International Journal of Engineering

Journal Homepage: www.ije.ir

A Hidden Markov Model for Morphology of Compound Roles in Persian Text Part of Tagging

H. Rezaei[a], H. Motameni*[a], B. Barzegar[b]

[a] *Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran*
[b] *Department of Computer Engineering, Babol Branch, Islamic Azad University, Babol, Iran*

PAPER INFO

Paper history:
Received 16 April 2021
Received in revised form 30 July 2021
Accepted 23 August 2021

Keywords:
Hidden Markov Model
Morphology
Independent Roles
Dependent Roles
Tri Gram

ABSTRACT

Nowadays, data mining has become significant given the popularity of social networks as well as the emergence of abbreviated words, foreign terms and emoticons in Persian language. Meanwhile, numerous studies have been conducted to identify the type of words. Identifying the role of each word in a sentence is far more important than identifying the type of word in the sentence. Meanwhile, the spelling-grammatical similarity of Persian to Arabic has enabled the newly proposed method in this paper to be applied to Arabic. In this paper, we adopted the Hidden Markov Model (HMM) and Tri-gram tagging with the aim of identifying the morphology of composition roles in Persian sentences. Then, a comparison was made between the technique developed in this paper and the Hidden Markov Model, Uni-gram and Bi-gram tagging. The proposed method supports the results obtained by the word role identification through "independent" and "dependent" roles and several factors that have a contribution to the words roles in sentences. In fact, the simulation results show that the average success rates of independent composition roles with HMM and Tri-gram tagging were 20.56% and 17.67% compared to Uni-gram and Bi-gram methods, respectively. Regarding the dependent composition role, there were improvements by 24.67% and 32.62%, respectively.

doi: *10.5829/ije.2021.34.11b.12*

1. INTRODUCTION

Nowadays, the expansion and popularity of social networks among the public has led the divergence of writing and speech styles in every language, turning phrases into a rather abridged, colloquial form. For that reason, it becomes increasingly crucial to conduct research into linguistics essential to numerous functions such as machine translation [1], smart filtering [2] speech recognition [3, 4], text processing and summarization. Persian is a language officially spoken in many countries including Iran, Afghanistan, Tajikistan and unofficially in certain regions of Uzbekistan. Furthermore, Persian is akin to Arabic in terms of alphabet, and in some extent grammar. This in turn explains the significant contribution of Persian to morphology of other languages. On the other hand, Persian language experts have more frequently investigated the processing of word types in sentences. Nevertheless, the word roles in machine translation or speech recognition tend to be more effective than word types for achievement of better results. Widely applied by many scholars in language processing, Hidden Markov Model (HMM) [5] with N-gram tagging is a key statistical measure in identification of words [6]. Due to insufficient investigation into the semantic aspect of words in sentences (i.e. compounds), there has been a significant gap in morphology through HMM with a focus on the scope of word roles in a Persian sentence. Raising the success rate in sentence role identification in any language will help to achieve better results in practical areas. The morphology of compound roles in Persian sentences is carried out through HMM with Tri-gram tagging given 1) the importance of identifying word roles in a sentence based on the word type, 2) the significant dispersion of Persian language across several countries, and 3) expansion of Persian usage [7-9].

To solve the above problems, this paper makes three

* Corresponding Author Institutional Email: motameni@iausari.ac.ir (H. Motameni)

Please cite this article as: H. Rezaei, H. Motameni, B. Barzegar, A Hidden Markov Model for Morphology of Compound Roles in Persian Text Part of Tagging, International Journal of Engineering, Transactions B: Applications, Vol. 34, No. 11, (2021) 2494-2507

contributions:

- The output of this research can be used in all data mining projects mentioned above.
- The method of this research can be used in all Persian-speaking countries and in Arabic-speaking countries.
- Morphology using fuzzy Hidden Markov Model Within the role of words in Persian sentences.

The remainder of this paper is organized as follows. Section 2 introduces background and literature review. Section 3 introduces the problem description. In Section 4, the proposed algorithm is described and its performance is evaluated. Section 5 gives the performance study with simulation. Finally, Section 6 concludes the paper.

2. LITERATURE REVIEW

This investigates the fundamental concepts in lingusitics and morphology. Furthermore, the related prelimineries are discussed in detail.

2. 1. Linguistics Scientists belive that language is a generic term interacting with wide array of issues including social factor, regional class, language acquisition, neurolinguistics, application of linguistic knowledge to the forensic context of law which is known as forensic linguistic, analysis of language disabilities, and language usage. Literature also shows that computer sicence and analytical techniques have great impacts on linguistics along with the aforementioned concept in recent years.

Linguistics includes the fields of reflecting the various dimensions of linguistics [12, 13]. Generally speaking, language is the study of language and its structure, and a linguist should be expert in one of the specific branches in linguisitc including dialectology, computational linguisitcs, applied linguistics, and etc [14]. Although lingusitcs is a legacy field of studying developed by Indian Panini during fifth century, the modern linguisitcs dates from the beginning of 20th century. The famous linguists in the beginning era of linguistics are Ferdinand de Saussure and Noam Chomsky. Ferdinand de Saussure trusted in theory of structuralism in linguistics while Chomsky identified that sentence as the unit of study in linguistics. The theories of Noam Chomsky have still strong followers and fans in North America today. In contrast, there exists a method called discourse approach that do not recognize sentence as the unit of study in linguistics.

Three semantic levels are identified for every text basically which examine three distinguished concepts namely, content, interaction within equation of this content, the level of sentence impact in formulation of the content. A discourse approach does not restrict the text to have pre-specified length as it even recognizes a word as a text [15]. The Iranian Linguistics Foundation proposed the initial Persian grammer and later advanced by Bateni through introducing new structures within Persian grammar. One of the main achievments of Bateni was formulating the conversion between different types of sentence. According to Meghdari et al. computational linguistic is an interdisciplinary field relating to two branch of knowledge namely, computer science and linguistics in which the natural lanuage is evaluated using statistical and machine learning techniques [16, 17]. This method includes grammar for making sentences and making words. One of the researchers on using Grammar in Natural Language Processing (NLP) in non-Persian-Arabic languages is Chomsky. He was a pioneer in this research.The aim of computational linguistic is to implement and build artifacts which improve the relationship between computer and the language. The initial applicability of computational linguistics was in machine translation using machine learning and data analytical techniques. The first journal publishes the machine translation related issues was Mechanical Translation in 1954 (renamed later as computational linguistics) prior to establish the Association for Computational Linguistics in 1962 [18]. The appliability of computational linguistics is not restricted to machine translation and is used wide range of IT-related issues [19, 20]. Literature also reveals many researches regarding Persian computational linguistics [2, 21-23].

With the establishment of a web technology research laboratory in recent years at Ferdowsi University of Mashhad [24] and the Linguistics Research Institute at Sharif University, extensive research has been done in this regard [16, 19]. They show the concentration and direction of computational linguistics research in Iran.

Fuzzy intelligent systems deal with fuzzy rules to describe vague, inaccurate concepts. In recent research, fuzzy theory has been used in Persian linguistics as well as its combination with Arabic [26]. In another study [6, 25, 34], the fuzzy method was adopted to identify composition roles in Persian sentences. Considering all advances in fuzzy data mining it seems that fuzzy morphology [34], especially in Persian and Arabic, and in particular the combination of words in the sentence, which deals with the meaning of words in sentences, has been paid less attention by the researchers.

