Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 9, No. 4, November 2018), Pages: 145-155
www.jacr.iausari.ac.ir

Terminology of Combining the Sentences of Farsi Language with the Viterbi Algorithm and BI-GRAM Labeling

Ebrahim Akbari✉, Homayun Motameni
Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran
akbari.ebrahim@iausari.ac.ir; motameni@iausari.ac.ir

Received: 2018/11/18; Accepted: 2019/01/14

Abstract

This paper, based on the Viterbi algorithm, selects the most likely combination of different wording from a variety of scenarios. In this regard, the Bi-gram and Unigram tags of each word, based on the letters forming the words, as well as the bigram and unigram labels After the breakdown into the composition or moment of transition from the decomposition to the combination obtained from the types of sentences, the educator is used in 194 different wording types, and the sum of them is obtained by the amount of the advance of each wording state and the MAX value is considered as the output of the system. And at the end, the success rate of these methods and the effectiveness of these two types of labeling are compared with each other.

Keywords: *Viterbi Algorithm, Terminology, Independent Roles, Dependent Roles*

1. Introduction

The idea of the emergence of artificial intelligence in the computer world began with the passing of hard and hard work to computers and Robots. Of the more important parts of the goal for scientists, artificial intelligence, linguistics. The reason for this importance can be extensive and important applications in the field of intelligence in every sector of our world is artificial humans. Terminology Part of linguistics is part of data mining. The Researchers Persian language linguistics, to stay ahead of the quick moving Studies have offered valuable impact. The paper also tries to take a step, however small, to strengthen AI in Persian rather than its benefits in other sectors will also benefit.

2. Related Works

The '80s began seriously terminology in English. Can be Petr Trojanskij Russian researcher in the year (1933) was the initiator of this research. Data mining research and terminology in Persian by, "Dr Mostafa Asi" (1371), the Institute for Humanities started. Of course, (2000) "dr Asi and Haji Abdolhosseini" in the tag word in Persian literature studies have done. "Dr Mehrnoush Shams Fard" Year (1993), Research Extensive to date in the field of language Have been computerized. Linguistics scholar in Persian, "doctor Mahmoud B. Jen Khan" from 1375 to date have been impressive research, including the results of the research building "body text Persian language"

known as the "body of Jen Khan "is. Pierre-year research study in linguistics Persian can be "Azimi" (1375), who produce and understand speech in Farsi and "Ghasemi" in (1377) to the first data mining research with Speech recognition issue. And further, the "eslami" (1379) research focuses on the structure of musical compositions in Persian. "Gholampoor" in the year (1379) Using matching sounds of words such as input, with a variety of phonetic models, detection is done sentences. "Riazati "(1997), phonological and morphological terminology on two levels, has done in the Persian language. The first machine translation project called " Shiraz " For automatic translation of Persian texts into English by "Amtrupt And colleagues "(1998-2000) , and further this research project by "Megerdoomian "In the years 2000 and 2004 Based on morphological description of Persian vocabulary was developed in 2008 to the construction of large-scale universal bank term. Data mining research in recent years, mainly in the field of speech recognition, such as research, "Ghasemzadeh and Rahimi" (2006), "Sharafi et al" (2007) and "Shams fard and fadaei" (2008) "Farhad Aroumchian" (2007), grammatical Labeling TNT Based on the hidden model of Markov. Investigating "the body " (2012) With the theme Labeling Persian language sentences based on the law as well as research oriented "Momeni et al"(2015) fuzzy method and HMM , Based on the statistical-based method, used to identify the role of words . In addition they can be used in morphological research analyst "Keeper" (2004), "Dehdari, Lonsdale "(2008) pointed out. Ultimately, including recent research on the subject Translator automated "Feili, Ghassem Sani" (2004) as well as "Saw Et al" (2009) pointed out.

3. Proposed Method

In this first part of the requirements for the proposed method is then applied to all stages of the process will be explained with an example.

