Comput Syst Sci & Eng (2020) 6: 423–430

International Journal of
Computer Systems
Science & Engineering

An Accurate Persian Part-of-Speech Tagger

Morteza Okhovvat[1,*], **Mohsen Sharifi**[2,†], **Behrouz Minaei Bidgoli**[2,‡]

[1] *Health Management and Social Development Research Center, Golestan University of Medical Sciences, Gorgan, Iran*
[2] *School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran*

The processing of any natural language requires that the grammatical properties of every word in that language are tagged by a part of speech (POS) tagger. To present a more accurate POS tagger for the Persian language, we propose an improved and accurate tagger called IAoM that supports properties of text to speech systems such as Lexical Stress Search, Homograph words Disambiguation, Break Phrase Detection, and main aspects of Persian morphology. IAoM uses Maximum Likelihood Estimation (MLE) to determine the tags of unknown words. In addition, it uses a few defined rules for the sake of achieving high accuracy. For tagging the input corpus, IAoM uses a Hidden Markov Model (HMM) alongside the Viterbi algorithm. To present a fair evaluation, we have performed various experiments on both homogeneous and heterogeneous Persian corpora and studied the effect of the size of training set on the accuracy of IAoM. Experimental results demonstrate the merit of the proposed tagger in achieving an overall accuracy of 97.6%.

Keywords: Hidden Markov Model, Maximum Likelihood Estimation, Morphology, POS Tagger, Viterbi Algorithm

1. INTRODUCTION

Part-of-Speech (POS) tagging is one of the most crucial tasks in Natural Language Processing (NLP) applications and refers to the assignation of lexical tags to the words of a corpus. These tags show the syntactic roles of words in a sentence. Many words are ambiguous with regard to POS. Thus, POS tagging is used to disambiguate the POSs based on their contexts [1], [2].

The POS tags cover various grammatical information such as a person, gender, quantity about a word and its neighbors [3], [4]. Since POS taggers and annotated corpora with POS tags have usually been used in the most areas of NLP such as automatic speech recognition, text-to-speech and spell checker systems, formation of such corpora has been studied in various languages simultaneously with the development of NLP methods.

POS tagging is the process of selecting the correct syntactic tag for a word according to the context or morphological features. In this process, the input data is a text and the output is its words supplied by their POS tags. Taggers are commonly categorized into three classes: statistical-based, rule-based, and transformational-based. Statistical taggers are trained by a labeled corpus and a model is formed that given a word yields the label with the utmost probability. Rule-based taggers contain a wide database of grammatical rules. These taggers make a hypothesis and then the rules are selected from their database. The third class is transformational-based learning (TBL). TBL is a rule-based algorithm for automatic POS tagging. To tag each word, TBL transforms one state to another using transformation rules. It mines linguistic information automatically from corpora. There are also hybrid approaches that apply a combination of the afore mentioned approaches.

POS tagging has various usages in NLP and can be used as a feature in some of the fields such as Duration and Intonation Model [5], Break Phrase Detection [6], [7] and many other fields of linguistic research [8]. Although various tagging methods have been applied in many languages so far, few works on Persian language have been subjected to tagging in recent years. In this paper, we propose a part-of-speech tagging system for the Persian corpus using the Hidden Markov Model (HMM) that is applied to both homogeneous and heterogeneous Persian corpuses. The results demonstrate that IAoM is more accurate than other approaches studied so far.

*Morteza Okhovvat, m-okhovat@goums.ac.ir(Corresponding author)
†Mohsen Sharifi, msharifi@iust.ac.ir
‡Behrouz Minaei Bidgoli, b_minaei@iust.ac.ir

The rest of this paper is organized as follows. In Section 2, notable related works are presented. Main challenges of POS tagging for Persian are presented in Section 3. Our approach is proposed in Section 4. Section 5 presents experimental results, and Section 6 concludes the paper.

2. RELATED WORK

Due to the individual challenging features and constraints of languages, applying existing general-purpose POS tagging methods to all languages is inapplicable [9]. There is however, a comparatively rich set of POS tagging researches for statistical and rule-based languages like English, but there are far fewer works on the Persian language. Jabbari and Allison [10] have presented the most recent work on POS tagging of the Persian language. They use a transformation approach similar to the one used by Brill and Hepple [11] for the in English language. Their tagger includes a trained learner machine that uses approximation rules. They use an Error-Driven Transformation-Based Learning and achieve 93% accuracy. Our objective in this paper is to show that a more accurate tagger for the Persian language is feasible.

