Morteza B. Hasan Alizadeh[1], Seyyedi Amin H. Seyyed[2]
Department of Computer of Maku Branch I.A.U, Maku, Iran
[1]xanaraz.mh@gmail.com,[2]amseyyedi@gmail.com
DOI: 10.25045/jpit.v10.i1.06

AUTO STEMMING OF AZERBAIJANI LANGUAGE

One of important features in natural language processing is to find the root of a word. Stemming means to remove prefixes, suffixes, and infixes for finding the root of the word. Its aims are about to information retrieval, exploring text, machine for translation, and word look up based on its root. Stemming increases document retrieval by 10-50% in most of international languages, it also compresses the size of web-based table indexes documents up to 50%. In this paper, by analyzing stemming approaches, using structural methods, and deterministic finite automaton machine, applying 274 existing prefixes in language (linkage), a stemming system for Azerbaijani language is generated. Experimental result demonstrates that the proposed algorithm performs more than 97% accuracy.

Keywords*: Natural language processing, Stemming, information retrieval, machine translation, Azerbaijani language.*

1. Introduction

Optimal access to required information, extracting information from texts and making connection between them is one of important factors in text processing area. Generating required devices in order to access to these aims, are essential. One of ways which can Improve the capabilities of information retrieval systems, machine translation and summarization is using of word stemming. Because different derivations of a word will be change to its root form, consequently, the search will be based on word root [1]. Stemming is a linguistic process which tries to clarify the root and the base of every word in text. Nowadays, different and various approaches of stemming in scientific literatures are introduced. Index approaches, statistical or structural or based on structure [1–2].

Index approach is the simplest way to auto stemming. So that each word and its root has been saved in a structural data. Its weakness is that requires lots of hard space. For every word, the table must be updated manually [2]. Statistical approach is stemming by statistical approaches. It stems words by limited information of a language. This approach does not require any special knowledge, does not depend on language formation [3]. Structural approach relates to word formation in language. In this approach the word formation rules have been used. These algorithm finds the roots of words by analyzing formation rules [2]. For example, in English language after removing "ING" from the end of a word, we can find its root. For example: working changes to work. This methods, because of clarifying ways of word creation provides a way of best mechanization. Porter stemming is one of algorithms of this method which is generated for English language, this stemming is based on removing the biggest fix. This does not provide method for prefixes and uses a couple of rules to identify the root [4]. But the problem occurs when to complete words with big differences create a compound word. For example: the word "goes" go +s can be mean honey bees or go + es means a form verb. One of advantages of this method has 400 line of codes which are written in BCPL. It is fast, easy, and efficient and has a capability of reflection. In Persian language algorithm of Kazem Tagavi is used, which is trying to find out especial suffixes and passes a couple of clear suffix creation which shows that the accuracy of this stemming tool on 26913 is 47.08% [4]. The time complex is these stemming tools are up and on the other hand. The level of success in them according to their limited rules are down [5]. Worth mentioning that most of providing method on text processing depends on target language, so providing general solutions which can be applied to most languages does not seem to be easy. So that because of importance of search and web translation, generating a logical framework for Azerbaijani language seems to be important. This article presents an automatic stemming algorithm with high accuracy, with taking account the rules of Azerbaijani language.

The rest of this article organized as follows. In the section one, the Azerbaijani language has been analyzed. In the section two, the existence linkage of Azerbaijani language has been analyzed. In the section three, the proposal algorithm has been described in details. Section four presents experimental results and analysis of proposed algorithm and conclusion is presented in section five.