3.1 Input

The inputs of this system are sentences separated from each other in Persian, as well as parsed sentences Is. For parsing sentences, one can use (Pars Pardaz) or other similar software that parses the sentences of the sentences.

3.2 Weight of words by labeling Bi-gram

This parameter is derived from the combination of words forming words. Basically a statistic of bi_gram The use of any of the words Excel 2013 The result is a table with 44 x 44 dimensions , including44 alphabetical characters and separator characters. The formula that can be derived from the letters of the alphabet, each of the words of the input, and the letters that make them the weight of each word in each part is given in the expression 1. After, gaining the weight of each word per role using expression 1. The number of words, including the two - dimensional array with dimensions (number of words, 24), forms and is used in calculations.

$$\text{Word weight} = \sum_{i=1}^{n} \text{The word i} + \text{I word} + \text{Letter 1}$$

In phrase 1, i the number is the letter of each word and the number of letters is each word.

3.3 The sum of the consequences of each role is combined in bigram The role of decomposition

The amount of the event, if the word n in decomposition, it is a type of word property, what role does it play in the word n + 1 I will combine this relationship in the phrase 2 is coming. in this case bi-gram For each pair of words, the value of this came from the table values in the Excel program It has been calculated and added to the program database with dimensions 22 × 10 in the program and is summarized as in the table below.

Relation bigram Composition and analysis = $\sum_{i=1}^{n}$ *Word type in parsing i* ***+ I word +the role of word in composition 1***

The value for the entire sentence is added for each state by adding to the value of the state in which the value of the words of that state has already been added. This event is calculated for sentences of more than one word. And is added to the weight of each word added in the composition matrix.

In the dimensions of this matrix, there are 10 lines to the number of separating characters and the type of words in the Farsi sentences, and 22 are the number of separating characters and the role of the sentences of the Persian language. It is noteworthy that this matrix, a fixed amount of 193 such samples have been obtained by statistical calculations.

3.4 Types of wording scenarios

In fact, given the number of words in the input sentences, the number of possible states will vary. In fact, in this system, to avoid the interference of the role of sentences, these sentences are divided into two independent and dependent roles. The role of independent (subject, Subject, predicate, object, Amendment), the affiliate (adjectives, adverbs, Genitive, possessive, added, fake, turned, inclined, call, proclaimed) in There are both verb forms and separator characters and prepositions. Given the number of roles in each category, the number of possible scenarios will vary. The number of possible states for independent roles 10(number of words) and for dependent roles 17(number of words). For example, Table 4, part of the possible scenarios of independent roles, is given for a 5-word sentence.

3.5 Output

After obtaining the weight of the words and the outcome of each role, after each breakdown, the resulting number of possible states is obtained and their values for each input sentence are stored in each of the states in its matrix, and at the end of the state that most the total amount of the earned as output is shown. In fact, the Viterbi algorithm is used to decide output and the event is shown as output. Of course, it should be noted that if the sentence contains only one word, then the method uni-gram to decide on the role of the word in the sentence is used.

4. Processing Steps Including “Sara Lives in the Village” with Tag Bigram

In this section, including the hypothetical "سارا در روستا زندگی می کند" is considered as input all the steps listed above that can be found in the weight of words, the event, the weight of each scene may be output as independent states.

Table 1 Weight word "سارا" in the subject, Table 2 Weight of the word "روستا" in the role of complementary and Table 3 the word "زندگی می کند" in action with hypothetical values show.

Table 1 - Example of weight calculation word tag Bi-gram for the word "سارا" in the role of "subject"

the amount of B-gram	The letter i+1	The letter "I"	Row
0.8	' /a/ا '	' /S/س'	1
0.56	' /r/ر'	' /a/ا'	2
0.66	' /a/ا'	' /r/ر'	3
2.02			Total

Like table 1 for each individual role in this study, we calculate the amount of word weight in the role, and then with the probability value, each component of the composition is summed up instead of the fractional role and its value is placed in the possible state tables.