Santos and Zadrozny [12] have used a deep neural network to perform POS tagging. Their model learns character-level demonstration of words and associates them with typical word representations. However, a weak point of their model is the introduction of additional hyper-parameters to be tuned.

Assi and Abdolhossini [13] use the Schuetze hypothesis [14] to propose their POS tagging method. They have expressed grammatical tasks that are reverberated in co-occurrence patterns, and have approximated the POS tags for a specified window size. The latter is achieved by sorting context vectors of each word and clustering of all similar vectors. Afterward, each cluster is manually annotated. This method has been utilized to annotate the FLDB corpus [15]. The correctness of various categories of nouns and verbs has been stated to be 69 to 83%, and the total accuracy of the automatic part to be 57.5%. Nevertheless, their proposed method is not applicable to Persian language with loads of ambiguous words.

Brants [16] has proposed a tagger for Persian corpus using the TNT POS tagger that is based on Hidden Markov Model. He uses 2.5 million tagged words as the training data set and considers a tag-set size of 38. His proposed tagger achieves 96.64% accuracy.

Megerdoomian [17] has proposed another POS tagger for Persian corpus. His report only studies some of the linguistical challenges for the development of POS tagging for the Persian language, and does not include any experimental results.

Given the above breif background on POS tagging, we propose IAoM based on HMM that supports properties of text to speech systems. We show that IAoM achieves higher accuracy than the aforementioned approaches in Persian POS tagging.

3. CHALLENGES

Persian language is classified as an Indo-European language with a basic word order of Subject-Object-Verb [18]. According to the different structure of Persian language, there are some challenges that do not exist in other languages like English. Hence, firstly, we explain some aspects of this language and then the challenges of POS tagging are presented.

In Persian, there are fewer tenses than in the English language. Persian has wide derivational and inflectional morphology. Persons inflect verbs and the syntax is not influenced by gender. Like the English language, Derivational Persian words are extracted by prefixing and suffixing their stems [19]. Considering the mentioned features of Persian, the most important POS tagging challenges can be categorized as follows:

I. There are numerous categories of a verb in the Persian language with various inflections in relation to persons leading to variety forms of words.

II. The same forms can mean various morphemes. For example, the suffix "ی" can be considered as a connecting part for the second person e.g., "خوردی" singular or as the indefinite piece of a word e.g., "دستی". This challenge is known as ambiguity in the Persian morphology.

III. In the Persian texts, blanks create serious problems in the process of POS tagging, making it difficult to detect word boundaries. As an example, the plural morpheme "ها" can emerge in various forms for nouns, e.g., the plural form of the word "دستگاه" can emerge in three forms: "دستگاهها", "دستگاه‌ها", and "دستگاه ها".

IV. In Persian, if different affixes such as possessive, indefinite, and plural pronouns emerge in a single word, all of them attempt to join to each other like "کتابهایم" which means "my books". This challenge is attributed to the morphology of the Persian language.

However, proposition of an accurate POS tagger is quite complicated considering the aforementioned challenges though possible as we also show in this paper.

4. CORPUS

One of the famous and well-known corpus in the Persian language is Peykare [20], which is a textual corpus of the persian language. It is organized into two categories: annotated and unannotated parts. The annotated part includes 10 million words that constitute 10% of the corpus. The texts in this corpus can be separated into formal and colloquial forms. Persian newspapers, journalism and books are used to extract formal texts that are a big part of this corpus. Persian storybooks, interviews, and plays are also used to extract colloquial texts, constituting the other part of the corpus. Out of 90 single tags of the corpus, 16 tags are major categories such as a noun, verb, adverb, and adjective. The structure of the words' tags in Peykare is hierarchical based on the EAGLES model [21]. Applying this hierarchical structure, the tagged words can trace the major category, subtype, clitics, inflectional affixes, and other properties of words. Below is an example of a one-tagged word.

N, COM, SING, 1 گلم (my flower) (1)

Table 1 Various forms of suffixes.