2. Azerbaijani Language

Current language of Azerbaijani people from historical connection and origin point, is one branches of Turkish language from Altaic language. Its main place in different historical periods: at first, it was in central Asian vast flats and yenichey beaches. Then, in different times and centuries because of Turkish tribe's immigration and habitation the language spreaded. its dominant was increased to Soria and Mesopotamia and in other side up to Balkan peninsula. it was the talking language of different nations and tribes. [6]. Linguistics processed language from three independent parts based on their structure: single syllable languages, analytical languages, and agglutinative languages [6]. The Azerbaijani language is an agglutinative language. In agglutinative language the root of a word might be multiple syllables. In such languages, word root connection with some of words which are not a root, derives different meaning for the root that might be so different from root meaning. So the extent of words in such languages becomes more extensive. In such languages the root does not change, and can be easily extracted. Linkages in such languages can be prefixes or suffixes. For example, in Azerbaijani: "göz" (eye), "gözəl" (beautiful), "gözəlim" (my fair), "gözəlimin" (my fair's property), "gözəliminki" (my fair's properties), "gözəliminkilər" (once my fair have), "gözəliminkilərdən" (these are of my fair properties) and "gözəliminkilərdəndi" (these are for my sweety). The root is that comes at the beginning of the word. Because there is no prefix in Azerbaijani, and word are created by adding linkages (that's good not to say suffixes) to the end of the root form. Then, in order to find the root we have to look up at the beginning of a word. Knowing that the root of more than 90% of Azerbaijani words are single syllable, makes our so much easier so that we can follow the word up to find a single syllable word[6].

Linkage. Although prefixes are being used in Azerbaijani language, but in most cases they use as negative prefixes and There are not frequent. So that in this paper our main focus is on suffixes and their structure. Suffix is the last added part to the root. In Azerbaijani language it may be possible that multiple adjuncts exist at the end of a root. For example: look at adjuncts at the end of a word "dağ" (Mountain), "dağda" (In the Mountain), "dağdakı" (At the mountain), "dağdakılar" (At the mountains), "dağdakılardan" (they stay at the mountain), "dağdakılardanımış" (they stay at the mountain), "dağdakılardanımışlar" (they stay at the mountains). As notice that in Azerbaijani language there isn't a single adjunctive at the end of a stem which can be called suffix, but it may be possible that more than ten adjunctive placed at the end of a root. So for such reason, we will call any kind of adjunctive which comes at the end of a linkage.

3. Structural approach

In this approach, according to some of predetermined rules, in order to find the stem some links will be deleted. These kind of stemming's depend on language and the expertness of designer. These types stemming finds the root by removing prefixes of suffixes, algorithm of these stemming tools gathered from a couple of rules that by finding the first possible and applicable rule uses that rule. This process goes until there is no possible alphabet to remove. Because of logical structure of Azerbaijani language, we use the proposed approach [2].

These systems in order to find roots use the omission of links. In these systems after entering every word in to the stemming tool by using existing rules, removable links will be recognized and will be removed. These systems also use a minimum type of possibilities to find the minimum length of root's alphabets [7]. These algorithm finds word's stem by using ways of word creations. These methods because of word's creation's clear structure, provide way of mechanical use [5].

3.1. Language linkages

Number of roots in every language are limited to its progress, and are not enough to call every things movement's modes and its attributes. Therefore, creating words and derivations based on existing stems in every language seems to be inevitable. This process takes place in every language based on its informal structural system and in relational languages, this task is beholding to linkages. In Azerbaijani language, linkages are divided in to two multiform and inflectional and descriptive linkages [8].

3.1.1. Single form and multiform linkages

Stems are appeared on three type:

1) Single form: one of important single form linkage is "kən" like "qaçarkən" (when fled) [6].

2) Double form: we can point out a ouple of such linkages one of them is "ma / mə" for example: "dolma" (Full of), "qovurma" (broil) [6].

3) Forth mode linkages: "–çi/–çı–/çü/–çu" "balıqçı" (fisherman), "dilənçi" (Beggar), "yolçu" (passerby), "sözçü" (speaker) [6].