2. 2. Morphology The literature offers various definitions for the term ‘Morphology’. Among them, the most commonly-referred definition indicats that morphology is the study of words and morphemes so that morphemes refer to the stem of words. Morphology is kind of complex since different languages involve different phonemes, alphabet, grammar, and speech. For example, English morphology meaningfully differs from Arabic and Persian morphology [27-29]. Buckwalter [33] implemented morphological analysis of Arabic and applied rule-based method to improve root search and clarification of Arabic words, and performed system evaluating root search of Arabic by that grammar and the relevant rules. Arab and Azimazadeh [35] used HMMs to predict the tags of unknown words. Tri-grams were used in their model and they tried to solve the ambiguity problem. Okhovvat and Minaiee [36] applied HMMs for Part-of-Speech (POS) tagging. They trained a model by both homogenous and heterogeneous corpora. They determined the sentence boundaries to make the model more precise. Another study was designing a dependency parser for Persian language and discovering the linguistic dependencies to ease NLP tasks [37]. Kardan and Imani [38] used maximum entropy as a classifier for POS tagging. They chose those types of features that can show the most important characteristics of a word. Pakzad and Minaee [39] also used dependency grammar and joint probability for Persian and English annotation. Table 1 compares the most important studies in the field of Persian morphology.

3. PROBLEM DESCRIPTION

In this section, we explore the HMM, N-gram tagging (particularly Tri-gram) and sentence roles in Persian prior to presenting our new method.

TABLE 1. Comparison of morphological research

Ref.	Methodology	Advantages	Disadvantages
Bijankhan el al. [21]	Eagles standard	One of the basic Persian morphological corpora	Tagging words and constructing corpora merely based on word types, nonstandard corpora textual documents, eagles-based methodology
Assi and Abdolhosseini [23]	Manual	One of the basic Persian colloquial morphological corpora	Tagging words and constructing corpora merely based on word types, nonstandard corpora textual documents, manual methodology
Motameni and Peykar [8]	Fuzzy HMM	Using fuzzy system to determine word roles	Tagging standard sentences, high computational complexity
Motameni et al. [25]	Classified fuzzy	Low computational complexity, determining word roles, using fuzzy system	Only tagging standard sentences
Motameni [34]	Deep Learning Fuzzy Neural Network	memory of the fuzzy GRU method allows to select irregular values relative to the input states	Tagging standard sentences, high computational complexity

3. 1. Hidden Markov Model In HMM, observations are probabilistic functions of states. The output is a stochastic model involving an underlying random hidden process observable only to a set of random processes that generate the sequence of observations [32]. As can be seen in this scenario, the observation time (t) is defined by variable Y_t. This model has been demonstrated in Equation (1), where S_1 is the initial probability value, $P(Y\_(\ t)\ |S\_t\ )$ is the extent of observation probability, and $P(S\_t\ |S\_(t-1)\ )$ specifies the level of state transition.

$$P(S\_(1{:}T).Y\_(1{:}T)\ ) = P(S\_1\ )P(Y\_1\ |S\_1\ )\ \prod\_(t = 2)\hat{}T ▒ P(S\_t\ |S\_(t-1)\ )P(Y\_t\ |S\_t\ ) \qquad (1)$$

The HMM consists of forward and backward models. The forward model computes the probability of a state according to subsequent states. Meanwhile, the backward model computes the probability of a state according to previous states. The newly proposed method adopts the forward model [5].

3. 2. N-gram Tagging This model was presented at the IBM linguistics lab for the first time in an effort to recognize speech through the Tri-gram model and roughly 20000 dictionaries with over 8 trillion modeling parameters [4].

Note that an N-gram in the fields of computational linguistics and probability is in fact a contiguous order of *N* items from a certain speech or text sample. Depending on the application in use, the items could be phonemes, syllables, letters, words, or base pairs. In general, researchers gather the N-grams searching in a text or speech corpus. In cases where the items are in the form of words, the N-grams might also be termed 'shingles'.

In this model, an increased level of *n* leads to higher accuracy, whereas it may decrease the reliability of parameters fulfilled from a peace of limited training text. The HMM is used as a decision-making model in N-gram tagging.

3. 3. Uni-gram The Uni-gram tagging is the primary level of N_gram. Uni-gram is not dependent on any of previous or subsequent words/letters. Meanwhile, all calculations are made through Equation (2) based on occurrences independent from other words/letters. In this scenario, M represents the number of words/letters, while *i* represents the number of current words/letters [11].

$$p(w) = \sum\_(i = 1)\hat{}M ▒ (\mathrm{F}i) \qquad (2)$$

3. 4. Bi-gram The Bi-gram tagging method examines the occurrence probability of a role according to the previous or subsequent word/letter. The word weights are obtained through Bi-gram method according to Equation (3), where *M* indicates the number of words/letters, and *i* indicates the number of current words/letters.

$$p(w) = \sum_{(i=1)}^{(M-1)} \text{▒}(\mathrm{F}i \quad after \quad \mathrm{F}i + 1) \qquad (3)$$

3. 5. Tri-gram

In this paper, the Tri-gram tagging has been employed for the letters composing each word, as well as for the occurrence probability of each sentence role. Therefore, the word weights and the occurrence probability of roles can be obtained through Tri-gram tagging according to Equation (4), where M indicates the number of words/letters, and i indicates the number of current words/letters [11].

$$p(w) = \sum_{(i=1)}^{(M-2)} \text{▒}(\mathrm{F}i \quad after \ \mathrm{F}i + 1 \ and \ after \ \mathrm{F}i+) \qquad (4)$$

3. 6. Independent and Dependent Roles

Roles in the Persian language are classified into two categories of independent and dependent. Given the two categories, the compound roles will be as follows.

- Totally independent of other roles, primary roles are significant in terms of word type and position in a sentence. These roles include subject (agent), predicate, object, complement and verb.

Dependent roles generally come in four classes, namely adjective, governing genitive, apposition and bending. A precise examination of dependent compound roles reveals more than four classes, extending the number to nine, including noun, genitive, governing genitive, apposition, retroactive exclamation, governing transducer, dependent adverb, annunciator, bending, etc. [21].

4. PROPOSED METHOD

The new method is composed of various elements. This section first discusses the essential elements for the newly proposed method and then offers the general algorithm for sentence processing. Finally, a practical example explains all processing stages.

4. 1. Input

In this system, inputs are represented by words or phrases that are seprated by space characters "؟،!", sentences separated by "." as well as parsed Persian sentences. The sentence parsing can be completed by Pars Pardaz [7] or any other similar software.