Separate	With Half-Space	Connected
می خرم	میخرم	میخرم
گل ها	گل‌ها	گلها

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

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

5. PROPOSED APPROACH

Equation 2 represents the Persian POS tagging method of this paper using the Markov Model and Viterbi algorithm.

$$\widehat{t_{1,n}} = \arg\max p(t_{1,n}|w_{1,n}) \approx \prod_{i=1}^{n} [p(w|t) * p(t|t_i)] \qquad (2)$$

Equation 2 contains two types of probabilities; probabilities of words and probabilities of tag transitions. $w_{1,n}$ represents a sequence of words to find the most likely sequence of tags from the set of possible set of tags, $t_{1,n}$. $\{w^1, w^2, w^3, ..., w^w\}$ is a set of words and $\{t^1, t^2, t^3, ..., t^t\}$ is the set of possible tags for the words. In Equation 2, $P(w_i|t_i)$denotes the probability of a certain tag to a given word. P(ti|ti-1) denotes the tag transition probability indicating the probability of a tag given the previous tag.

As shown in table 1, there are three forms of writing of affixes in Persian, such as connected, with half interspaces, and separated that can well decrease the accuracy of tagging. For example, "می خرم (mikharam) that means "I am buying" and "گل‌ها" (golha) that means "flowers" may be written in three forms, as shown in Table 1.

To overcome the above challenge, the corpus is normalized in our approach as explained in Section 5.1.

5.1 Text Normalization

Detection of word boundaries is one of the significant challenges in the tagging of Persian corpus. As mentioned before, affixes may be written in three forms: connected, with half interspaces, and separated. For example, the word "گلها" (golha) that means "flowers" may emerge in one of the following forms: "گلها" [in connected form], "گل‌ها" [in half-space form], and "گل ها" [in a separate form]. This challenge however may cause some problems. In the Persian orthography, using the half space form is not recommended since it may lead to various difficulties in Token-to-Word transformation. This problem relates to the fact that usually words are diagnosed, with their interspaces, by tokenizer systems. Consequently, affixes that do not appeare in the connected form are recognized as two words and this error affects the results that are obtained in the upcoming steps of NLP. To answer this problem, it is necessary to attach the affixes to their stems. However, this solution is not simply employed for all affixes considering the fact that some affixes are homographs with some words. For example, the word "می" can be pronounced in two forms, "mi" or "mey". If it is pronounced "mi", it should be recognized as an affix for verbs, but if it is pronounced "mey", the meaning is completely changed and it would be a big error since it means wine. To solve this problem it is necessary to consider the context of the words.

In our proposed approach, the Persian letters are firstly changed to English letters that are formerly mined by spaces and other punctuation characters, and in a next stage, affixes are reconnected to their stems. Hence, we may encounter two kinds of affixes to reconnect:

- Dissimilar affixes to stems like "تان" (tan) that signify yours.
- Similar affixes to stems like "تر" (tar) that signify moisture or may be an elaborative suffix.

However, a decision tree is constructed that connects the affixes and words. The primary set of words that are reconnected to the former words is insignificant, but the next set of words that are made to elucidate the context is important. For example, the record of training data for the word "می" in the decision tree is as follows: ((Boolean attach_sign) ("گل" (flower)) ("می") ("خرم" (I buy))).

5.2 Prediction of Unknown Words' Tags

Determinations of the words that were not seen before in the training set, is one of the main challenges in POS tagging. These words are named as "unknown words". We use maximum likelihood estimation (MLE) method to approximate tags of unknown words. To tag each word placed in training set, the more frequent tag compared to other tags is assigned to that word. These tags are called *selected* tags for the words.

MLE-N_SING and MLE-DEFAULT are the two models of MLE methods. MLE-N_SING assigns the "N_SING" tags to unknown words, but MLE-DEFAULT assigns "DEFAULT" tags.

To evaluate the accuracy of MLE and to determine the impact of the size of test sets and training sets on the accuracy of MLE, we run MLE on different percentages of test set and training set of the reduced-tags of Peykare corpus, which was generated by randomly dividing the corpus into test sets and training sets with various distributions. In order to compute the accuracy, the number of times that selected tags are assigned correctly to the words is calculated. Then, the accuracy is calculated by using Equation 3.

Table 2 Accuracy of MLE-DEFAULT.