3.1.2. Discriptive and linkages

As notice from the above mentioned linkages, that some of them do not change the nature and the spirit of the word, like pluralization linkage of "lar". Words like "ağalar"," bacılar" (Gentlemen's and sisters) do not separated from their main root and keep their own form "ağa"," bacı" (gentleman and sister). We call such linkages as inflectional linkages. The transient and level of coverage of such linkages are palmate. Some of inflectional linkages in Azerbaijani language are as follow: plural linkage, possessive linkages, conditional linkages, interrogational linkage, form, time, and personal linkages. There are other linkages beside such linkages, which change the nature and spirit of the root and become nature of the new word. Like "ar" in "açar" (opening, key) although it creates a new word which relates to the root "açmaq" (open) but it has a new and independent identity from the root. These kind of linkages are discriptive. These linkages in opposite of inflectional linkages, are not transient and permanent, and stay forever beside the new word but in comparison to inflectional linkages have limited coverage [8]. The following examples clarify my explanation.

Possessive linkages: "başım" (my head), "başı" (his or her head), Free conditional linkage: "dil" (Language), tongue extended linkage: "dilin" (the language), effective: "dilə" (a language), dependent: "dili" (his language), specific: "dildə" (in the language), departed: "dildən" (from language), question linkages: "neçə?" (how much), "hankı?" (which one), time linkages: "indilik" (Now), "hələlik" (right now), identifier linkages: "mənlik" (Mine), "sənlik" (yours) manner linkages: "yazam" (I write), "yazasan"(you write).

3.2. Stemming rules in Azerbaijani language

In order to find stemming rules, we have to pay attention to grammar and structural form of this language. In the following, three groups of nouns, verbs and pronouns are being analyzed in order to identify the rules of word formation.

3.2.1. Noun

Noun is one of important parts of speech creation, and points to a couple of descriptive words that used to call living, lifeless, events and meanings. There are couple of noun linkages in Azerbaijani language, which comes at the end of a noun or a verb, and make a new noun.

1) Nominal noun creation linkages: these kinds of linkages are linkage which added to the end of a root (whether noun or verb) and makes a new nouns or verbs. the following are examples:

1- ac/əc 2– arı,əri 3 -aq,ək 4- alaq, ələk 5 -av,ov 6- it,ut 7- iz,uz 8- iş,uş 9– il,ul 10 - man, mən [9].

2) Verbal noun creational linkages: these kind of linkages comes at the end of verbal root and changes it to the noun. Examples are follow: 1- a,ə 2- ar,ir 3 -araq, ərək 4- arı, əri 5- ası, əsi 6- aq, ək 7- aqan, əgən 8- alaq, ələk 9- am, um [9].

3.2.2. Verb

Verbal changes in Azerbaijani language in order to show creatures movements and actions occurs by adding inflectional and descriptive linkages to the root and there are no changes in the root. In this language, there are also linkages for verbs. Verb creator nominal and verbal verb creators [8-9].

1) Nominal verb creator linkages: there are linkages that comes at the end of nominal (noun) roots and creates verbs. So that after such linkages, there will be negative, unknown forms and pronouns, verbal linkage places at the middle of a word (verb). We can call verb maker and its linkages as verbal basement. Verbal linkages usually show the form of verb. There are:

1 -a,ə 2 -at,ət 3 -ar,ər 4 -aş,əş 5 -al,əl 6 -an,ən 7 -ay,u 8 -ir,u 9 -irqa,irgə [9].

2) Verb-Creator verbal linkages: there are linkages which are added at the end of a verbal root and by making changes in the meaning of root creates new verbs with new form and meaning. The examples are: 1 -a,ər 2 -ar,ər 3 -arlama,ərləmə 4 -alama,ələmə 5 -it,ut 6 -i,u 7 -ir,u 8-irga,irgə 9- iz,u [9].