4. 2. HMM Parameters

The HMM involves two types of probability distributions namely discrete and continuous. Since this research intends to obtain the roles of individual words, this paper adopts the discrete probability distribution through Equation (5).

$$\lambda = (A . B . \pi) \qquad (5)$$

The triad set in discrete HMM is the main parameter directly involved in HMM's decision-making. The rest of parameters indirectly participate in HMM's decision-making. Therefore, it is crucial to first obtain the required parameters in order to engage the HMM computations as well as to achieve the best possible solution to the role of each word in Persian sentences. These risks include:

- Number of possible states: In independent roles, $11^{(\text{number of words })}$ is true, where 11 is the number of compound roles in primary roles (subject, predicate, object, complement, verb) + letters in sentences + three spacing characters "؟،!". In addition, in each sentence, the number of words will be variable. In spite of the fact that predicate may contain subject and other items in our observations, in cases where the new system decides that the word in the sentence is predicate while failing to detect the type of predicate, the role will only be reported to be predicate. Moreover, the dependent compound roles are explored separately. In this section, the number of possible states is $17^{(\text{ number of words })}$, where the value of 17 is made of 11 dependent compound roles (noun, adjective, bending, genitive, governing genitive, dependent adverb, apposition, governing transducer, exclamation and annunciator), three spacing characters "؟،!", one unspecified, one verb and one letter. However, the number of words in this number of states will vary according to each input sentence.
- The number of observations depends on the number of words in a sentence. Therefore, the possible outputs that may be accepted by each words in the input sentence will appear as subject, predicate, object, complement, verb, noun, adjective, bending, genitive, governing genitive, dependent adverb, apposition, governing transducer, unspecified, letter, exclamation and annunciator.
- There are specific symbols to the number of all roles and even the type of words. Moreover, the states in each input sentence can be observed with more difficulty. Hence, the sequence of observations is achieved through Equation (6).

$$O = \{o_1 . \ldots . o_t\} \qquad (6)$$

4. 3. Initial Probability Distribution π={π_i }

This distribution is stored by a single-dimensional matrix with one line of information, indicating the probability of each role starting a sentence among all 194 training sentences. The initial value of each Persian sentence role is obtained according to Equation (7).

$$\pi_i = p\{q_1 = i\} . 1 \le i \le N \qquad (7)$$

The values of initial probability distribution have been displayed in Table 2. Since sentences are not likely to begin with certain roles, the phrasing states initiated by these roles do not require HMM calculation, since they are excluded from the set of possible states.

TABLE 2. Initial probability distribution (π)

Percentage of being a first word	Role	Percentage of being a first word	Role
4.895	Adjective	0.34	Verb
0	Governing genitive	4.545	Letter
0.349	Genitive	4.545	Adverb
0	Genitive	0	N.A.
1.748	Governing transducer	18.531	Subject
0	Bending	0.349	Subject
5.244	Retroactive	4.895	Predicate
1.398	Exclamation	0	Object
0.699	Annunciator	0	Complement
		0.699	Noun

4. 4. State Transition Matrix A=[a_ij]

This matrix comprises a set of transition probabilities between states, which is called MatrixA in this paper. Having determined different types of possible states, the percentage of transition in possible states is calculated through forward Tri-gram according to Equation (8), where *N* is the number of words, a_ij is the percentage of presence for each role after another role and then another role. Moreover, *i* is the number of matrix rows, which is equal to q_t followed by $q\_(t+1)$, i.e. current role, while *j* is the corresponding number of columns, which is equal to $q\_(t+2)$, i.e. two roles after the current role.

$$\begin{aligned} &A = [a\_ij\ ] \\ &a\_ij = p\{j = q\_(t+2) \mid i = q\_t.\, q\_(t+1)\ \},\ \ 1 \le i.j \le N. \\ &a\_ij \ge 0.\ \ 1 \le i.j \le N \\ &\sum\_(j=1)\hat{}N ▒〖a\_ij = 1.\ \ 1 \le i \le N〗 \end{aligned} \tag{8}$$

Since this method calculates the possibility of one or two subsequent roles, a large matrix is created with dimensions of 26×703. Furthermore, another reason behind the large size of matrix is the number of word types and roles, as well as three spacing characters in the matrix, which have led a total of 26 rows.

On the other hand, two possible states are assumed in rows equivalent to 26×26 since the Tri-gram model is involved. In addition, the Bi-gram method is explored for 2-word sentences adding 26 more items. As for 1-word sentences, one row covers Uni-gram, leading to a total of 703 matrix rows. The values of this matrix are derived from 194 sentences which are used to train the new system.

4. 5. Probability Distribution of Observations/ Matrix B

The individual letters composing each word in a sentence are examined to determine the weight of each word and the probability distribution of observations through Tri-gram tagging. For that reason, each role is distinguished in the database, followed by adoption of Tri-gram method, which displays the percentage of three letters occurring together in words larger than two letters.

As for 2- and 1-letter words, Bi-gram and Uni-gram are used respectively to calculate the weights of words corresponding to the number of letters. The dimensions of this matrix are 44×1936 for each of the roles. The reason behind the large number of matrix rows is that the table involves a possible compound of two Persian letters together, in addition to 1- and 2-letter words, making it a total of 1936 rows and 44 columns, including the Persian alphabet as well as the Arabic alphabets such as “ء، إ، أ، ي، ة، ؤ،”, letters “آ، (،)” and the spacing character. Therefore, Equations (9) and (10) are employed to calculate the distribution in each role.

$$\begin{aligned} &B = \{b\_j\ (k)\} \\ &b\_j\ (k) = p\{v\_k = o\_ + \mid j = q\_t\ \}.\ \ 1 \le j \le N.\ 1 \le k \le M \end{aligned} \tag{9}$$

where, v_k represents the *k*th symbol observed in the alphabet, $o\_ +$ is the vector of current input parameters, N is the number of words, J is the word counter, M is the number of letter in each word, and k is the letter count for each word [6, 10].

$$\begin{aligned} &b\_j\ (k) \ge 0.\ 1 \le j \le N.\ 1 \le k \le M \\ &\sum\_(k=1)\hat{}M ▒〖b\_j\ (k) = 1.\ \ 1 \le j \le N〗 \end{aligned} \tag{10}$$

4. 6. Output Calculation and Viterbi Algorithm

The Viterbi algorithm is the final decision-maker in this research. In fact, the independent roles and $17\hat{}(Number\ of\ character)$ possible states of dependent roles at this stage determine which state provides the best possible solution. In the best possible state, given the values, the initial probability distribution, transition state matrix, and observations probability distribution are calculated by HMM method, where the largest value is obtained with the Viterbi algorithm.

