5th SASTech
Iran | Mashhad
May 12 - 17 | 2011
5th Symposium on Advances in Science & Technology

# A study on part of speech tagging

**N. Jahangiri, M. Kahani, R. Ahamdi, M. Sazvar**
**Professor of linguistics Department, Mashhad, Iran+ jahangiri398@yahoo.com; professor of computer Department, Mashhad, Iran+ kahani@um.ac.ir; M.A in computational linguistics,Mashhad, Iran+ keran_ra@yahoo.com; M.A in computer engineering, Mashhad, Iran+ majid.sazvar@yahoo.com**
1412-436
Razieh

## Abstract

Part of Speech (POS) tagging has high importance in the domain of Natural Language Processing (NLP). POS tagging determines grammatical category to any token, such as noun, verb, adjective, person, gender, etc. Some of the words are ambiguous in their categories and what tagging does is to clear of ambiguous word according to their context. Many taggers are designed with different approaches to reach high accuracy. In this paper we present a new tagging algorithm with a Hybrid algorithm. This algorithm combines the statistics and the rule based tagger to tag Persian unknown words. These algorithms use morphological and syntactical rules for tagging. These algorithms are applied in Gate package.

This package has two parts in tagging; part of tokenization and part of tagging. Many problems depend on part of tokenization. Tokenization is detecting of tokens in a text. In this part, morphological analysis is very important and makes some problems in computational analysis. Persian morphological makes some problems in computational analysis. There is another case which causes some problems in tokenization and is called Persian script.

In this paper, we elaborate some problems in Persian morphology in tokenization and Persian script.

The purpose of this paper is to improve tagging and also to study problems in Gate package in tokenization part according to study of linguistics.

After improving and studying of problems, this package was evaluated with two kinds of texts; standard and non standard texts. Accuracy of Gate package with the standard text and non standard text are 97 and 92%, respectively.

**Key words**: rule based, statistical based, tagging, tokenization, unknown word

## 1. *Introduction*

One of the fundamental works in natural language processing is Part-of-Speech-Tagging. It determines grammatical category to any token, such as noun, verb, adjective, person, gender, etc.

Many taggers are designed with different approaches to reach high accuracy. In this paper we present a new tagging algorithm with a Hybrid algorithm. This algorithm combines the statistics and the rule based tagger to tag Persian unknown words.

These algorithms use morphological and syntactical rules for tagging. These algorithms are applied in Gate package. This package uses for POS tagging for many languages. In this paper, we use this package for the first time in Persian language.

This package has two parts in tagging; part of tokenization and part of tagging. Many problems depend on part of tokenization. Tokenization is detecting of tokens in a text. In this part, morphological analysis is very important and makes some problems in computational analysis.

The purpose of this paper is to improve tagging and also to study problems in Gate package in tokenization part according to study of linguistics for reach high accuracy. After study of problems, this package was evaluated with two kinds of texts; standard and non standard texts.

### *1-1.Literature Review*

In Persian, there has only been an activity in this area in the last few years. One of the first Persian POS taggers is in works of Assi and Abdolhosseyni (2000) that is in turn based on the Schuetze hypothesis. This hypothesis states that syntactic behavior is reflected co-occurrence patterns. They assume that, for a given window size, by storing both the left and the right context vector of each word, clustering the all similar vectors and then manually annotating each cluster, the POS tags can be estimated by observing the cluster to which the new words belong. This system uses tag-set with 45 tags and perform at 57.7 % accuracy.

Amiri et al. (2007) used TNT tagger that has 2.5 million tagged words as training data and the size of tag-set is 38. It has an overall accuracy of 96.64%, specifically, 97.01% on known words and 77.77% on unknown words. The accuracy for known words is much higher than unknown words (about 24%). In comparison with other languages, the accuracy of TNT for Persian, is less but it is close to the accuracy of English and Germany and more accurate than the accuracy of Spanish

Quchani et al. (2008) used Hidden Markov Model. This tagger is a part of Persian TTS system called Pars Gooyan that is implemented in festival TTS software. It is implemented in this environment by Scheme (SIOD) script language and makes use of the Viterbi Decoder by Edinburgh speech Tools. The overall average accuracy for this tagger is 95.11%. The accuracy of known and unknown words is 96.136% and 60.25%, respectively.