Table 2 - weight calculations word tag Bi-gram for the word "روستا" in the role of "supplement"

the amount of B-gram	The letter i+1	The letter "I"	Row
0.31	'و/o/'	'ر/r/ '	1
0.89	'س/s/'	'و/o/ '	2
0.37	'ت/t/'	'س/s/ '	3
2.07	'ا/a/'	'ت/t/ '	4
1.64			Total

The word "در" Because the preposition and its role remains so preposition instead of the word analysis and processing without accepting the new role as prepositions are shown.

Table 3 - weight calculations word tag Bi-gram for the word "زندگی می کند" in role "Verb"

The mount of Bi-gram	i+1 The letter	The letter "i"	Row
0.2	'ن/n/'	'ز/z/ '	1
0.5	'د/d/'	'ن/n/ '	2
0.57	'گ/g/'	'د/d/ '	3
0.24	'ی/y/'	'گ/g/ '	4
0.63	'/ / '	'ی/y/ '	5
0.17	'ی/i/'	'م/m/ '	6
0.78	'ک/k/'	'ی/i/ '	7
0.23	'ن/n/'	'ک/k/ '	8
0.29	'د/d/'	'ن/n/ '	9
3.61			Total

After gaining the weight of each word in each role, it is the turn of the chance to proceed Bi-gram The role of the indicators that word is so hypothetical values for the "سارا در روستا زندگی میکند " and parse input " the name, words, the name, verbs, verb" to Table 4 Will be.

Table 4 - Possibility of occurrence bi-gram for incoming entries "سارا در روستا زندگی میکند" and Log in The decomposition of" The name, the word, the name, Verb"

The mount of Bi-gram	Type the word i+1 in Compound	Type the word “i” in Analysis
0.68	Letter	Name
0.9	Complement	Letter
0.5	Verb	Verb
2.08	Total	

Values can be summed together for each possible state. For example, for 4 selectable modes and hypothetical values of table 5 How to calculate the stage for the entry " سارا در روستا زندگی میکند" and parse input " nouns, words, nouns, verbs, verb" show.

Table 5 - Sample calculation of the values of any hypothetical state for the incoming sentence " سارا در روستا زندگی میکند" and the decomposition input "noun, letter, noun, and verb"

total	Bi-gram occurrence of each word in the respective roles	The weight of each word in the role of the state	The role of the state	State symbol
20.95	1.63	3.61+1.98+2.23+10+1.5	Subject , letter , predicate , predicate, verb	CHDDA
18.62	0.78	3.61+0.5+2.23+10+1.5	Subject , letter , predicate , Subject, verb	CHDAA
20.25	2.01	3.61+1.98+0.63+10+2.02	Subject , letter , Subject, Predicate, verb	AHCDF
22.98	2.08	3.61+3.61+1.66+10+2.02	Subject , letter , complement, verb, Verb	AHJFF

The highest value of a variety of states is sent to the output as the response of this method; therefore, according to Table 5, the sentence "سارا در روستا زندگی میکند" subject, word, complement, verb, verb" respectively.

5. The Results of the Proposed Method

By reviewing and comparing the results of the method bi-gram and unigram both read the Viterbi algorithm used and the results it can be expressed. And the rate of momentum conversion from decomposition to compound (Uni-gram) Or Bi-gram Split analysis will be checked. So the results of Uni-gram Instead of parsing each letter and word, that letter and word combinations are also achieved with this method, the output of the Viterbi algorithm labeling Bi-gram Compared.

5.1 The overall success rate in the combination of sentences in Persian language

Of the total of 415, the total number of independent and dependent roles in Bi-gram 230 number of role sare properly identified and Uni-gram 305 number is correctly specified. The percentage of success in the role of independent labels Uni-gram against

73.49% and Bi-gram it is equal to 55.42%. Therefore, the diagram of these successes can be shown in Figure 1.