3.2.3. Pronoun

Pronouns are descriptive words that shows people things by pointing or allusion. In Azerbaijani language pronouns are as follow: subjective pronouns, indication pronouns, and interrogational pronouns. In stemming subjective pronouns play the most important rule, because the rest pronouns are repeated and exist in nominal linkages. If some of subjective pronouns are repeated, they will be mentioned. Subjective pronouns which clarifies, the subject is being connected to the end of a verb and is like:

First singular person: subjective pronoun is m which added to the end of a verb. “Yazdım” (I wrote), “yazacağam” (I will be write).

Second singular person: in perfect past and conditional form the verb and its symbol is “n”. “qaldın” (you have stayed), “gördün” (did you see). Conditional form: “alsan” (unless you buy), “gəlsən”(if you come).

Third singular person: It doesn’t any form and usually its form is on its formlessness.

First plural person: its form based on sounds relation is: “q/k” “alaq” (Lets get it), “vurduq” (we hit), “Gördük” (we saw), “görsək” (if we saw), “görmüşdük” (we have seen),

Second plural person: whenever the “Z” letter added to the end of second single pronoun it will change to second plural pronoun. “Yazdın” (did you write)+ z =”yazdınız” (You wrote), “qazırsan” (You are digging) +z= “qazırsanız” (you are digging).

Third plural pronoun: its linkages are “lar/lər” which is added to end of third single pronoun. “almışdılar” (They bought), “görmüşdülər” (They have seen), “baxacaqıdılar” (they will look), “Gələcəkidilər” (They will be coming)

4. The proposed auto stemming algorithm

Auto stemming is one of important and basic part of natural language processing. Most of proposed methods in text processing depend on target language, so proposing general solutions language which can be useful in every section seem to be impossible. In this article an auto stemming algorithm with high accuracy for Azerbaijani language has been proposed. The proposed method has a great influence on indexing. The proposed algorithm started the auto stemming of words and texts in Azerbaijani language in two phases. The general diagram block of the proposed algorithm is shown in figure 1.

4.1. First phase of proposed algorithm

In the first phase the inflectional rules of words and linkages in Azerbaijani language has been analyzed and has been interrogated. The inflectional structures is being processed based on structural approach, and is being divided in different groups (fourteen groups). In order to remove linkages, they will be recognized.

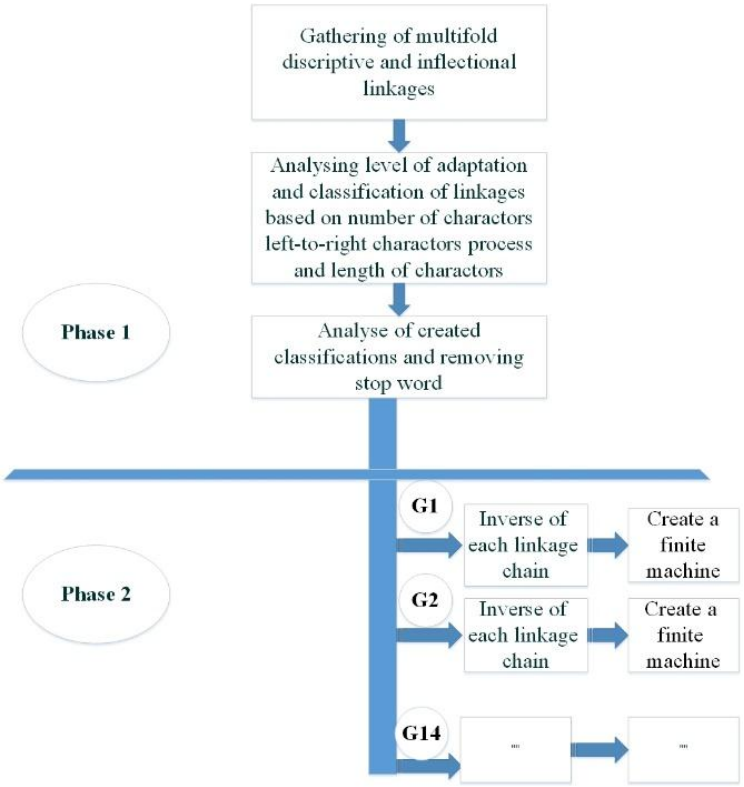

Fig.1. Block diagram of the proposed algorithm.