Run	Percentage of Training Set	Percentage of Test Set	Accuracy of MLE-DEFAULT	
			Known Words	Unknown Words
1	60%	40%	96.45%	0.14%
2	70%	30%	96.51%	0.18%
3	80%	20%	96.60%	0.20%
4	90%	10%	96.81%	0.25%

Table 3 Accuracy of MLE-N_SING.

Run	Percentage of Training Set	Percentage of Test Set	Accuracy of MLE-N_SING	
			Known Words	Unknown Words
1	60%	40%	96.45%	56.49%
2	70%	30%	96.51%	56.53%
3	80%	20%	96.60%	56.62%
4	90%	10%	96.81%	56.67%

Figure 1 The steps carried out by IAoM.

Input: **Running Text**

1. Reduce very rare irrelevant tags and the tags that indicate semantic concept in a unique group
   a. Reduce tags with more than two levels to two-level tags
   b. Reduce two-level tags to one-level tags
2. Normalize the text //*described in 5.1*
3. Determine the boundaries of sentences
4. Tag the words with their most common suffixes and prefixes (shown in Table 3)
5. Analyze all the words of the annotated chunk of the corpus inflectionally
6. Decrease the frequency by determining a record for every word
7. Substitute each word by the most frequent analysis of that word to tag a new text
8. Run POS tagging

Output: **Tagged text**

Tables 2 and 3 demonstrate the accuracy for MLE-DEFAULT and MLE-N_SING, respectively.

$$Accuracy = (Number\ of\ correctly\ assigned\ tags)/Number\ of\ words \tag{3}$$

As it is shown in Tables 2 and 3, we have the same accuracy for both MLE models since these models act differently just with unknown words. However, the accuracy of MLE is not acceptable for unknown words since it has 0.19% accuracy under MLE-DEFAULT and 56% under MLE-N_SING. In order to improve the accuracy of MLE for unknown words, we propose the POS tagger called IAoM as an Improved and Accurate Tagger using the Maximum Likelihood. Figure 1 shows the steps that are followed by IAoM.

IAoM, in the first step, removes the unrelated single tags. These tags are very rare implying that the semantic concept is not appropriate for POS tagging. These tags usually arise for nouns and adverbs. Hence, in the noun category, the tags of DIR (direction), DAY (day), MON (month), SES (season), TIME, SURN (surname), and LOC (location) are removed. In the adverb category, the single tags such as EXM (example), LOC (location), NEGG (negative), REPT (repetitive), ORD (ordinal), and TIME are removed. In this step, the tags with three or more levels are changed to two- level tags. For example, "N_PL_DAY" and "N_PL_LOC" are considered as three-level tags. These are two sample tags of plural nouns and related to time and location, respectively. After the first step of IAoM, these two tags are changed to new two-level tags named "N_PL" as a two-level tag. On the other hand, some of the rare and unnecessary tags such as various tags of adverbs, prepositions, and conjunctions are reduced to single-level tags. Also, the five non-repeated tags that are "N", "V", "V_SNFL", "MORP" and "NP_INYA", are eliminated.

Because in the Persian corpus, affixes can be written in three forms of connected, separate and half interspaces, and the training data contains incorrectly tagged words, the POS lexicon includes noise. Hence, IAoM normalizes the text in the second step as described in Section 5.1. Since the input of IAoM is a text and the tagger is the work on sentences and the boundary of sentences that are not verified in the corpus, IAoM uses some rules to determine the sentences boundary. Relation (4) shows the rules extracted by examination of the configuration of Persian text:

$$\text{Verb} + (\text{conjunction, preposition '.', '?', ',', ':' or ','}) \tag{4}$$

The POS lexicon contains noise because the training data includes incorrectly tagged words. These can introduce errors in the decision-making process of the system. Therefore, to reduce the error rate, a linguistic specialist has manually corrected the lexicon. Another source of error is abnormal trigram values. Some trigrams have very large values that cause errors in the candidate-choice process. These trigrams always get high heuristic scores in the Viterbi decoder. We have clipped all trigram values above the threshold to the threshold value.