Raja et al. (2007) used other tagger also inspired from M.M, and is based on Memory and Maximum likelihood approach. A POS corpus was created for these experiments and taggers were trained on 85% of the corpus and were tested on the remaining 15%. The size of tag-set is 40. In this study simple heuristics that could be applied in post-processing of the output of the tags was experimented. These heuristics were based on modifying a few prefix or suffix characters of the words. Result show that these simple heuristics have significant impact on improving the tagging of the unknown words especially for the weaker models. The overall and unknown word performance of memory based approach with post-processing and the TNT system without post processing are similar to that of other languages such as English, German and Spanish.

Jabbari and Allison (2007) used an implementation of error- driven Transformation based learning. The system learns tagging rules for every coarse part of speech categories and subsequently for a full, complex tag-set.

Shamsfard and Fadaee (2008) used hybrid approach. This algorithm combines the features of probabilistic and rule-based tagger to tag Persian unknown words. The proposed algorithm only deals with the internal structure of the words and pay attention not to the situation of the word among the other words in the same sentence. This algorithm uses morphological rules to tag the words. It applies syntactic rules to reduce the ambiguities of this tagger. In this research there are four steps in tagging an unknown word;

- Detecting the probable affixes in the word.

- Constructing word's parse tree.
- Pruning the parse tree.
- Calculating the truth probability of the remaining derivations.

Megerdoomian (2004) used. This tagger is based on a combination of Symbolic (rule-based) and statistical approaches. These approaches are known as hybrid taggers. This paper present an overview of the main challenges encountered in the developing of a POS tagger for Persian. The paper describes problems arising from encoding issues, detached inflectional morphemes, as well as attached word-like elements forming complex tokens, the discrepancy between orthography and phonetics in application of phonological rules, the interdependency between non-adjacent morphemes in a word, and the recognition of phrasal boundaries. In addition it introduces a certainty to be considered in designing an annotation set for POS tagging.

Aleahmad et al. (2008) created a tagger based on OWA (Ordered Weighted Average) method. This method is a famous method to fuse the final result of the POS tagging systems. In this paper OWA method is used to fuse the result of three different POS tagging system, namely MLE (maximum likelihood estimation), TNT tagger and PTT (Persian Tree Tagger). Results show that although OWA fusion technique has better result than both MLE and PTT system but it cannot over perform TNT system. The results also show that the overall accuracy of the tagger is about 96.59% and the accuracy for known words is much higher than unknown words (about 24%).

The other approach which Bidgoli and Mohseni (2008) used is a bout automatic tagger that based on the full second Markov mode and second Morkov Model with first order output probabilities for Persian language. These two models have used with viterbi decoding for tagging. In both models the tag of each word is based on the two previous tags of the word but the differences between them is based on two hypotheses. In full second order Hidden Morkov Model, tag of word like W is based on tag of previous word and tag of word but in two Morkov Model with first order output probabilities, tag of word like W is based on tag of word. Results show that the accuracy for the unknown word in fuul second order Markov model is 58.2% and for known word is 79.3% and the overall accuracy is 78.3% in the second order Morkov Model with first order output probabilisties. The accuracy for the unknown words is 60.1% and known word is 97.4% and the overall accuracy is 95.7%. The system accuracy for unknown words in first model is less than the second model.

Another study is the tagger based on finite-state Morphological Analysis of Persian. Megerdoomian (2004) said that work describes a two-level morphological analyzer for Persian using a system based on the Xerox finite state tools. The paper presents the problems arising from detached inflectional morphemes, as well as attached word-like elements forming complex tokens, the discrepancy between orthography and phonetics in application of phonological rules, and the interdependency between non-adjacent morphemes in a word.

Mohtarami et al. (2008) used other tagger which is based on Using Heuristic Rules is to improve Persian POS tagging Accuracy. To evaluate the effects of those rules, MLE approach is being selected because of its simplicity and the ease of implementation. Results show that the MLE approach has produced a tagging with the accuracy around 95.29% by using heuristic rules.