4. 7. Tri-gram Morphology General Algorithm

1. In the first stage, various word types in each of the roles are extracted separately.
2. Various types of possible sentence phrasing are obtained in Persian grammar.
3. State transition matrix (A): The value of Tri-gram tagging in the structure of sentence obtained in Stage 2 is calculated through level 3 N-gram. In fact, this stage statistically analyses what roles fall in t+1 and t+2 positions after/before each role at *t* the position. Equation (11) displays how the state transition matrix is calculated.

$$\mathrm{A} = [a\_ijk\ ]$$

$$\mathrm{a\_ijk} = \mathrm{p}\{k = \mathrm{q\_(t+2)}\ |j = \mathrm{q\_(t+1)}\ \big|\ i = \mathrm{q\_t}\ \}.\ \mathrm{a\_ijk} \geq 0.\ \ 1 \leq \mathrm{i.j} \leq \mathrm{N}\ . \qquad (11)$$

$$A(ijk) = \sum\_(k = 1)\text{^}(No\ roles\ \ )\text{▒}\sum\_(j = 1)\text{^}(No\ roles\ \ )\text{▒}\sum\_(i = 1)\text{^}(No\ roles)\text{▒}⟦\text{roles } i \text{ after } j \text{ after } k\ \ ⟧$$

4. Initial probability distribution (π): In order to calculate the initial probability distribution for the sentences obtained from Stage 2, we find out the percentage of cases where one of the roles occurred in the starting position of a sentence. Equation (12) shows how to calculate the initial probability matrix as one of the components required in HMM computations.

$$\pi = \{\pi\_i\ \}\ , \pi\_i = p\{q\_1 = i\}.1 \leq i \leq N \qquad (12)$$

5. Observations probability distribution matrix (B): At this stage, Tri-gram tagging is adopted to weigh the words. These calculations serve to obtain the word weights in each role, covering the letters of each word. In this matrix, the values of word weights in each role are obtained based on the percentage of cases where and what letters in *t*th position fall in t+1 and t+2. Hence, Equation (13) is used at this stage.

$$B(ijk) = \sum\_(k = 1)\text{^}(No\ char)\text{▒}\sum\_(j = 1)\text{^}(No\ char)\text{▒}\sum\_(i = 1)\text{^}(No\ char)\text{▒}⟦\text{word } i \text{ after } j \text{ after } k⟧ \qquad (13)$$

6. After completing the statistical calculations from Stages 1 to 5, the HMM is employed to specify the role of each word for each possible state in the sentence. Given the role of each word, the letters composing each word, and the possibility of presence for each role as the first word in a sentence, HMM can obtain values to determine whether each state is true as follows. These calculations are completed through Equation (14), where *n* represents the number of input sentence words.

$$\mathrm{MA}{=}\sum\_(i = 1)\text{^}n\text{▒}⟦\text{Frequency percentage of roles after } i \text{ and } i+1 \text{ and } i+2⟧ \qquad (14)$$

In Equation (15), the weight of each word in the input sentence is obtained through Matrix B, where the frequency percentage of letters is extracted and *m* is the number of letters in each word.

$$\mathrm{MB}{=}\sum\_(i = 1)\text{^}m\text{▒}⟦(\text{Frequency percentage of word } i\ after\ i + 1\ after\ i+2)⟧ \qquad (15)$$

The occurrence probability of each phrasing state might be obtained through HMM according to Equation (16).

MB: The level of transition state for each possible phrasing, MA is the level of observations probability for input sentence, and π: is the level of initial probability distribution for each phrasing.

$$P = \pi \times\ 〖MB〗\ \_(1..N) \times\ 〖MA〗\ \_(1..N) \qquad (16)$$

7. Once the occurrence probability of each sentence is calculated through Equation (17), the largest P is sent as the largest possible occurrence, indicating the role of each word in the sentence. Where, *s* indicates the number of phrasing states, and P is the occurrence probability of each state.

$$OutPut = \mathrm{max}┬s\ (\mathrm{P\ s}\ ) \qquad (17)$$

8. Due to using defuzzifier y Max (product), the effect of any lower value is less and the effect values are more [6].

4. 8. An Example of the Newly Proposed Morphology Tri-gram Algorithm

At first, the state transition matrix (A) is calculated according to Table 3 using Tri-gram tagging, statistical calculations and tables obtained from these calculations in Stages 1 and 2.

As noted in Stage 2 of the algorithm, possible phrasing states are created as many as possible for each sentence. The value of state transition matrix should be calculated for all states. For instance, the value of transition matrix for input Persian sentence “او به مدرسه رفت” literally translated into "he to school went" will be according to Table 3 for the possible state of "subject, preposition, complement, verb".

Following the state transition matrix calculations, the initial probability distribution of statistical calculations obtained from 194 different sentence phrasing types is calculated as training data in the system at Stage 4. Table (2) provides the possibility of each role starting a sentence (i.e. initial probability distribution).

In Stage 5, the probability distribution matrix (B) is computed according to Table 4. As noted earlier, however, Uni-gram and Bi-gran tagging methods are adopted in words where the number of letters is fewer than 3. Table 4 displays hypothetical sentence “رفت او به مدرسه” and hypothetical state "subject, preposition, complement, verb".

TABLE 3. Example of state transition matrix calculations for hypothetical state (A)

I	Tri-gram words	Tri-gram roles	Tri-gram occurrence value
0	مدرسه به، او،	Subject → preposition → complement	0.236
1	رفت مدرسه، به،	preposition → complement → verb	0.367
Conclusion:			0.236+0.367 = 0.603

TABLE 4. Example of Calculations for probability distribution matrix (B)

I	Tri-gram word	Tri-gram role	Tri-gram calculations of word in role	Tri-gram occurrence value
0	او	Subject	Occurrence value "ا" and then "و" as subject role	0.333
1	به	Article	Occurrence value "ب" and then "ه" as preposition role	0.9
2	مدرسه	Complement	Occurrence value "م" and then "د" and then "ر" + occurrence values of "د"و"ر"و"س" + occurrence values of "ه" و"س"و"ر" as complement role	0.742+0.456+0.389=1.587
3	رفت	Verb	Occurrence value "ر" and then "ف" and then "ت" as sentence verb role	0.649

Depending on the spacing characters, the type of words and the zero initial probability distribution are filtered to reduce the number of possible states and curtail the computation time of the project. Then, the HMM calculations are completed. The overall occurrence calculated for sentence "او به مدرسه رفت" and the value of hypothetical state "Subject, preposition , Complement, Verb" is calculated by Equation (18). Then, Tri-gram state transition is calculated for phrasing state "subject-proposition-complement-verb" according to Equation (18) using Table 3.

$$\mathrm{P(Subject, preposition\ , Complement, Verb)} \\ = \mathrm{P(subject, preposition\ , complement)} \\ * \ \mathrm{P(preposition\ , Complement, Verb)} * \ \mathrm{P(subject)}$$

Given the values in Tables 1 to 3 and Equations (18) to (19), the value of HMM for input sentence "او به مدرسه رفت" and hypothetical state "subject, preposition, complement, verb" will be according to Table 5.

As shown in Table 5, the calculations of Equations (18) and (19) are performed similar to Table 5 for each input sentence and all phrasing states. Moreover, the largest result is indicated as the input sentence compound.

5. PERFORMANCE STUDY WITH SIMULATION

In addition to the roles provided in previous section, there are "verb-letter" roles, which were discarded because of their shared decomposition and composition.

$$\text{Success percentages} = (\text{Success } \times 100)/(\ Total\ ) \quad (18)$$

Relying on Equation (18), we obtained the success rate in each section. Parameter *Total* in Equation (18) changes in each section. In addition, parameter *Success* in this regard varies according to each section.

Also the main software used to exploit the proposed method of visual studio 2019- visual basic software, for primary statistical work from the office 2019 collection Excel 2019 software was used and then in the programming environment the results obtained from Excel to SQL server 2017 is used.

This section makes a comparison between the simulation results obtained by the proposed method and those of other methods. Accordingly, the success rate of the new method is comparatively examined with the HMM Uni-gram and Bi-gram tagging techniques.

