) is the subtype common for nouns, third one shows that this noun is singular and the last tag is for attached connected pronoun for person 1 namely "م" ("کتاب" + "م" = "کتابم").

Using hierarchical combination of single tags to annotate the words, 586 different tags are obtained in the corpus. This is because of morphosyntactic features of Persian words and the need for hierarchical combinations of tags to represent these features.

3.2 Morphology Analyzing[2] and Tagging

Morphosyntactic features of Persian words cause two problems: the number of tags is increased in the corpus (586 tags) and the form of the words is changed. This high number of tags debilitates any taggers to work efficiently. From other side the change of word forms reduces the frequency of words with the same lemma and the number of words belonging to a specific tag reduces as well. This problem also has a bad effect on statistical taggers. To eliminate these problems, our idea is to analyze words inflectionally before tagging (Figure 1).

In Figure 1 the improvised morphological analyzer works in conjunction with a lexicon. Because both of lexicon and morphological analyzer have a compact effect on each other we encircled them in a dashed box. Analyzed the word to their elements, two above mentioned problems cause by morphosyntactic features of Persian words are solved: the tags is reduced to a manageable number and words with the same lemma are no longer interpreted differently.

[2] In this paper Morphology uses for adding clitics and inflectional morphemes to words. Therefore Morphological Analyzing indicates analyzing word according to inflectional morphemes and clitics.

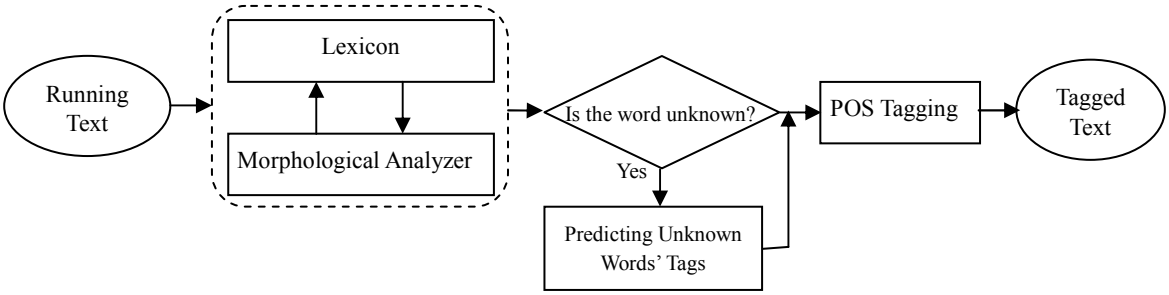

Figure 1: Morphological Analyzer and POS Tagging

But how is a morphological analyzer with a high accuracy created? Due to the fact that there are many inflectional affixes and clitics, Persian is a rich inflectional language in term of its strong morphology. Therefore, methods which try to use morphological rules cannot be successful (like (Shamsfard & Fadaee, 2008) which discussed in section 2). Following we explain our methods to deal with these problems.

Our method utilizes that part of Peykare which is tagged manually in order to create an automatic morphological analyzer. The method includes 6 steps:

Step 1: In the first step we remove those irrelevant single tags that are very rare or indicate semantic concept which are not suitable for POS tagging and in fact many of them are not POS tags. These tags occur more for two main categories, noun and adverb. In the noun category single tags DAY (day), LOC (location), DIR (direction), SES (season), MON (month), SURN (surname) and TIME are removed. In the adverb category the tags TIME, LOC (location), EXM (example), ORD (ordinal), REPT (repetitive) and NEGG (negative) are eliminated. Most of these tags can be added to words simply when tagging is finished. By doing so, the number of distinguished tags in the corpus reduced to 471 tags. This number of tags is still too much for tagging systems.

Step 2: In this step inflectional morphemes and clitics are classified into their representative tags as Table 1 shows. In cases that formal and colloquial forms of morphemes are different from each other colloquial ones are mentioned separately.