There is also a work which Mohseni and Bidgoli (2008) did is based on Morphological analysis in Persian POS tagging system has also been studied. Persian morphology changes the forms and the tags of words. This causes some problems for any natural language processing systems like POS taggers. To show the efficiency of the method, a trigram tagger is applied on the corpus. The results show that using morphological

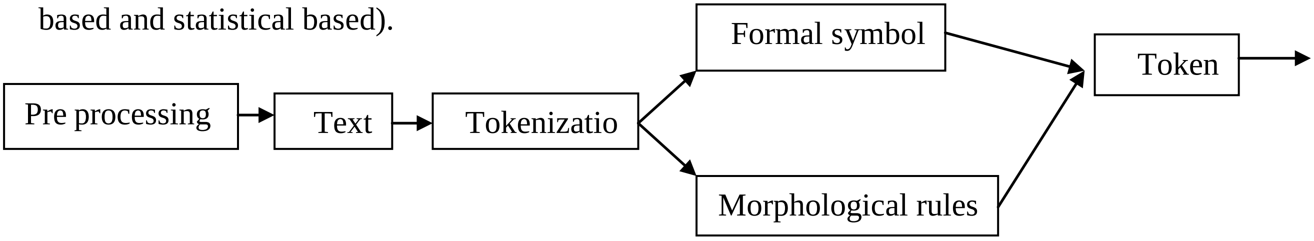
analyzer the tagging system can cover a large number of different tags in the corpus and simultaneously the accuracy is kept higher.
A different view which Azimzadeh and Arab (2007) used is based on the Persian Morphological Parser which is designed based on Finite-State Transducer (FST) is used in some parts of Persian text-to-speech system called Pars Gooyan such as POS of the words without their affixes, and POS of the words in their contexts. These two features are received from a POS tagger that is based on Hidden Markov Models. The accuracy of the final system is 83.51%

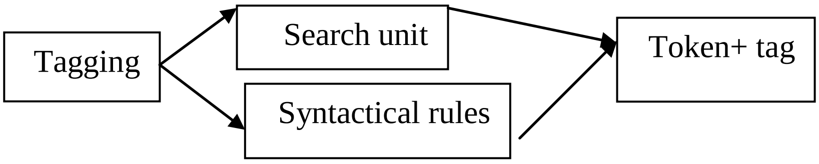
## 2. *Data and material*

The corpus which was used in this work is a part of the Bijankhan's tagged corpus, which is maintained at the linguistics laboratory of the University of Tehran. The corpus is gathered from daily news and common text. It contain 2598216 tokens and tagged with 550 different tags. Most of the tools for POS tagging do not work with a large set of tags. In order to make the tagging process more feasible, we used 40 tags that reduced from Bijankhan's corpus by Oroumchian. This corpus is used for learning rules and tags.
This package in tokenization part has 12000 errors that we use from these errors instead of data and in next section analysis theme. There is another error in tagging part that consist of syntactic errors that we use them for analysis.

## 3. *Research methodology*

In this section we show Gate order in tagging with diagram and use of algorithms (rule based and statistical based).





*3-1. Rule based approach*: this approach use contextual and morphological information to assign tags to unknown and ambiguous words. Some of these rules are written by someone or are learned from corpus. These rules are different for every language and should be underline for tagger designing.
*3-2. Statistical approach*: this approach use probabilistic algorithms and need to be trained on a pre-tagged corpus. Exit of tagging part is token and its tag. There are some tags for each token and tags order significantly for each word and statistical approach select probable tag for word.
*3-3. pre processing*: in Persian language, there are letters that are written some forms in text, like ( ی , ئ , ي)(y) and ( آ , ا , إ , أ )( a) and ( ك , ک )(k). Gate, not recognize words that consist of these letters so we selected from each group one sample, e.g. from ( ی , ئ , ي), (ی) and from ( آ , ا , إ , أ ),(ا ) and from( ك , ک),(ک ).
*3-4. Formal symbol*: in this section for tokenization and separation of tokens, formal symbol like punctuation employ. e. g. space

*3-5. Search unit*: in this section, Gate search words that have wrong tag according to formal symbol and morphological rules.

*3-6. Morphological rules*: morphological rules only deal with the internal structure of the words and syntactic rules are applied for clearing of word according to their context that means the situation of the word among the other words in the sentence.