In this phase, linkages are classified based on their last alphabet and their first, second and number of characters. In order to prevent mistakes in other episodes of stemming because of existing similar words in Azerbaijani language, the length of characters is being processed. For example: " əl" is a word and also is a linkage. These words and their similar cases should not be deleted unexpectedly. If a linkage does not on its class, because of diversity in words, it will place on other classes, in a way that it doesn't make problem in word process. In words like: gözəl (beautiful), "keçəl" (bald), özəl" (private), and genəl"(wider)," əl" (Hand) acts as a linkage. In other words, it is a suffix and most be deleted in stemming. So, the length of line in order to improve accuracy and functioning of proposed algorithm is important. The length of selected line at least have to be equal with length of smallest existing word. In table 1 a view of classification has been shown, that in the next phase for each of them deterministic finite automaton tool should be prepared.

Table 1

Third group, linkages that end to t

T linkages	**Inverse linkages**	**Example**
İT	Tİ	ərit (əritmək)
UT	TU	Unut\qurut
AT	TA	bayat

4.2. The second phase of proposed algorithm

The second phase, which is phase of removing linkages. According to verb declension, Azerbaijani language is one of logical languages [2]. The proposed algorithm in order to expecting or rejecting auto stemming creates deterministic finite automaton for each of the groups , at the end, all of 14 created machine gathered in the finite automaton. In this part according to classifications in the first phase, in order to recognize and remove linkages deterministic finite automaton is required. Fourteen different machine with machine number fifteen, which is the collection of machines, should be prepared. Lines that are accepted by the machine, is recognized as a linkage and is deleted. The

remaining line is reexamined by the machines, to remove any of linkages. Because there is sequence of linkages in Azerbaijani language and it is possible in order to find the root of the word several linkages remove simultaneously. Whenever the machine stops in a non-final state, the root of word is extracted. Every deterministic finite automaton Ma describe as: that is,

$$M_a = (Q, \sum, \sigma, S, F)$$

Q _indicates the number of machine's states. In every linkage (table) number of condition is bigger or equal with number of unrepeated existing characters within linkage.

$\sum$ _Azerbaijani language alphabets that contains 32 character's which is from A-Z.

σ _is actuation function that its principle is like $Q \times \sum \to Q$ which shows the way of transforming from one state to another one. so that this transformation will be completed by reversing existing linkages in every table. It is describe completely by an example.

S _ indicates the start state of machines.

F _ is the collection of final states of machine. Final states are the last existing characters in 14 groups. If stop in each state indicates the recognition of linkage, which should be removed to find the root.

For example imagine the T linkage group, which its deterministic finite has been shown in figure 1.

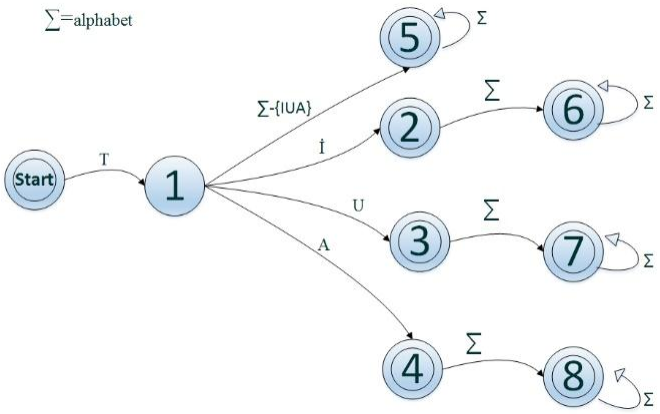

Fig.2. A definite finite machine for the T group's transcription.