Description	Tag	Formal	Colloquial
Connected possessive pronoun (person 1-6)	1	م، ام، یم	_____
	2	ت، ات، یت	_____
	3	ش، اش، یش	_____
	4	مان، امان، یمان	مون، امون، یمون
	5	تان، اتان، یتان	تون، اتون، یتون
	6	شان، اشان، یشان	شون، اشون، یشون
Infinitive marker	YE	ی، ای، یی، ئی	_____
Plural morpheme	PL	ها، ان، یان، جات، گان، ات، یون، ون، ین	ا، ون
Progressive marker (verbs)	PRG	می	_____
Present marker (verbs)	PRES	می	_____
Past participle (verbs)	PAST-P	ه	_____
Copula (verbs for person 1-6)	1	م، ام، یم	_____
	2	ی، ای	_____
	3	ست	ه
	4	یم، ایم، ئیم	_____
	5	ید، اید، یید، ئید	ین، این
	6	ند، اند، یند	ن، ان
Negative marker (verbs)	NEG	ن، م	_____
Subjunctive	SUB	ب	_____
Imperative marker (verbs)	IMP	ب	_____

Table 1: Classified Inflectional Morphemes and Clitics According to Their Tags

Step 3: According to collected morphemes in the previous step all words in the annotated part of the corpus are analyzed inflectionally. In this way the number of different tags in corpus reduced dramatically (from 471 to 105 tags) because any two words with the same lemma are no longer considered as two different words. Therefore, the two problems indicated in the beginning of the section are solved.

Step 4: In this step a lexicon is created. For each word a record is added to the lexicon including different analyses of the word in order of descending frequency. For words with no analysis the record contains only the word. When the lexicon is created, the lexicon can be searched for each word to retrieve its analyses.

Step 5: If now we want to tag a new text, before tagging each word is replaced by the most frequent analysis of the word which is stored in the lexicon. When the word is replaced with its frequent analysis, there are some cases which they were not supposed to be analyzed. By adding a tag to the tagset we allow the tagger to handle these cases. As shown in the experiments we accept a low fault about 5% in analyzing words.

Step 6: In the last step a tagger can run on the analyzed words.

4. Trigram Tagger

According to our experiments (Mohseni, 2008) trigram tagger (second order Markov tagger) is efficient in Persian POS tagging.

If we assume that $\{w^1, w^2, \ldots, w^w\}$ is a set of words in the lexicon and $\{t^1, t^2, \ldots, t^\tau\}$ is a set of possible tags for words, given a sequence of words from the set of words, $w_{1,n}$, the purpose is to find most likely sequence of tag from the set of tags, $t_{1,n}$. Applying Bayes rule, the second order Markov model is defined as follows:

$$\hat{t}_{1,n} = \arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) = \prod_{i=2}^{n+1} [P(w_i \mid t_{i-1} t_i) \times P(t_i \mid t_{i-2} t_{i-1})] \quad (1)$$

Smoothing method is also important when using a Markov model for tagging. We apply no smoothing on output array in the Markov tagger. This is because using some smoothing methods on the output array no improvement achieved in our experiments. The smoothing method which we use for transition array the taggers is inspired from (Thede & Harper, 1999). The smoothing method for trigram tagger is as follows:

$$P(t^k \mid t^i t^j) = k_3 \times \frac{C(t^i, t^j, t^k)}{C(t^i, t^j)} + (1 - k_3) \times k_2 \times \frac{C(t^j, t^k)}{C(t^j)} + (1 - k_3) \times (1 - k_2) \times \frac{C(t^k)}{\sum_{\forall t^m} C(t^m)} \quad (2)$$

In the above relations:

$$k_2 = \frac{\log(C(t^j, t^k) + 1) + 1}{\log(C(t^j, t^k) + 1) + 2} \quad (3)$$

and

$$k_3 = \frac{\log(C(t^i, t^j, t^k) + 1) + 1}{\log(C(t^i, t^j, t^k) + 1) + 2} \quad (4)$$

5. Experimental Studies

In this section using trigram tagger the efficiency of explained method in section 3 is shown. The evaluation method which we used for experiment is a 5-fold cross validation. As we uttered above the aim of this paper is to use the annotated part of Peykare to develop a method based on morphological analyzing for tagging known words.

To tag unknown words the tagger uses estimated probabilities of POS tags for unknown words (Table 2) as we used in (Mohseni et al., 2008).

Tag	Probability
COMmon Noun	39%
PRoper Noun	18%
SIMple ADJective	25%
Verb	2%
RESidual	11%
Others	5%

Table 2: Estimated Probabilities of POS Tags for Unknown Words

Applying the method explained in section 3, in the first step the number of tags from 586 is reduced to 471. This number of tags is very high for any tagger to work efficiently. After analyzing word the number of tags is reduces to 105 tags. 1692775 words out of 8875679 words are analyzed i.e. 1692775 words in the corpus have inflected by prefixing and/or suffixing their lemmas and now are analyzed by morphological analyzer.

If step 5 is applied, 1610985 out of 1692775 words are analyzed correctly. This statistic shows the accuracy of morphological analyzer for inflected words is 95.1%.