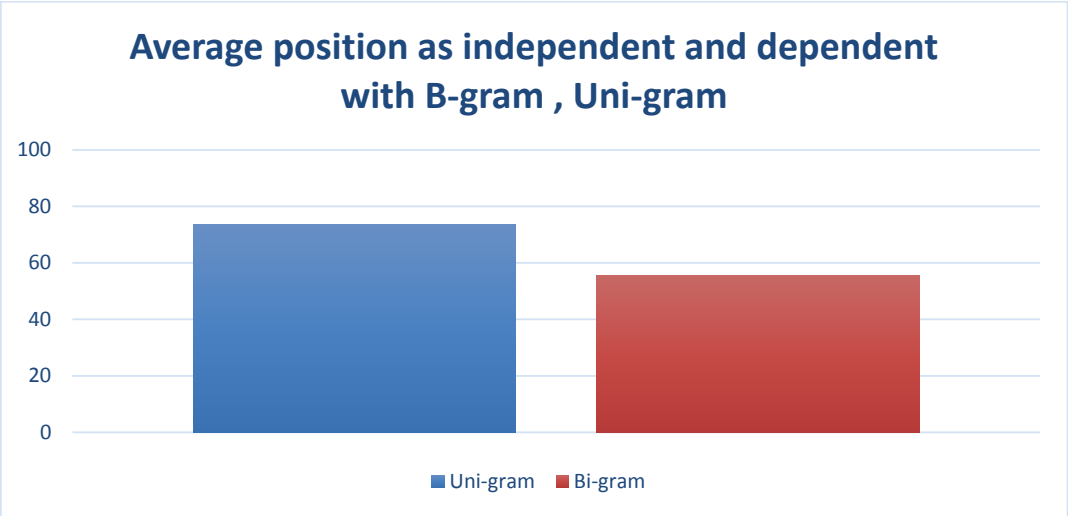

Figure 1: The overall success rate in two methods of labeling Uni-gram and Bi-gram with the Viterbi algorithm

So as you can see in Figure 1. The success rate of the method left column Uni-gram and the right side of the method Bi-gram is. So the value of the moment of transition from decomposition to the combination means method Uni-gram more important than after the transition from decomposition into a combination means a method Bi-gram is.

5.2 The success rate on the role of independent

In general, there are 260 independent roles in this 73 input sentences in the method Uni-gram There are214 correctly identified cases in the method Bi-gram 149 number is correctly specified. So the success rate on the role of independent methods Uni-gram 82.3% in the method Bigram Is 57.3%. Thus, according to the results of Figure 2 show these values with better clarity.