5. Evaluation of the efficiency of the proposed algorithm

In this section, some experiments are carried out to assess the efficiency of the proposed scheme. The proposed method has been simulated using C# Programming tool on windows 10 platform.

Analyzing of stemming algorithms, depending on their functioning can be different. We can compare stemming tools based on their functioning and accuracy. As mentioned before, the goal of this paper is to propose an algorithm in order to use in machine translation. The main concentration is on the accuracy results in stemming [10]. In general, there are two conventional methods for analyzing accuracy of auto stemming algorithms:

5.1. First assessment method

Number of words has been entered to stemming algorithm and the results will be evaluated. In this method the generated roots by the system has been compared with correct roots of words and the functionality of system is being analyzed. The table two shows a couple of words that generated by proposed algorithm [7].

Table 2

Examples of word stemming

Entry word	main root	root derived from the algorithm
düşür	düş	*düş*
darıxdırıcı	dar	*dar*
qorxmuşdur	qorx	*qorx*
əzənlər	əz	*əz*
dinləməlisən	din	*din*
çalğı	çal	*çal*
evcik	ev	*ev*
boyunduruq	boy	*boy*
qurtum	qurt	*qurt*

5.2. Second assessment method

Some documents with different subjects has been selected and after extracting there words and removing repeated words they will be entered to the algorithm. Results of algorithm stemming will be compared with correct stem and the accuracy of algorithm will be evaluated [7]. Evaluating system functionality, for evaluating of stemming functionality in second method of evaluation, different texts in civil subject with 707 words [11], sports 884 [12], economy 480 [13] gathered from Azerbaijan websites. The results shown in table 3. Shows %97 average of accuracy in stemming tool.

Table 3

Result of stemming on 3 documents

The correct answer rate	Wrong answers	Correct answers	Total number of words	NO
97.17	20	687	707	1
96.94	27	857	884	2
96.80	15	455	470	3

6. Conclusion

Analyzing of word stemming algorithm indicates that in order to find the real root of words removing prefixes and suffixes (linkages) is the best way. Azerbaijani language is a relational language and the best and the most important approach tool of stemming is a structural approach. In this approach the rule of contextual meaning in the sentence is not important because contextual analysis of sentence increases the time of stemming. Additionally, its effect on result is very low. Results of experiment shows that the proposed approach has great deal of effect on decreasing indexing. Experimental analysis has been done on different texts by 2 methods. One of them is entering a couple of word and analyzing their results and the was analyzing texts with different subjects which implies, 97 percent accuracy. Additionally, according to repeated tests and analyzing of different texts in databases, there are words that is not originally Azerbaijani and in the text processing and entering text the tool can recognize them. Because existing of unknown words make stemming process harder, so that we can decrease them logically by creating database.

References

1. Ehsan N., Fili H. Analysing Effects of Stemming in Information Retrieval in Persian Language, Pardis Electronic and IT College Tehran Scientific College, 2011(in persian).
2. Taghi Zadeh H., Sadrodini M.H., Deyanati M.H., Rasekh A.H. A Persian Stemming based on Structure by using methodical Phrases / The Eleventh Conference of Iran's intelligent Systems, Kharazmi University, February 29–30, 2012 (in persian).
3. Momeni P., Marjab F., Marjab S. Challenges in Persian Texts Stemming in information retrieval Systems / The third Conference of Progressing Scientifical Engineering,- Tankarisheh Ayandegan institution of High Learning, 2016 (in persian).
4. Arhaft S., Shadgar B., Ashakori M. Proposing a Persian Stemming Based on paradigmatical