TABLE 5. Final HMM calculations for sentence "او به مدرسه رفت". (rows 1, 2, 3, 4 total of percentage of occurrence for letters, and 5, 6, 7 percentage of occurrence probability of the state).

No.	Phrase	Matrix name	Percentage value
1	P(او ،subject)	Observations probability distribution matrix (B):	0.333
2	P(به ،preposition)	Observations probability distribution matrix (B):	0.9
3	P(مدرسه ،complement)	Observations probability distribution matrix (B):	1.587
4	P(رفت ،verb)	Observations probability distribution matrix (B):	0.649
5	P(subject, preposition, complement)	State transition matrix (A):	0.236
6	P(preposition, complement, verb)	State transition matrix (A):	0.367
7	P(subject)	Initial probability distribution (π):	0.634
8	P(Subject, preposition, Complement, Verb)	With values of rows 5, 6 and 7	0.236×0.367×0.634= 0.054
Result	P(او،به،مدرسه،رفت|Subject, preposition, Complement, Verb)	Using rows 1 to 8	0.333×0.9×1.587×0.649× 0.054= 0.016

5. 1. Results of Tri-gram Tagging with HMM

The average success rate of this method is 73.34608 percent in identifying the roles in Persian sentences.

As displayed in Table 6, the success rates of both categories of roles are less than 1%, which indicates the approximate equality of success rate in calculating the word roles in Persian sentences. According to Table 6, this difference has been demonstrated in Figure 1.

As displayed in Table 7, the success percentage lies within interval 40-98.55343, where the success rate follows an ascending order "complement, object, predicate, proposition, verb and subject".

According to Figure 2, the highest and lowest success rates were achieved by "subject" and "complement", respectively. It is worth noting that in reality, however, "agent and pronoun" are two roles categorized as "subject", the average success rate of which is 82.5%.

TABLE 6. Average success rates for both categories of compound roles through HMM and Tri-gram tagging

Category	Value %
Independent roles	73.9816
dependent roles	73.34608

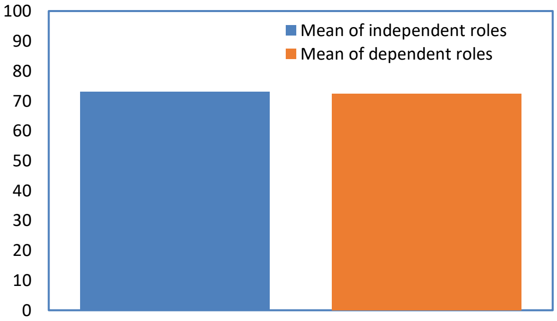

Figure 1. Average success in finding compound roles in two separate categories

TABLE 7. Average success rates for independent compound roles through HMM and Tri-gram tagging

No.	Role	Value %
1	Verb	91.64355
2	Predicate	67.76667
3	Subject	98.55343
4	Subject	79.567995
5	Object	56.98756
6	Complement	40
7	Letter	86.97294

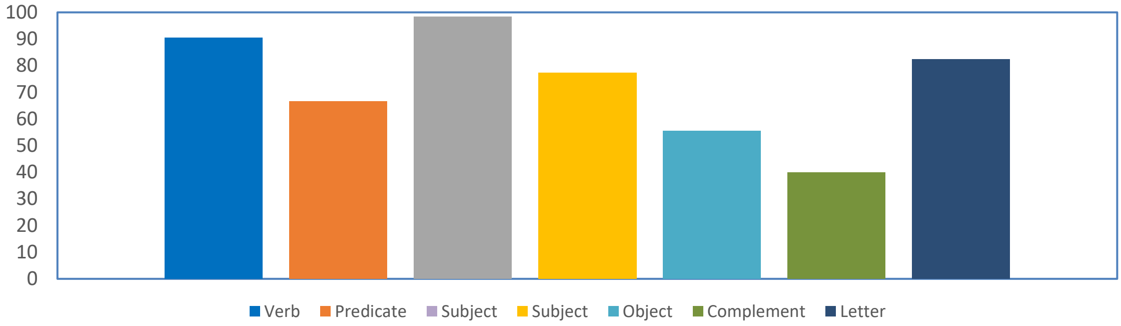

Figure 2. Success rate of each primary role with Tri-gram tagging and HMM

TABLE 8. Average success rates for dependent compound roles through HMM and Tri-gram tagging

No.	Role	Value in terms of %
1	Adjective	22.22222
2	Noun	72.72727
3	Adverb	55.17241
4	Unspecified	77.55102
5	Governing genitive	20
6	Genitive	20
7	Apposition	100
8	Governing transducer	100
9	Bending	100
10	Retroactive	100
11	Exclamation	100
12	Annunciator	100

Among these 12 roles in Table 8, the 4th role (unspecified) indicates the percentage of words without any specific dependent roles in the sentence. Table 8 provides the success rates of dependent role tagging in third-order N-gram and HMM. The range of variations in the success rate of tagging is 80% for HMM. Since these roles are overlapping, it is difficult to determine the success rate of dependent compound roles.

As can be seen in Figure 3, the success rate in dependent roles “apposition, governing transducer, bending, retroactive, exclamation, annunciation”,

which are basically rare roles in Persian sentences, is 100%. One reason is that filtering is used to alleviate the computational load in the new method. Meawhile, the lower frequency of such roles affects the results. According to Figure 3, however, the smallest values are related to "genitive, governing genitive and adjective".

600 dpi and without borders, with capital first letter of axis titles and write its unit (all the Figures and Tables should be placed on the top or in the bottom of the page; not in the middle of text).

5. 1. Comparative Overview of Uni-gram, Bi-gram and Tri-gram Tagging Techniques with HMM

One of the most important advantages of the proposed method is its improvement in tagging. Considering that previous studies have already calculated these values through Uni-gram and Bi-gram tagging methods, the current paper focused on Tri-gram tagging.

Table 9 displays the difference in improvement of success in detection of roles in Persian sentences through HMM. The shift from first-order to second-order N-gram is about 5.7%. However, there is approximately 20% improvement in the success of detecting compound roles from second to third order.

Figure 4 demonstrates the difference in success rates of tagging styles in an ascending trend. Table 10 displays the average success of independent roles by three tagging styles of Uni-gram, Bi-gram and Tri-gram in the HMM.