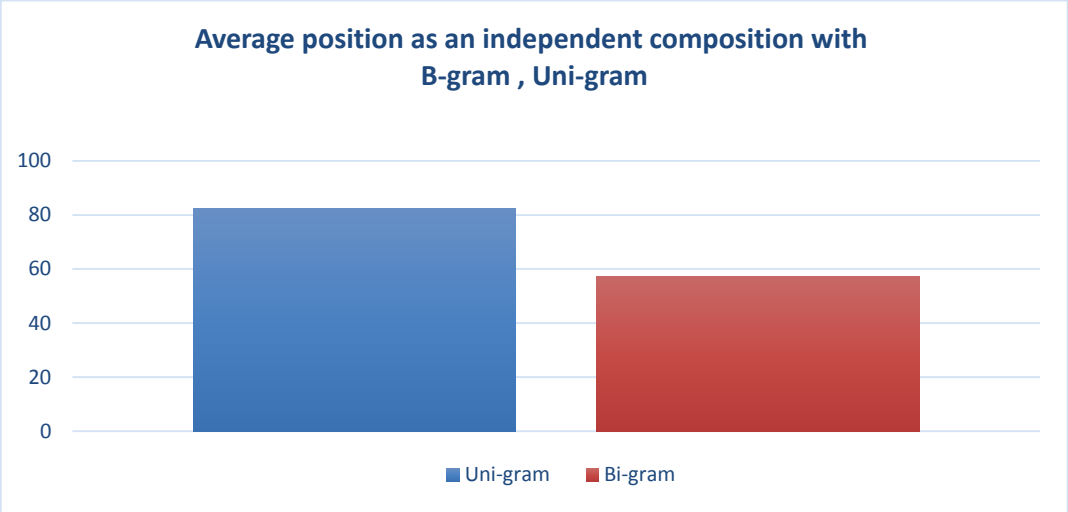

Figure 2: The success rate for each of the two independent roles

As shown in Figure 2, the left column corresponds to the method Unigram and the right side related Bi-gram which indicates the method's superiority Uni-gram than Bi-gram is.

In table 6 the success rate of each of the two tag values Uni-gram And Bigram the Viterbi algorithm is independent role.

Table 6 - The success rate of Uni-gram and Bi-gram labeling by independent roles

Bi-gram	Uni-gram	Independent roles
33.33333	66.66667	**Predicate**
20	98.33333	**Subject**
0	77.41935	**governing predicate**
100	55.55556	**Object**
0	40	**Complement**

As in Table 6 Visible only in the role of the percentage of the method's success Bi-gram Above Uni-gram Therefore, citing Table 6 The role of independent (subject, subject, Subject , complement) the moment of transition from analysis to combine more effective than the post-transition.

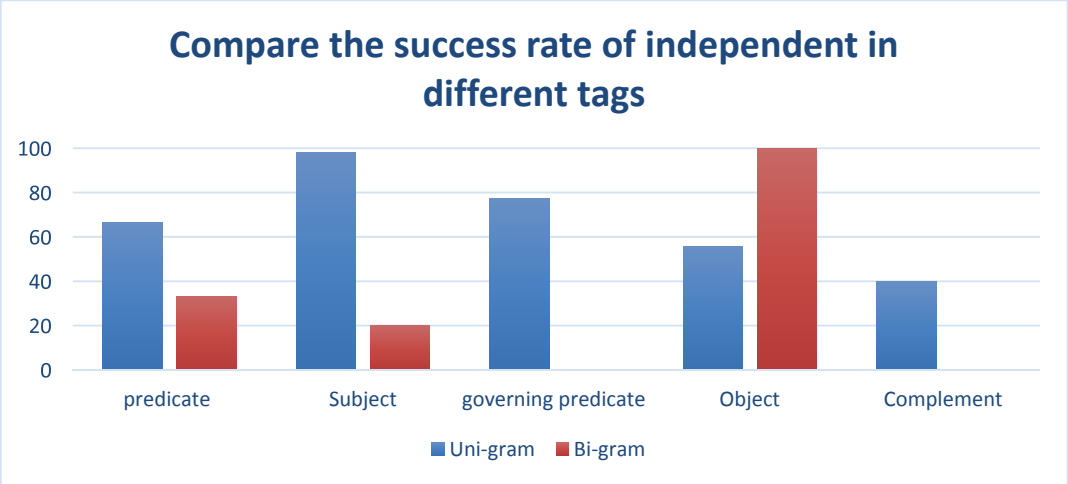

Figure 3: Percentage of successful label Uni-gram and Bi-gram in independent roles

Figure 3 shows the success rate of each role (predicate, subject, Subject, object complement) as the independent role is known, is. In Figure 3, the last pair of columns, the left column of the label Uni-gram And the right to label Bi-gram is.