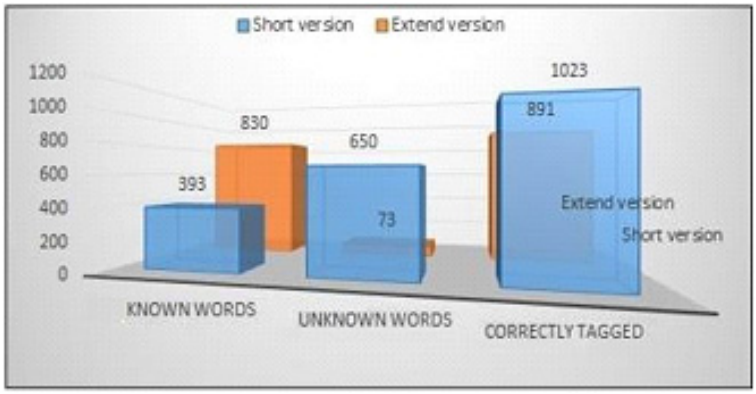

Figure 2 Average results of tagging.

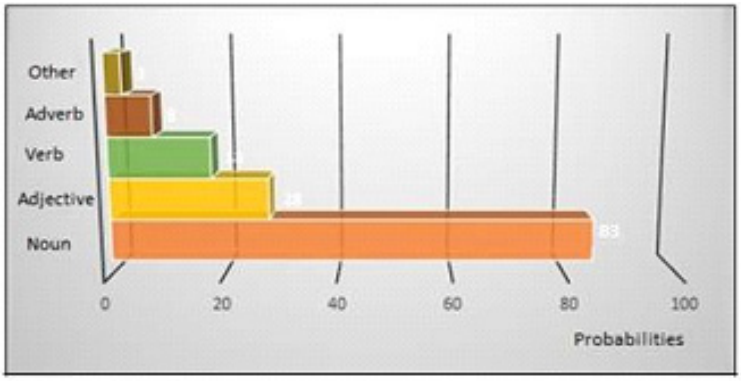

Figure 3 Probabilities of POS tags for unknown words.

As mentioned in the step 4 of IAoM, words with their most common suffixes and prefixes are tagged using rules showed in Table 3 that is based on MLE-N_SING. As shown in Tables 2 and 3, MLE-N_SING does better in the assignment of tags to unknown words compared with MLE-DEFAULT. In order to improve the accuracy of MLE-N_SING, we did rather a deep study. We figured out that some of the unknown words that were plural nouns ("N_PL"), are tagged wrongly.

In the Persian language most of the plural nouns finish with suffixes like"ها", "ان","ات", etc. For instance, میز (miz) which means "table" is a singular noun and its plural form ("N_PL") is "میز ها" (mizha) which means "tables". Therefore, the accuracy can be improved in such a way that an unknown word ends with any of the plural suffixes is tagged as "N_PL". Table 3 shows a classification of the features of unknown words. In cases that formal and colloquial forms of morphemes are dissimilar, colloquial ones are specified definitely.

At the end of step 6, the frequency is decreased. In fact, in step 5, a record is added to the lexicon containing different analyses. In the case of having no analysis, the record holds simply the word. After creation of the lexicons and recovering their analysis, they can be searched for related words. According to step 7, if we are going to tag a new text, every word is substituted by the most frequent analysis of the words saved in the lexicon. Since we may face cases wherein some words are not supposed to be analyzed, some tags are added to the tag-set. Finally, our POS tagger based on hidden markov is run. After performing IAoM, the number of different tags in corpus is dramatically reduced from 882 to 37 tags.

6. EXPERIMENTAL RESULTS

The efficiency of IAoM is reported in this section. We applied IAoM to two different Persian corpora. First, we use the Peykare [20] and the Hamshahri corpora as training and test sets, respectively. The Hamshahri corpus has been extracted from published papers in the Hamshahri newspapers during 2001 to 2005. We did our experiments using two versions of the corpus including 11400 words named as short version and 2487900 words named as extendeded version. We repeated the experiments on both short and extended versions of the corpus for three times. Tables 4 and 5 show the results of experiments and the average results of the experiments are shown in Figure 2.

To complement our experimental evaluation, we deployed IAoM on two separate parts of the same corpus developed by the Research Center for Intelligent Signal Processing, RCISP [22]. This corpus contains some texts with different subjects such as art, sport, economics, social, culture, and religious. The annotated part of the corpus that is used in our experiments has nearly 7.5 million annotated tokens containing 10 million words. We used 25 tags in the experiments that are the key tags from 168 single tags of the corpus. The remaining tags are 143 single tags and used for Persian morphology.