Morphological rules cause difficulty in tagging in many languages like Persian, English, Urdu, etc. Persian morphology raises some interesting issues for a computational analysis. One of the main challenges of Persian resides in the tokenization of the input text, since word boundaries are not always respected in written text. In this paper, we elaborate on some of the challenges presented by a morphological analysis of Persian and discuss the solutions for improvement of tagging in this language. These challenges are Persian script problems and unknown words. These words consist of multi-token units, derivation words, proper noun, Acronyms, loan words. In this section we first describe the Persian script problems and then describe unknown words. These problems are according to tokenization' errors (12000)

*3-6-1. Persian script and its problems*

The Persian orthography allows some morphemes to appear as bound or free affixes before and after a stem. Many words can be written as concatenated or non- concatenated. In the concatenated form, morpheme is joined to the stem. For instance, in the word كتابها (ketab-ha), suffix ها (ha) is joined to (ketab). In the non- concatenated form, the space or Zero Width Non- joiner character is inserted. For instance كتاب ها and كتاب‌ها.

Typists while typing texts does not insert zero width non- joiner between stem and morpheme. Therefore, a token can be a morpheme, a simple word, a compound and derivation word. Furthermore, the use of the Arabic script and the fact that short vowels are not written and capitalization is not used create ambiguities that impede computational analysis of text.

*3-6-2. unknown words*

Unknown words are words that there are not in the lexicon. Analysis of these words relies on morpho-syntactic information

*3-6-2-1. Multi-token units in Persian*

Bijankhan and Sharifi (2009) says Multi token units are words which made of multi token that should have one POS tag. Between of their tokens should inserted zero width non-joiner but contrary to usual inserted space character. These words are fundamental difficulties in tagging. These words consist of compound preposition, compound conjunction, compound noun, compound adjectives and compound adverbs. In this section we describe these multi-token units.

*Compound prepositions*

Compound prepositionʼs structure in Persian language is divided into two parts. Meshkatodini ( 2008) says First part is P+P such as از پيش (az piše), از روي (az ruye),“from on” . Second part is P+N such as به سبب (be sababe),“because”or به وسيله ( be vasileh),“by”. These second prepositions some times is joined each other, such as بسبب (be sababe), بوسيله ( be vasileh) that means the space character does not insert between P and N and they are tagged as one token but if space character was inserted between them, then they are tagged as two tokens and are tagged as P and N. Furthermore, inserting zero-width non-joiner character and joining them, they are tagged as one token means P.

*Compound conjunction*

Compound conjunction has additional words, such as اگرچه ( agarče), “althought” or درحالي كه ( dar hali ke), “while” or وقتي كه ( vaqti ke), “when”. Some of them such as وقتي كه and در حالي كه is joined such as (وقتيكه) and (در حاليكه). In this form and inserting zero-width non-joiner character, they are tagged one token, means as CON. But inserting space character between them, then they are tagged as two or three tokens such as در حالی که (dar hali ke) is tagged three tokens, means P, N and CON.

*Compound adverb*

Compound adverb usually consist of one preposition and one or two nouns. For instance (به خوبي) (be xubi), “well” or در واقع (dar vaqe), “really” and در نهايت (dar nahayat), “finally”. These with inserting zero-width non-joiner character between P and N are tagged as one token but with inserting space character are tagged two or three token, such as به خوبي is tagged as P and N.

*Compound adjectives*

Compound adjectives’ structure in Persian are different such as1: ADJ+N, e.g, گران قيمت (geran qeymat), “expensive” 2: N+stem, e.g, حقيقت جو(haqiqat ju) , “trust finder” 3: N+ past participle : توزيع كننده ( tozi konandeh), “distributioner” 4: group of adjectives(p+n+past participle)such as ازكارافتاده ( az kar oftadeh),“disabled” and 5 Many adjectives are formed by the bound prefixes such asبی (bi) in بي دقت ( bi deqqat), “careless” 6 number +N, سه نفره ( seh nafareh),“threesomes”. With inserting space character between parts of them, they are tagged separately, which means that in the first one its tag is ADJ+N, in second one is N+N, and in the third one, N+MORP( this tag is in Bijankhan’corpus). The fourth one is P+N+MORP, the fifth one is P+N and in the sixth one is CN+N. Tag of CN is for numbers tags in Bijankhan’ corpus.