In the last step we apply trigram taggers explained in section 4. The result shows that the accuracy of whole system is higher than 90%. In other words covering 471 tags of words in Peykare, the method can tag words by accuracy 90.2%. This is a fantastic result for Persian POS tagging. So far offered results with accuracy higher than 90% as shown in section 2 have been in present of tagsets with maximum 45 tags. For unknown words because the number of tags (105) is still high the accuracy of system cannot exceed 53%.

6. Conclusion

This paper represents the use of morphological analysis in Persian POS tagging system. This system is a main part of a process to expand a Persian corpus called Peykare or Textual Corpus of Persian Language. Used corpus and the tagset are described briefly. Persian morphology changes the forms and the tags of words. This causes some problems for any natural language processing systems like POS taggers. To eliminate these problems, a method based on morphological analysis of words is proposed. To show the efficiency of the method, a trigram tagger is applied on the corpus. The results show that using morphological analyzer the tagging system can cover a large number of different tags in the corpus and simultaneously the accuracy is kept high.

There are many ideas to improve the tagging system. Inspiring described method for tagging known words, a similar method should be developed to tag unknown words more accurately. Investigating errors one can discover those major categories and morphosyntactic features in which errors more occurs than others.

This information can lead to develop post processing methods which minimize manually attempt for modifying automatic tagging results.

7. References

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252--262.

Chercheur, J.L. (1994). *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufman Publishers.

Castor, A., Pollux, L.E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1), pp. 37--53.

Zavatta, A. (1992). Un Générateur d'Insultes s'intégrant dans un Système de Dialogue Humain-Machine. Thèse de Doctorat en Informatique. Université Paris-sud, Centre d'Orsay.

Grandchercheur, L.B. (1983). Vers une modélisation cognitive de l'être et du néant. In S.G Paris, G.M. Olson, & H.W. Stevenson (Eds.), *Fondement des Sciences Cognitives*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 6—38

Assi, S. M. (1997). Farsi Linguistic Database (FLDB). *International Journal of Lexicography*, Vol. 10, No. 3, EURALEX Newsletter p. 5.

Assi, S. M., and Abdolhoseini, M. H. 2000. Grammatical Tagging of a Persian Corpus. *International Journal of Corpus Linguistics,* Volume 5, Number 1, pp. 69-81(13).

Bijankhan, M. (2002). The Persian Language Modeling Plan. Stage Two. Linguistics Lab, Faculty of Literature & Human Science, University of Tehran.

Charniak, E., Hendrickson, C., Hacobson, N. and Perkowitz, M. (1993). Equation for part-of-speech tagging. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp 784-789.

Daelemans W., Zavral, J., Berck, P. and Gillis, S. (1996). MBT: A memory based part of speech tagger-generator. *Proceeding of the Fourth Workshop on Very Large Corpora*, pp 14-27.

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6(3): 225-242.

Leech G. & Wilson, A. (1999). Standards for Tagsets. In Halteren, H. V. *Syntactic wordclass tagging*, (pp: 55-81). Dondrecht: Kluwer academic publishers, The Netherlands.

Megerdoomian, K. (2004). Developing a Persian part-of-speech tagger. *In Proceedings of First Workshop on Persian Language and Computers*. Iran.

Mohseni, M. (2008). Automatic Part-Of-Speech Tagging and Disambiguation System for The Textual Corpus of the Persian Language. MSc. Dissertation. University of Science and Technology: Department of Computer Science. Iran.

Mohseni M., Motallebi H., Minaei-bidgoli B. and Shokrollahi-far M. (2008). A Farsi Part-Of-Speech Tagger Based on Markov Model. *23rd ACM Symposium on Applied Computing*, Brazil.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proceeding of the Conference on Empirical Methods in Natural Language Processing*, pp 133-142.

Raja, F., Amiri, H., Tasharofi, S., Sarmadi, M., Hojjat, H. and Oroumchian, F. (2007). Evaluation of Part of Speech Tagging on Persian Text. *Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages Linguistic Institute*, Stanford, California, USA, pp. 21-22.

Schuetze, H. (1995). Distributional Part-of-Speech Tagging From Texts to Tags: Issues in Multilingual Language Analysis. *Online Proceedings of the ACL SIDGAT Workshop*. On the Internet at http://xxx.lanl.gov/find/cmp-lg.

Thede, S. M., and Harper, M. P. (1999). A Second-Order Hidden Markov Model for Part-Of-Speech Tagging. *ACL1999*.