As mentioned before, unknown words lead to some problems in POS tagging as well as many NLP systems. Therefore, the possibilities of various POS tags for unknown words are determined by applying 5-fold cross validation (Figure 3).

In order to assess our proposed approach, we firstly applied IAoM to the Economic section of the corpus as homogenous

Table 4 Classified Features of Unknown Words.

Real Tag of the Unknown Word	Unknown Word's Formal Morphemes	Unknown Word's Colloquial Morphemes	Suffix/Prefix
N_PL (Plural Noun)	امان، یمون، مان، تان، اتان، یتان، ها، هاي، هایي، ان، هایم، هایت، هایش، هایمان، هایتان، هایشان، ین، ات، ان، یان، یون، ون، ویان، وجات، یجات، هجات، اجات، جاتي، جات، گان	مون، تون، اتون، یتون، ا، ون، شون، اشون، یشون	Suffix
V_PRE (Attributive Verb)	ست	—	Suffix
V_PRS (Present Verb)	می، نمی	—	Prefix
ADJ_SUP (Superlative Adjective)	ترین	—	Suffix
ADJ_CMPR (Comparative Adjective)	تری، تر	—	Suffix
CON (Conjunction)	با وجود اینکه، مادامیکه، در هر حال، گو اینکه، کما اینکه، با وجود اینکه، گذشته از اینکه، پیش از اینکه، بالتبع، از آن رو، به طوریکه، به صورتي که، بدین قرار، از طرف دیگر، شگفت اینکه، صرف نظر از اینکه، به این معنا که، به بیان دیگر، به عبارت دیگر، هر چه که، معذالک، به همین علت، از این رو، هر چه که، به عبارتي	با وجود اونکه، از اون رو، هر چي که،	Word
V_PA (Past Verb)	اي، ام، اند، ید، ند، یم، یند	ین، این، ان، ن	Suffix
V_SUB (Implicit Verb)	ب، ن، م	—	Prefix

Table 5 Experimental results of tagging (short version of corpus).

Experiment No.	Tagged Words	Known Words	Unknown Words	Correctly Tagged
1	91	34	57	88
2	520	214	306	514
3	2512	930	1582	2461

Table 6 Experimental results of tagging (extended version of corpus).

Experiment No.	Tagged Words	Known Words	Unknown Words	Correctly Tagged
1	91	66	25	70
2	520	373	147	399
3	2512	2049	463	2203

text. Results show the accuracy of our proposed approach for about 97.1% shown in Table 6.

Secondly, we carried out our experiments on a part of the corpus selected from five different genres of texts. The experimental results are shown in Table 7 in terms of accuracy.

Although the training data in the second experiment is heterogeneous, more accuracy is achived in tagging known

Table 7 Results of the first experiment.

Type	Number of Words	Accuracy
Known words	746198	97.5%
Unknown words	14582	69.9%
Total	760780	97.1%

Table 8 Results of the second experiment.

Type	Number of Words	Accuracy
Known words	1052362	98.2%
Unknown words	21562	65%
Total	1073924	97.6%

words because of the larger training data in the second experiments. Since the training data in the first experiment is homogeneous and selected from a single type, more accuracy was gained for unknown words. Our experimental results reveal the superiority of IAoM in Persian POS tagging.

7. CONCLUSION AND FUTURE WORKS

Part-of-Speech (POS) tagging is one of the important but challenging tasks in NLP applications. Because of the different structure of the Persian language, there are certain challenges that do not exist in some other languages like English. In this paper, we study the challenges of Persian POS tagging systems, and propose an accurate Persian POS tagger, named IAoM, based on the combination of MLE method, HMM, and Viterbi algorithm to tag Persian corpus. IAoM supports properties of text to speech systems such as Lexical Stress Search, Homograph words Disambiguation, Break Phrase Detection, and main aspects of Persian morphology. To improve the accuracy of tagging unknown words, we constrain the MLE by using some defined rules. We perform various experiments to present a fair evaluation of IAoM. To evaluate the performance of IAoM, we use Hamshahri corpus as the test set and apply IAoM to Peykare and RCISP corpuses. Experimental results demonstrate slight deviation in the accuracy rate of IAoM in both homogenous and heterogeneous corpora. The high accuracy (approximately, 97.6%) of IAoM indicates its merit in tagging of Persian corpora.