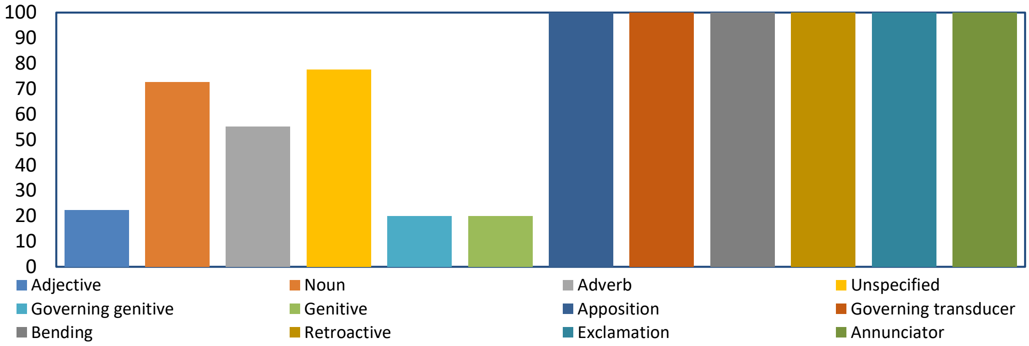

Figure 3. Success rate of each dependent role with Tri-gram tagging and HMM

TABLE 9. Average success rates of Uni-gram, Bi-gram, and Tri-gram tagging techniques through HMM

Tagging Method	Value
Uni-gram	45.49315
Bi-gram	50.02104
Tri-gram	73.84364

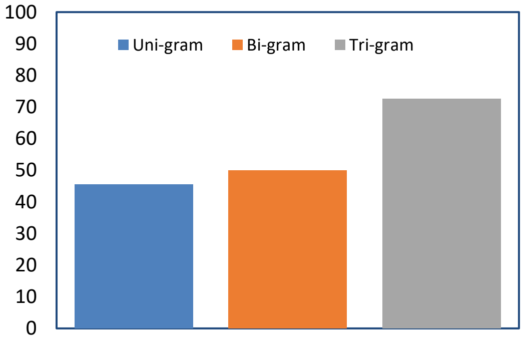

Figure 4. Comparison of overall results from Uni-gram, Bi-gram and Tri-gram tagging techniques

TABLE 10. Comparison of average success rates for independent roles through HMM and Uni-gram, Bi-gram and Tri-gram

Tagging style	Independent roles
Uni-gram	55.3028
Bi-gram	52.4129
Tri-gram	73.9816

As can be observed, there is a slight difference between the independent roles for Uni-gram and Bi-gram tagging styles, whereas the difference from Tri-gram is roughly 20%.

As shown in Figure 5, the slope of variations in independent roles is not identical in each stage. In addition, the lowest average values were found in Bi-gram. As displayed in Table 11, comparison of success rates for Uni-gram, Bi-gram, and Tri-gram tagging techniques in dependent roles reveals a difference of approximately equal from 11% to 20% in each stage. Nevertheless, these values arise from dependent roles, the detail of which do not indicate such difference in success rates.

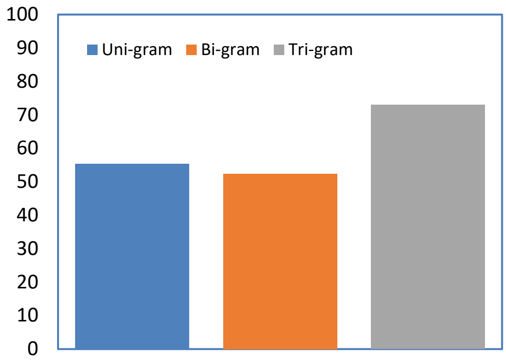

Figure 5. Average success rates for independent compound roles in Uni-gram, Bi-gram and Tri-gram

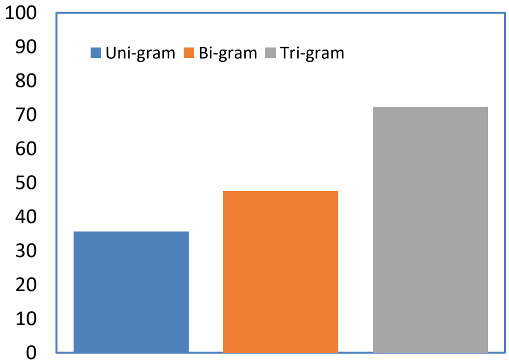

Figure 6. Average success rates for independent compound roles in Uni-gram, Bi-gram and Tri-gram

TABLE 11. Comparison of average success rates for dependent roles through HMM and Uni-gram, Bi-gram and Tri-gram

Tagging style	dependent roles
Uni-gram	35.683907
Bi-gram	47.629573
Tri-gram	73.34608

As shown in Table 7, the lowest and highest success rates in this category of roles were found in Uni-gram to Tri-gram, respectively. Average success rates for dependent roles through HMM and Uni-gram, Bi-gram and Tri-gram have been shown in Table 12.

The values of verb and letter, however, have been excluded since they were extracted from decomposition. The minimum value is for object in Uni-gram tagging, while the maximum value is for subject in Tri-gram tagging. As shown in Table 8, the highest value was found in "subject" while the lowest value was found in "complement", except Uni-gram tagging where the lowest value was found in "object".

Since "apposition" is rarely found in a Persian sentence, it was not calculated in Uni-gram and Bi-gram tagging styles. Nonetheless, rows 7 to 12 indicate rare roles because there is a little statistical population. In case of no value is identified, a large percentage is lost.

TABLE 12. Average success rates for independent roles through HMM and Uni-gram, Bi-gram and Tri-gram

No.	Role	Uni-gram tagging %	Bi-gram tagging %	Tri-gram tagging %
1	Predicate	24.39	60.975	67.66667
2	Subject	51.162	88.37	98.93333
3	Subject	51.351	48.648	77.511935
4	Object	18.818	31.818	55.65556
5	Complement	33.333	8.33	40

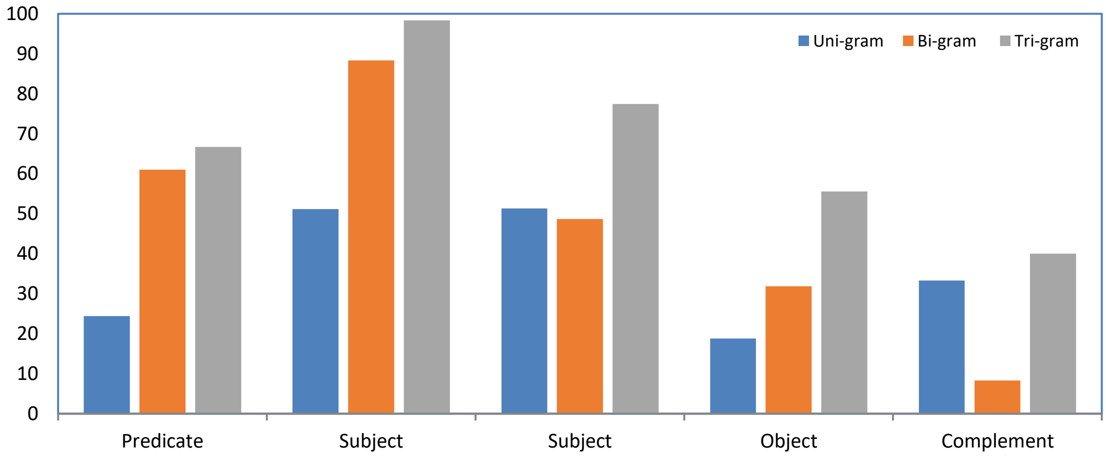

Figure 7. Success rates for independent compound roles in Uni-gram, Bi-gram and Tri-gram

Table 13 shows the smallest values were found in "genitive" following "exclamation, annunciator, apposition", indicating 0% in the two tagging techniques of the first stage. As can be seen from the results in Figure 8, the smallest values were found in "genitive" following "exclamation, annunciator, apposition", indicating 0% in the two tagging techniques of the first stage. Moreover, the highest values were found in “bending, retroactive and unspecified” [30, 31].