5.3 Percentage of success in dependent roles

The total number of associated roles in the 73 inputs is 155, of which the number of correct outputs in the Uni - gram labeling is 91, and in the bigram labeling is 81, so the success rate of the uni-gram labeling method is:

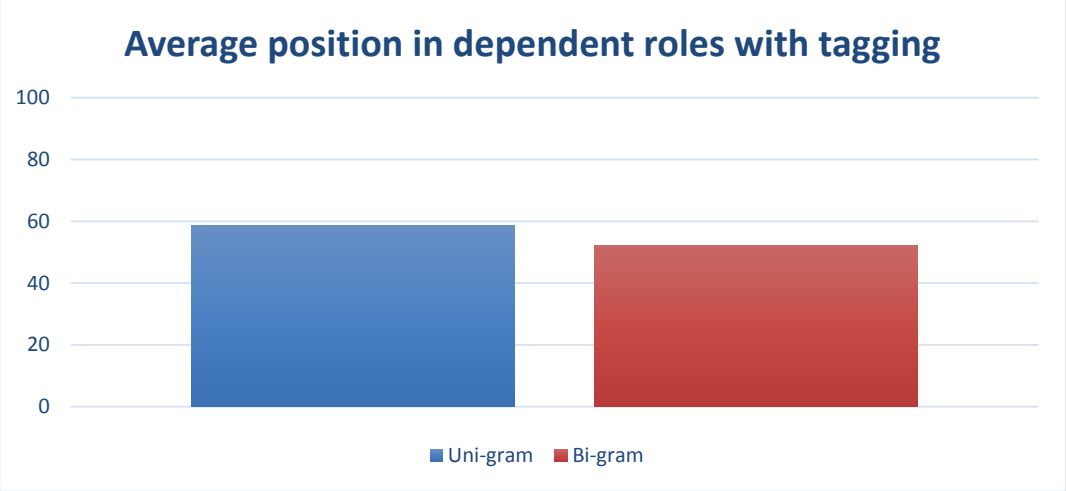

Figure 4: the success of the plan depends on the type of tag

Figure 4 percent each of the two methods for successfully portrays the role of an affiliate. In this column the left column corresponds to the method Uni-gram and the right column of the label Bi-gram is.

Table 7 - The success rate of Uni-gram and Bi-gram labeling is distinguished by dependent roles

Bi-gram	Uni-gram	depends roles
5.55	72.22	Adjective.
9.09	0	noun (Describing adjectives)
82.75	65.51	Adverb
93.87	100	Unknown
20	0	governing genitive
0	0	Genitives
0	0	Apposition
0	0	Apposition thereof
75	75	Bending
75	75	Retroactive
0	100	Exclamation
0	0	Annunciator

According to Table 7, it can be seen that the number of roles that have a success rate of zero is 6 in Uni-gram labeling, but in the Bi-gram method it is 4, while in the whole Uni-gram method is higher than the method Bi gram has succeeded. To better understand the difference in the success of the affiliated maps, two types of Uni-gram and Bi-gram labeling are shown in Figure 5.