*Compound nouns*

Compound nouns in Persian have different structure, such as N+N: آيين نامه ( ayin nameh), “bylaw”, or V_PA+ CON+V_PA: زدو خورد ( zad o xord),“conflict” or V_PA+stem: گفت وگو ( goft o gu), “ talk” and V_IM+ V_IM: بريزبپاش ( beriz bepash), “” and N+ADJCTIVE: چراغ راهنما ( čeraq rahnama),“traffic light” adjective/adverb+V_PA: گرامي داشت ( gerami dašt), “honoring”. These words in tokenization are as two token are tagged according to parts of their formations. For instance the compound word with structure V_PA+CON+V_PA (زد و خورد ) and its tag is V_PA+CON+V_PA but this is one word and one token and its tag should be N. The other words are one token and one word and their tags is N.

*Derivational words*

In Persian, the derived words structure is affix+ stem. These affixes come after and before stem. In Persian there are two affixes; Inflectional affix and the derivational affix. Katamba (2006) says Inflectional affixes serve a syntactic function and derivational affixes create new lexical items. Inflectional affixes are parts of open class words and derivational affixes are parts of closed class words.

Derivational affixes inserting in tagger are classified according to categories that they make. Such affixes make noun, adjectives, and adverb. But some of them are affixes that are Suffixe- like are very important in morphological analysis. These are made of stem of verbs that make noun and adjectives, e.g. ساز(saz) make the adjective,سازي(sazi) make the noun in derived wordsمجسمه ساز and مجسمه سازي(mojasameh saz, mojasameh sazi),“sculptor” and “sculpture” . The number of affixes is very much and should be investigate in morphological analysis. Another problem is common affixes between noun and adjective, e.g. the suffix ي(ye),e.g. خوبي (xubi),“goodness” or between adverb and adjective. e.g.آنه (aneh) for instance شجاعانه (shoja aneh) “bravely”.

*Proper nouns*

Proper noun in Persian is compound and simple. In morphological analysis in Persian, compound nouns such as last name for instance احمد پور (Ahmad pur) are very important. These nouns like other compound nouns with inserting space character act as two tokens but if inserted zero width non-joiner character then known as one token.

*Acronyms*

Acronyms are words that are made from initial letters of phrase. In Persian, there are two kinds. The first kind is made according to Persian. Second kinds are loan acronyms that come from other languages, especially English. One of the examples is BBC. These acronyms in Persian are written in the form of Persian orthography like بي بي سي or initial orthography like BBC. If these acronyms are written in the form of Persian orthography and space character is inserted between them, then they are as three tokens but if it is written in the form of initial orthography, then they are recognized as unknown words. But acronyms which are made according to Persian and are Persian such as ناجا (naja), then they are not creating any difficulties. They are recognized as one token with one tag of N

*Loan words*

Loan words are words which come from other languages specially, from Arabic and English languages to Persian. These Arabic words are Arabic expressions such as Arabic first name and English words are scientific nouns or proper nouns that are recognized as unknown words.

We used some affixes for driven words problem and we make a lexicon for Gate from errors, because a large lexicon is necessary for tagger and reduce unknown words. The tag which system select for unknown words is NN.

In addition to unknown words, there are three unknown tags (JJ, CD and NNS).

Syntactical rules are rules that we can write theme by tags because is simple method. We write 13 rules according to next words, previous words, two next words and two previous words and etc.

1 if tag of word is P then tag of next word is N

2 if tag of previous word of co-ordination conjunction is N, ADJ, ADV,V, PRO then tag of next word of co-ordination conjunction is N, ADJ, ADV, V, PRO

3 if tag of word is QUA then tag of next word is N

4 if tag of word is P_DEFI then tag of next word is N

5 if tag of word is ADJ_SUP then tag of next word is N

6 if tag of words are CON, V sequence then tag of previous word of theme is N

7 if tag of word is DET then tag of next word is N

8 if tag of previous words are QUA, CN then tag of next word is N

9 if tag of previous word is SUR then tag of next word is N

10 if tag of previous word is ADJ then tag of next word is N

11 if tag of two previous words are QUA, P then tag of next word is N

12 if tag of two previous words are QUA, GEN sequence then tag of next word is ADJ