TABLE 13. Average success rates for dependent roles through HMM and Uni-gram, Bi-gram and Tri-gram

No.	Role	Uni-gram tagging by %	Bi-gram tagging by %	Tri-gram tagging by %
1	Adjective	60	30	23.22222
2	Noun	25	66.66	74.72728
3	Adverb	96.299	81.482	56.17242
4	Unspecified	100	75	79.55103
5	Governing genitive	29.413	35.295	20
6	Genitive	14.281	21.429	20
7	Apposition	-	-	100
8	Governing transducer	33.333	66.666	100
9	Bending	75	100	100
10	Retroactive	75	100	100
11	Exclamation	100	0	100
12	Annunciator	0	0	100

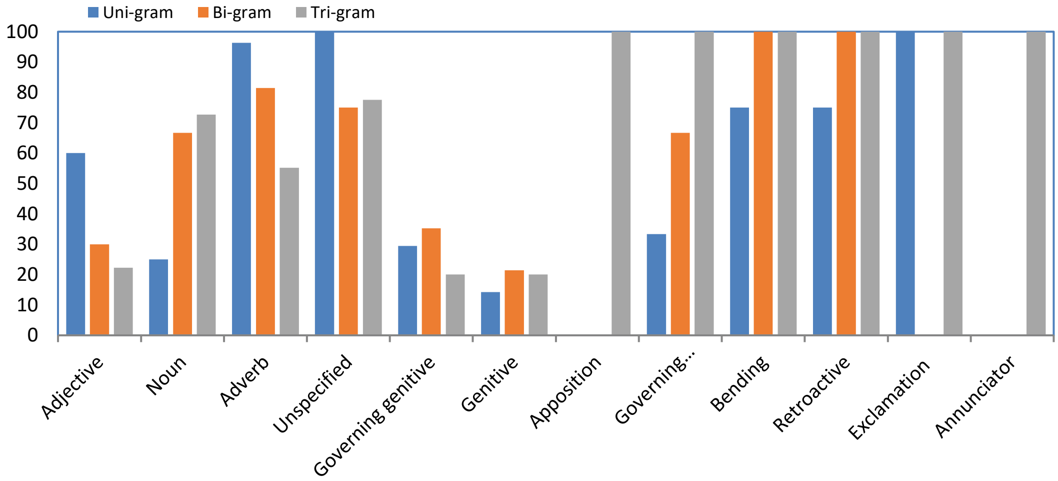

Figure 8. Average success rates for dependent compound roles in Uni-gram, Bi-gram and Tri-gram

6. CONCLUSIONS AND FUTURE WORK

The results of simulation in the proposed method suggest that Tri-gram tagging achieved improvement about 20% higher than Bi-gram tagging. The same amount of improvement was found in both independent roles and dependent roles. The results of Tri-gram tagging indicated that the lowest success rate is 40%, whereas success rates in Uni-gram and Bi-gram are 18% and 8%, respectively. Moreover, there are very different values for independent roles associated with rare roles. Nevertheless, these roles have been greatly improved by Tri-gram tagging method. There are several explanations for improvement by the new method: 1) high accuracy in matrices applied in HMM under Tri-gram tagging style, 2) the filtering applied on possible states making a key contribution to the calculations. In fact, calculations can be lightened through filtering according to Persian grammar. In addition to reducing the computational load in this scenario, this type of calculations shifts from statistical to rule-based or hybrid.

Future studies can focus on the new method in combination with other morphology techniques in an effort to compare its effects on success rate. Finally, different methods can be investigated to curtail the computational load of the statistical procedure.

7. REFERENCES

1. Yoonseok, H., Sangwoo, K. and Donghyun, Y., "Multimodal Neural Machine Translation with Weakly Labeled Images," ***IEEE Access***, Vol. 7, 54042-54053, (2019), doi: 10.1109/ACCESS.2019.2911656.
2. Alshammari, M., Nasraoui, O., and Sanders, S., "Mining Semantic Knowledge Graphs to Add Explainability to Black Box Recommender Systems," ***IEEE Access***, Vol. 7, 110563-110579,(2019),doi: 10.1109/ACCESS.2019.2934633.
3. Wu, B., Kehuang, L., Fengpei, G., Zhen, H., Minglei, Y., Chin, L., and Chin-H, L., "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition," ***IEEE Journal of Selected Topics in Signal Processing***, Vol. 11, No. 8, (2017), 1289-1300. DOI: 10.1109/JSTSP.2017.2756439.
4. Vani, H., Anusuya, M., "Fuzzy Speech Recognition: A Review," ***International Journal of Computer Applications***, Vol. 177, No. 47, (2020), 39-54. DOI:10.5120/ijca2020919989.
5. Xia, T., Chen, X. (2020). A Discrete Hidden Markov Model for SMS Spam Detection. ***Applied Sciences***. 10. 5011. 10.3390/app10145011.
6. Motameni, H., Peykar, A., "Morphology of Compounds as Standard Words in Persian through Hidden Markov Model and Fuzzy Method, 2015.," ***Journal of Intelligent & Fuzzy Systems***, Vol. 30, No. 10.3233/IFS-151865, (2016), 1567-1580. DOI: 10.3233/IFS-151865.
7. Peykar, A., Motameni, H., Aboutalebi, M. "Application of fuzzy identification method depends on the synthesis of the Persian language," in Conference iran data mining Iran, Tehran, 2014.
8. Peykar, A., Motameni, H., Aboutalebi, M. "Comparison of fuzzy and hidden Markov model to identify independent of synthesis words in Persian," in Conference iran data mining Iran, Tehran, 2014.
9. Asghari, R. "Application of N-gram modeling in language statistical modeling. (Persian)," in International Conference on Nonlinear Modeling & Optimization, Amol, Iran, 2012.
10. Keysers, D., Deselaers, T., Rowley, H., Wang, L. and Carbune, V., "Multi-Language Online Handwriting Recognition," ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 39, No. 6, (2017), 1180-1194. DOI: 10.1109/TPAMI.2016.2572693
11. Peykar, A., Motameni, H., Aboutalebi, M. "study of the role of labeling N_Gram, terminology and phrases in Farsi, hidden Markov models," in third national conference on computational linguistics, Tehran, 2014.
12. Obin, N., Lanchantin, P., "Symbolic Modeling of Prosody: From Linguistics to Statistics," ***IEEE/ACM Transactions on Audio, Speech, and Language Processing***, Vol. 23, No. 3, (2015), 588 - 599. DOI: 10.1109/TASLP.2014.2387389.
13. Sahraee Juybari, M., Bozorgian, H., "Cultural Linguistics and ELT curriculum: The case of ‘Prospect’ English textbooks in Iran",Vol.30,Issue3, (2020),https://doi.org/10.1111/ijal.12301.
14. Lücking, A., Driller, C., Stoeckel, M., Abrami, G., Pachzelt, A. and Mehler, A., "Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology", ***Language Resources & Evaluation***, (2021). https://doi.org/10.1007/s10579-021-09553-5.
15. Fang, K. L., "A Short History of Linguistics R. H. Robins," American Anthropologist, Vol. 70, (2009), 1186-1186.
16. Moniri, M., "Fuzzy and Intuitionistic Fuzzy Turing Machines.," ***Fundamenta Informaticae***, Vol. 123, No. 3, (2013), 305-315. DOI: 10.3233/FI-2013-812.
17. Meghdari, A., Alami, M., "Phrases from well-known social robotics," in Symposium on gateways to the field of cognitive science, Tehran, (2015).
18. Abid, M., Habib, A., Ashraf, J., Shahid, A., "Urdu word sense disambiguation using machine learning approach". ***Cluster Comput***, Vol. 21, (2018), 515-522 https://doi.org/10.1007/s10586-017-0918-0.
19. Austin, P. "Theory of language: a taxonomy". ***SN Soc Sci*** 1, 78 (2021). https://doi.org/10.1007/s43545-021-00085-x.
20. Bijankhan, M., Sheykhzadegan, J., Bahrani, M. and Ghayoomi, M., "Lessons from Building a Persian Written Corpus: Peykare," ***Language Resources and Evaluation***, Vol. 45, (2011), 143-164. DOI: 10.1007/s10579-010-9132-x
21. Baghaei P., Khoshdel-Niyat, F. & Tabatabaee-Yazdi, M."The Persian adaptation of Baddeley’s 3-min grammatical reasoning test", ***Psicologia: Reflexão e Crítica*** Vol. 30, No. 16, (2017), https://doi.org/10.1186/s41155-017-0070-z.
22. Yusupov, A., Yusupova, N., Sibgatullina, A. Grammatical Absorption and Functioning of Arab and Persian Conjunctions in Old Tatar Language in the 19th Century. ***International Journal of Society, Culture & Language***, 8, 3 (Special Issue on Russian Culture and Language)) (2020), 80-88.
23. Web, A. F., "Natural Language Processing Software of Ferdowsi University of Mashhad Version 1.3.(persian)," Web Technology Lab of Ferdowsi University of Mashhad, Mashhad, (2012).
24. Sadeghi, H., Motameni, H., Ebrahimnejad, A. and Vahidi, J., "Morphology of composition functions in persian sentences through a newly proposed classified fuzzy method and center ofgravity defuzzification method," ***Journal of Intelligent & Fuzzy Systems***, Vol. 36, No. 6, (2019), 5463-5473. DOI: 10.3233/JIFS-181330
25. Safari, A., Mazinani, M. and Hosseini, R., "A Novel Type-2 Adaptive Neuro Fuzzy Inference System Classifier for Modelling Uncertainty in Prediction of Air Pollution Disaster", ***International Journal of Engineering, Transactions B: Applications***, Vol. 30, No. 11, (2017), 1746-1751. doi: 10.5829/ije.2017.30.11b.16
26. Gardani, F. "Borrowing matter and pattern in morphology. An overview". ***Morphology***, Vol. 30, (2020), 263-282, https://doi.org/10.1007/s11525-020-09371-5.
27. Alexis Amid, N., Éric, L., "Pattern-and-root inflectional morphology: the Arabic broken plural". ***Language Sciences***, Vol. 40, (2013), 221-250, https://doi.org/10.1016/j.langsci.2013.06.002.
28. Pakendorf, B. "Lamunkhin Even evaluative morphology in cross-linguistic comparison". ***Morphology,*** Vol. 27, (2017), 123-158 https://doi.org/10.1007/s11525-016-9296-1.
29. Sagot, B., Walther, G. "A Morphological Lexicon for the Persian Language," in LREC 2010, Valletta, Malta, 2010.
30. Megerdoomian, K. "Finite-State Morphological Analysis of Persian," in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages., Stroudsburg, 2004.