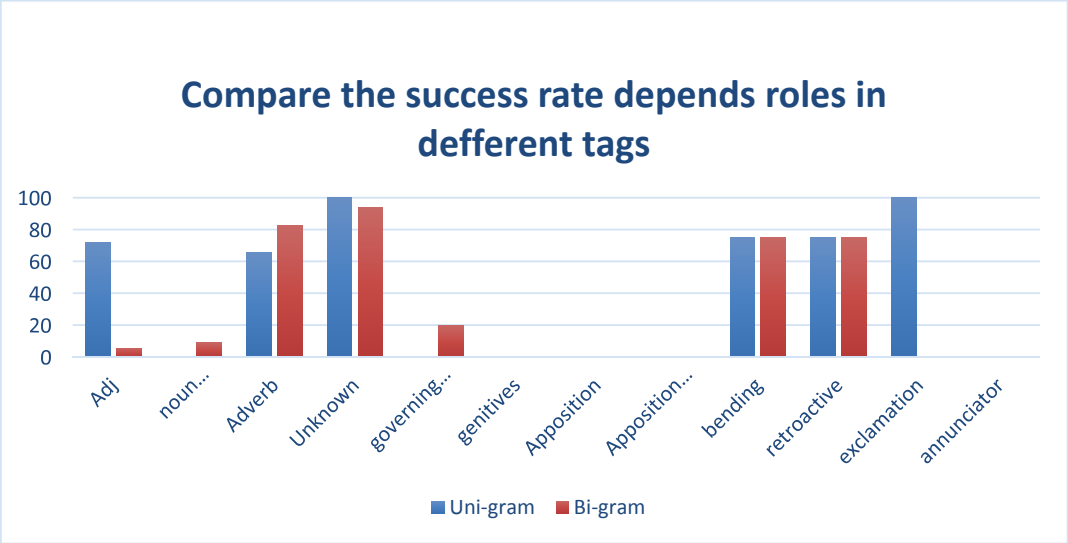

Figure 5 - Breakdown success rate two types of tagging in depends roles

In Fig. 5, each column pair represents the label that the left column corresponds to the Uni-gram tagging method and the right column for the Bi-gram tag. As shown in Fig. 5 Apposition, genitives, Apposition thereof, annunciator roles are equal to zero in both types of labeling, and in addition to the fact that the statistical values extracted from the teachings of the sentences in this system are influential, the scarcity of these roles in the language sentences Persian is also influential.

6. Conclusion

Clearly, the Uni-gram labeling method is more successful than the Bigram labeling method. Since the Uni-gram labeling method uses two items of weighting words with the Uni-gram tag and the word that each word is parsed in a combination of words, On the other hand, due to the fact that the use of the Uni-gram tagging method in the associated terms has had weaker results relative to Uni-gram, it can be concluded that labeling the Uni-gram letters of the sentences of the sentence sentence and also what type of word in the analysis is what kind of word in The compound at the same point of decomposition is more important than the bi-gram tagging of the words of the sentences of the input sentence and that instead of any of the i words in the decomposition of what kind of combinatorial in the i + 1 position is located. Therefore, we can suggest that:

Use these two labels at once.

As well as research to examine the magnitude of the effect of the letters of each word, the input sentence, and the magnitude of the outcome of each role of the analysis with the labeling of Uni-gram and Bi-gram, which can be ignored from items that have a lower success rate, and computational burden Heavy statistical methods are also mitigated.

Reference

[1] Peykar, H. Motameni and M. Aboutalebi, "Application of fuzzy identification method depends on the synthesis of the Persian language," in *Conference iran datamining Iran*, Tehran, 2014.

[2] M. Assi and M. Haji ABDOLHOSSEINI, "Grammatical Tagging of Persian Corpus," *Internatinal Journal of Corpus Linguistics.*, vol. 5, no. 1, pp. 69-82, 2000.

[3] س. راحتی قوچانی, ع. عظیمی زاده و م. م. عرب, ‘‘برچسب‌زن دستوری واژه‌گان فارسی به کمک مدل پنهان مارکوف,’’ در *سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران*, جزیره کیش, ۱۳۸۶.

[4] عاصی, ‘‘پردازش دستوري زبان فارسي با رايانه,’’ *ویژه نامه دستور فرهنگستان زبان و ادب فارسی*, pp. ۲۹-۵۱, ۱۳۸۳.

[5] M. Shamsfard and H. Fadaee, "A Hybrid Morphology-Based POS Tagger," in *LREC 2008*, Marrakech, 2008.

[6] M. Bijankhan, J. Sheykhzadegan, M. Bahrani and M. Ghayoomi, "Lessons from Building a Persian Written Corpus: Peykare," *Language Resources and Evaluation*, vol. 45, pp. 143-164, 2011.

[7] Sagot and G. Walther, "A Morphological Lexicon for the Persian Language," in *LREC 2010*, Valletta, Malta, 2010.

[8] K. Megerdoomian, "Unification-Based Persian Morphology," in *CICLing 2000*, Mexico, 2000.

[9] K. Megerdoomian, "Finite-State Morphological Analysis of Persian," in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.*, Stroudsburg, 2004.