In future, we would like to extend our study in more depth to Persian Text-to-Speech (TTS) systems and try to achieve more accurate results of Persian NLP by using a multi- approach system.

REFERENCES

1. Zeroual, I., Lakhouaja, A., Belahbib, R., (2017). Towards a standard Part of Speech tagset for the Arabic language. Journal of. King Saudi University – Computer and Information Science, 29 (2): 171–178.
2. Bach, N.X., Linh, N.D., Phuong, T.M., (2018). An empirical study on POS tagging for Vietnamese social media text. Computer Speech & Language, 50: 1–15.
3. Chen, W., Zhang, M., Zhang, Y., Duan, X., (2016). Exploiting meta features for dependency parsing and part-of-speech tagging. Artificial Intelligence, 230, 173–191.
4. Elahimanesh, M.H., Minaei Bidgoli, B., Kermani, F., (2014). ACUT: An Associative Classifier Approach to Unknown Word POS Tagging. In International Symposium on Artificial Intelligence and Signal Processing.
5. Antony J, B., Mahalakshmi, G.S., (2015). Content-based Information Retrieval by Named Entity Recognition and Verb Semantic Role Labelling, Journal of Universal Computer Science, 21 (13): 1830–1848.
6. Sun, S., Luo, C., Chen, J., (2017). A review of natural language processing techniques for opinion mining systems. Information Fusion, 36: 10–25.
7. Pakzad, A., Minaei Bidgoli, B., (2016). An improved joint model: POS tagging and dependency parsing. Journal of AI and Data Mining, 4 (1): 1–8.
8. Bellegarda, J. R., Monz, C., (2016). State of the art in statistical methods for language and speech processing. Computer Speech & Language, 35, 163–184.
9. Okhovvat, M., Minaei Bidgoli, B., (2011). A Hidden Markov Model for Persian Part-Of-Speech Tagging. Procedia Computer Science, 3: 977–981.
10. Jabbari, S., Allison, B., (2007). Persian Part of Speech Tagging. In Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages (CAASL-2).
11. Hepple, M., (2000). Independence and commitment: Assumptions for rapid training and execution of rule-based pos taggers. In Proceedings of the 38th Annual Meeting of the ACL.
12. Santos, C. D., Zadrozny, B., 2014. Learning character-level representations for part-of-speech Tagging. In Proceedings of the 31st International Conference on Machine Learning (ICML'14).
13. Assi, S. M., (1997). Farsi Linguistic Database (FLDB), International Journal of Lexicon, 10 (3): 1–9.
14. Steube, A., (2008). The Discourse Potential of Underspecified Structures. Walter de Gruyter, Berlin, Germany.
15. Schuetze, H., (1995). Distributional Part-of-Speech Tagging From Texts to Tags: Issues in Multilingual Language Analysis. In Proceedings of the ACL SIDGAT Workshop.
16. Brants, T., (2000). TNT – a Statistical Part-of-Speech Tagger. In Proceedings of 6th conference on applied natural language processing (ANLP).
17. Megerdoomian, K., (2004). Developing a Persian part-of speech tagger. In Proceedings of first Workshop on Persian Language and computer.
18. Mousavian, A., Ebrahimzadeh, M. H., Birjandinejad, A., Omidi-Kashani, F., Kachooei, A.R., (2015). Translation and cultural adaptation of the Manchester-Oxford Foot Questionnaire (MOXFQ) into Persian language. The Foot, 25 (4): 224–227.

19. Carneiro, C.C. H., M.G. França, F., Lima, M.V. P., (2015). Multilingual part-of-speech tagging with weightless neural networks, Neural Networks, 66: 11–21.
20. BijanKhan, M., (2004). The Role of the Corpus in Writing a Grammar: An Introduction to a Software. Iranian Journal of Linguistics, 19 (2).
21. Leech, G., Wilson, A., (1999). Standards for Tag-sets. Syntactic Wordclass Tagging, Springer, 55–80.
22. Research Center for Intelligent Signal Processing, available at http:// http://www.rcisp.com (visited on 6/12/2018)