diversity / International Conference Research on, Science and Technology Engineering, Istanbul institute of Idiological Manager of Vira Capital, http://www.civilica.com/Paper-RSTCONF01-RSTCONF01_454.html, 2015 (in persian).
5. GhasemSani Gh., Hesami .. A Stemming Algorithm for Farsi Language / In Proceeding of 11 International CSI Computer Conference (CSICC'2006), 2006.
6. Naebi M. Azerbaijani Turkish Grammar Learning, 2008, p.80 (in persian).
7. Nojavan Agdaragh B., Rezaey M., Feyzi Derakhshi M.R. Auto Stemming of Persian Words by using useful combination of word structure rules and date bases.the Eighth International Conference of Persian literature fomentation. Iran, Zanjan Persian language fomentation assemblage, Iran. http://civilica.com/Paper-ISPL08-ISPL08_345.html, 2014 (in persian).
8. Farzaneh M.A. Nitty-gritty of Azerbaijani Turkish Language Grammar. Kaveyan,1987, p.61 (in persian).
9. Hadi A. Turkish is Art. Iran, Tabriz: Ahrar,1995, p.291 (in persian).
10. Frakes W.B. Stemming algorithms, in Information Retrieval Data Structures and Algorithms, Ed. Prentice-Hall, 1992, pp.131–160.
11. www.icherisheher.gov.az/qanunlar,154/lang,az/
12. www.maliyye.gov.az/sites/default/files/store/13/AASMN_qerarlar_1.doc
13. http://facemark.az/files/telebe/628941319836404291011.doc

Morteza Hasan Alizadə B.[1], Səyyadi Səyyad Amin H.[2]
Kompüter elmləri bölməsi, Maku şöbəsi I.A.U, Maku, İran
[1]xanaraz.mh@gmail.com,[2]amseyyedi@gmail.com

Azərbaycan dilinin avtomatik söz köklərini tapmaq

Təbii dil emalında vacib xüsusiyyətlərdən biri sözün kökünü tapmaqdır. Stemming, şəkilçilərin ayrılması ilə sözün kökünün tapılması deməkdir. Onun məqsədi sözün kökünə əsasən informasiya axtarışı, mətnin öyrənilməsi, maşın tərcüməsi və sözün axtarılması ilə əlaqədardır. Stemming, bir çox beynəlxalq dillərdə olan sənədlərin axtarışını 10-50% artırır, həmçinin web-cədvəl indeksli sənədlərin ölçüsünü 50% -ə qədər sıxır. Bu məqalədə kökün tapılmasına olan yanaşmalar analiz olunmuş, struktur metodlardan və determik sonlu avtomatlardan istifadə edilərək Azərbaycan dilində mövcud 274 ön şəkilçinin tətbiqi ilə sözün kökünü avtomatik təyin edən sistem yaradılmışdır. Eksperimentlər nəticəsində təklif olunan alqoritmin 97%-dən çox dəqiqliyə malik olduğu göstərilmişdir.

***Açar sözlər**: Təbii dilin işlənməsi, stemming, bilgi alma, maşın tərcümə, Azərbaycan dili.*

Мортеза Хасан Ализаде[1], Сейеди Сейед Амин[2]
Кафедра компьютерных систем филиала Маку, Исламский Азиатский университет, Маку, Иран
[1]xanaraz.mh@gmail.com, [2]amseyyedi@gmail.com

Автоматическое наблюдение за происхождением корня слов в азербайджанском языке

Одной из важных особенностей обработки естественного языка является поиск корня слова. Чтобы найти корень слова, необходимо удалить префиксы, суффиксы и инфиксы. Цель наблюдения связана с подбором информации, изучением текста, машинного перевода и поиском слов на основе его корня. Нахождение корня увеличивает поиск документов на 10−50% в большинстве международных языков. Он также сжимает размер веб-табличных документов с индексами до 50%. В этой статье предложено автоматическое наблюдение за происхождением слов в азербайджанском языке на основе анализа подходов к поиску корня, используя структурные методы и детерминированную конечную автоматическую машину, применяя 274 существующих префикса. Экспериментальные результаты показывают, что предложенный алгоритм имеет точность более чем 97%.

***Ключевые слова**: обработка естественного языка, нахождение корня, получение информации, машинный перевод, азербайджанский язык.*