13 if tag of previous word is QUA then tag of next word is ADV

## 4. *Results and Analysis*

Purpose of each tagger is to reach high accuracy. After improving and solving problems, this tagger with two texts was evaluated, standard and non standard text. Standard text selected from Hamshahri newspaper with different topics (sports, religious, scientific, etc). Number of tokens that have tagged manual are 2500 and Accuracy of system is 97%. Non standard texts are with different topics. Number of tokens that have tagged manual are 675 and accuracy of system is 92%. This low accuracy is dependent on unknown words.

| accuracy | Tag set | token |
|---|---|---|
| 97% | 40 | 2500 |

Table1. Standard text

| accuracy | Tag | token |
|---|---|---|

| | set | |
|---|---|---|
| 92% | 40 | 675 |

Table2. Non standard text

## 4. *Conclusions*

Persian script has some problems in tokenization contrary to Arabic and English script. One of the problems contrary to Arabic is short vowel are not written in text and cause ambiguities in tagging, for example the word (**مرد**) (man) in Persian

Text is pronounced to two forms, (mard, mord). There are some words in Persian which are common between affix and preposition, e.g. ( با )(ba), in ( باهوش ) ( ba hush)( clever) is prefix and in ( من با علی رفتم )( I went with Ali) is preposition.

In contrary to English script, in Persian, proper nouns are not written with capitalization, so tokenization of proper noun in Persian cause problems. In tagging and tokenization for Persian the best method is that was created a large lexicon of proper nouns, loan words, Acronyms, etc.

## *References*

Aleahmad, A., & Ramezani,Y., & Oroumchian, F.(2008). Using OWA for Persian Part of Speech Tagging .*Electerical and Computer Engineering Department, University of Tehran*

Amiri, H., & Raja, F., & Sarmadi, M., &Tasharofi, S., & Hojjat, H., & Oroumchian, F .(2007). A survey of part of speech tagging in Persian. *Data base Research Group,* University *of Tehran*

Assi, M., & Hajiabdolhosseyni, M. (2000). Grammatical tagging of a Persian corpus. *International Journal of Corpus Linguistics 5(1), ( 69-81)*

Azimzadeh, A., & Arab, M. (2007). The Persian morphological parser by pos tagger. *The second workshop on computational approaches to Arabic script – based languages , linguistic institute Stanford university,California,USA*

Bidgoli,B.,&Mohseni,M.(2008).7thcongressoflinguistics, Iran.

Jabbari, S., & Ben, A. (2007). Pos tagging for Persian. *The second workshop on computational approaches to Arabic script–based languages , linguistic institute Stanford university,California,USA*

Katamba, F., & Stonham, J. (2006). Morphology. New York, NY: PALGRAVE MACMILLAN

Megerdoomian, K .(2004). Finite-state morphological analysis of Persian. *In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, University of Geneva, Iran*

Megerdoomian, K. (2004). Developing a Persian part-of-speech tagge. *In Proceedings of the First Workshop on Persian Language and Computers. Invited talk. TehranUniversity*

Meshkatodini, M. (2008). dasture zabane farsi: vajegan va peyvanhaye sakhti. Tehran: Samt publisher.

Mohseni, M., & Bidgoli, B. (2008). A Persian pos tagger based morphological analysis. *Iran university of science and technology,LREC conference, 1253-1257*

Mohtarami, M., & Amiri, H., & Oroumchian, F. (2008).Using Heuristic Rules to improve Persian pos tagging Accuracy. *in processing of the 6th international conference on information and systems, faculty of computers and information cairo university,pp-NLP34-NLP38*

Quchani, S., & Azimzadeh, A., & Arab, M.(2008). 13th international conference of Iran computer association

Raja, F., &, Amiri,H., &, Tasharofi, S,. & Hojjat, H.,& Sarmadi, M., & Oroumchian, F.(2007). Evaluation of part of speech tagging on Persian text. *the second workshop on computational approach to Arabic script – baed languages. Linguistic institute Stanford university*

Shamsfard, M., & Fadaee, H. (2008). A Hybrid Morphology – based pos tagger for Persian. *NLP Research Laboratory , Faculty of Electrical & computer engineering shahid beheshti university, Tehran, iran,3453-3460*

Sharifi-Atashgah, M., & Bijankhan, M. (2009).Corpus-based Analysis for Multi-Token Units in Persian. *linguistics Department the Faculty of letters and Humanities, Tehran University*