31. Mor, B., Garhwal, S., Kumar, A. "A Systematic Review of Hidden Markov Models and Their Applications". ***Archives Computational Methods in Engineering,*** Vol. 28, No. 3 1429–1448, (2021). https://doi.org/10.1007/s11831-020-09422-4.

32. T. Buckwalter, "Buckwalter Arabic Morphological Analyzer.," the Linguistic Data Consortium,, Pennsylvania, 2002, https://doi.org/10.35111/050q-5r95.

33. Motameni, H. Determining the Composition Functions of Persian Non-standard Sentences in Terminology using a Deep Learning Fuzzy Neural Network Model. ***International Journal of Engineering, Transactions C: Aspects***, (2020); 33(12): 2471-2481. doi: 10.5829/ije.2020.33.12c.06

34. Azimizadeh, A., Arab, M., Quchani, S. Persian part of speech tagger based on Hidden Markov Model. In JADT 2008 : 9th international conference on textual data statistical analysis, pages 121–128, March 2008.

35. Okhovvat, M, Minaei Bidgoli, B, "A hidden Markov model for Persian part-of-speech tagging", ***Procedia Computer Science***, Vol. 3, (2011), 977-981, https://doi.org/10.1016/j.procs.2010.12.160.

36. Seraji, M., Megyesi, B., Nivre, J. "Dependency parsers for Persian". In Proceedings of the 10th Workshop on Asian Language Resources, (2012), 35-44.

37. Kardan, A., Imani, M. "Improving Persian POS tagging using the maximum entropy model". In 2014 Iranian Conference on Intelligent Systems (ICIS), (2014), 1-5, doi:10.1109/IranianCIS.2014.6802567.

38. Nourian, A., Rasooli, M., Imany, M., Faili, H. "On the importance of ezafe construction in Persian parsing". In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 2: Short Papers), Vol. 2, (2015), 877-882.

39. Pakzad, A., Minaei Bidgoli, B., "An improved joint model: POS tagging and dependency parsing". ***Journal of AI and Data Mining***, Vol. 4, No. 1, (2016), 1-8, ISSN 2322-5211. doi:10.5829/idosi.JAIDM.2016.04.01.01.

Persian Abstract

چکیده

امروزه به دلیل محبوبیت شبکه های اجتماعی و نیز ورود کلمات مختصر، کلمات خارجی و شکلک‌ها در زبان فارسی، اهمیت داده‌کاوی افزایش یافته و پژوهش هایی در رابطه با شناسایی نوع کلمات انجام شده است. این در حالی است که تشخیص نقش کلمه در جمله مهمتر از تشخیص نوع کلمه در جمله می باشد. از طرفی تشابه املایی- دستوری زبان فارسی به زبان عربی، موجب شده تا بتوان از روش پیشنهادی این مقاله در زبان عربی نیز استفاده کرد. در این مقاله، جهت واژه شناسی نقش‌های ترکیب جملات زبان فارسی از روش آماری مدل مخفی مارکوف و برچسب گذاری Tri-gram استفاده شده و با روش مدل مخفی مارکوف و برچسب‌گذاری Uni-gram و Bi-gram مقایسه شده است. در روش پیشنهادی با استفاده از دو دسته نقش‌های "مستقل" و "وابسته"، و عوامل پذیرش نقش کلمات در جملات، نتایج شناسایی نقش کلمات را بهبود می‌بخشد. به طوری که نتایج شبیه سازی نشان می دهد که میانگین موفقیت نقش های مستقل ترکیب با مدل مخفی مارکوف و برچسب گذاری Tri-gram نسبت به روش Uni-gram و Bi-gram به ترتیب ۲۰.۵۶ و ۱۷.۶۷ درصد و برای نقش های وابسته ترکیب به ترتیب ۲۴.۶۷ و ۳۲.۶۲ درصد، بهبود داشته است.