[10] W. Amtrup, K. Megerdoomian and R. Zajac, "Rapid Development of Translation Tools," in *Proceedings of Machine Translation Summit VII*, Singapore, 1999.

[11] A. J. W., M. R. H., M. K. and Z. R., "Persian-English Machine Translation: An Overview of the Shiraz Project," Memoranda in computer and cognitive science, 2000.

[12] Riazati, B. Comp and G. Dip, "Computational Analysis of Persian Morphology," Department of Computer Science RMIT, 1997.

[13] S. Jabbari and B. Allison, "Persian Part Of Speech Tagging," *CAASL-2 Proceedings Stanford University*, pp. 67-74, 2007.

[14] A. Peykar, H. Motameni and M. Aboutalebi, "identify the independent synthesis, in Farsi sentences with fuzzy systems," in *the first national conference of meta-heuristic algorithms and applications in science and engineering*, Fereidoon'kenar, 2014.

[15] Peykar, H. Motameni and M. Aboutalebi, "Comparison of fuzzy and hidden Markov model to identify independent of synthesis words in Persian," in *Conference iran datamining Iran*, Tehran, 2014.

[16] Peykar, software analysis and synthesis of Pars Persian sentences processor / Pars Process, Gorgan: Golestan University Faculty of Science, 2012.

[17] F. Oroumchian, A. Aleahmad, P. Hakimian and F. Mahdikhani, "N-GRAM and local context analysis for persian text retrieval," in *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, Sharjah, 2007.

[18] ه. فیلی و غ. قاسم ثاني, ‘‘استفاده از گرامر اتصال- درخت براي ترجمه ماشيني انگليسي به فارسي,’’ *نشريه علمي پژوهشي انجمن کامپیوتر ایران*, جلد ۳, شماره ۱, pp. ۲۳-۳۱, ۱۳۸۴.

[19] H. Wettig, S. Hiltunen and R. Yangarber, "HIDDEN MARKOV MODELS FOR INDUCTION OF MORPHOLOGICAL STRUCTURE OF NATURAL LANGUAGE," Department of Computer Science,University of Helsinki, Finland, Helsinki, 2010.

[20] M. Mohseni and B. Minaei-bidgoli, "A Persian Part-Of-Speech Tagger Based on Morphological Analysis," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

[21] M. Bahrani, H. Sameti, N. Hafezi and S. Momtazi, "A New Word Clustering Method for Building N-Gram Language Models in Continuous Speech Recognition Systems," in *New Frontiers in Applied Artificial Intelligence, 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE*, Wroclaw, Poland, 2008.

[22] م. ح. معطر, "مدل مخفي ماركوف و الگوريتمهای آموزش," دانشگاه صنعتي اميركبير (پلي تكنيک تهران) دانشكده مهندسي كامپيوتر و فناوري اطلاعات, تهران, ۱۳۸۵.

[23] Geeraerts and H. Cuyckens, The Oxford Hand Book Of Cognitive Linguistics, Oxford: Oxford University press, 2007.

[24] M. Haspelmath, Understanding MORPHOLOGY, London: ARNOLD (A member of the Hodder Headline Group, 2002.

[25] د. ح. انوری, دستور زبان فارسی, اصفهان: چاپ اول اصفهان, ۱۳۹۰.

[26] گرگان: دانشگاه گلستان- دانشکده علوم ,Pars Processع. پیکر, نرم افزار تجزیه و ترکیب جملات فارسی پارس پرداز/ پایه, ۱۳۹۰.

[27] طیب زاده, "وابسته هاي فعل در زبان فارسي بر اساس نظريه وابستگي," دستور, جلد ۱, شماره ۱, pp. ۲۸-۱۳, ۲۰۰۵.

[28] غ. كاشف, دستور زبان فارسی, اسلامبول مطبعه الشمس, ۱۳۲۸.

[29] A. Peykar, H. Motameni and M. Aboutalebi, "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.