Lexicon-free stemming for Kazakh language information retrieval

Ualsher Tukeyev
Information Systems Depatment
KanNU named after al-Farabi
Almaty, Kazakhstan
ualsher.tukeyev@gmail.com

Aliya Turganbayeva
Information Systems Depatment
KanNU named after al-Farabi
Almaty, Kazakhstan
turganbaeva.aliya@bk.ru

Balzhan Abduali
Information Systems Depatment
KanNU named after al-Farabi
Almaty, Kazakhstan
balzhanabdualy@gmail.com

Diana Rakhimova
Institute of Information and Computational Technologies
Almaty, Kazakhstan
di.diva@mail.ru

Dina Amirova
Information Systems Depatment
KanNU named after al-Farabi
Almaty, Kazakhstan
amirovatdina@gmail.com

Aidana Karibayeva
Information Systems Depatment
KanNU named after al-Farabi
Almaty, Kazakhstan
a.s.karibayeva@gmail.com

***Abstract*— In paper are considered existing algorithms for automatically isolating the bases for a number of natural languages and possible ways of synthesizing a normal form of a word for the Kazakh language. In paper are described the complete system of endings of the Kazakh language. The paper proposes a new approach to constructing a lexicon-free stemming algorithm for the Kazakh language on the basis of a complete set of endings of the Kazakh language. The stemming algorithm will be used for Kazakh language information retrieval to finding specific words in the documents by stemming base of word.**

Keywords— lexicon-free, stemming, algorithm, Kazakh, language, information, retrieval.

I. INTRODUCTION

Questions of word processing of the Kazakh language are poorly developed. Firstly, this is related to the specifics of Kazakh language as a language with a complex morphology, and secondly, with not an insufficiently active research of Kazakh language in this area. However, questions of word processing in Kazakh language, in practice, are very actual. One of the actual issues is the area of information retrieval in the Kazakh language texts. In this area, an important issue is the problem of the quick search of specific words in documents. One of the ways to quickly find words is to search the basics of words among the key words of documents, which allows you to select the appropriate document as the desired one. Therefore, one of the ways to solve searching by the basics of words is the allocation of basics in the word, namely, stemming [1]. Stemming of words can be done through the use of dictionaries or without the use of dictionaries, namely, lexicon-free. This approach, initiated by Porter[2] got active development and application in the field of information retrieval.

In this paper, is proposed the approach to build algorithm of the lexicon-free stemming for Kazakh language based on the complete set of endings of Kazakh language. The completeness of the endings of the language ensures a sufficiently high level of implementation the stemming of words.

II. RELATED WORKS

In the area of development algorithms of stemming presented quite a lot of work. It should be noted, some of them quite close to this paper, firstly, on the investigated languages, and secondly, on the approach to the construction of the algorithm. Therefore, it is the Porter's work, the disadvantage of which is a fairly high percentage of errors. In the work of Segalovich[3] is proposed an efficient algorithm for the Russian language the use of the dictionary that improves the quality of stemming. In the work of A.M. Fedotova and others proposed algorithm of lemmatization for the Kazakh language, in which examined the system of the endings of the Kazakh language, which is not complete.

The stemmer is the implemented algorithm of stemming. Widespread stemmers are the following:

- Porter's stemmer (Snowball);
- Stemmer MyStem.

Porter's stemmer - the algorithm of stemming, published by Martin Porter. The algorithm does not use the bases of words, but works by consistently applying a series of rules for cutting off endings and suffixes [2, 4].

The basic idea of Porter's stemmer is that there is a limited number of form and word-building suffixes, and word stemming occurs without using any bases of bases: only a lot of existing suffixes (complex compound suffixes are divided into simple suffixes) and manually defined rules.

The advantage of the stemmer is the speed, since dictionaries and the base of bases are not used, and the disadvantage is not always an accurate selection of the stem [5].

The Stemmer MyStem was developed by Ilya Segalovich in 1998. Here the algorithm works on the basis of dictionaries. The advantages of this algorithm are that for all variants of the normal form all grammatical information is offered (synthesized for non-existent words), these data can be used later to select one normal form from the set proposed by the program. The disadvantages of this algorithm are that, in the absence of a word, it can not always cope with the task [3, 10].

In the work of A.M. Fedotova and others, formally for nouns, the following model of wordform formation is constructed [9].

Through Pi, the following types of endings (affixes) are indicated for i = 1,2,3,4:

1. P1 - the end of the plural;
2. P2 - possessive ending;
3. P3 is the case end;
4. P4 is a personal ending.

The following combinations of noun endings are possible:

1. the end of the plural + the possessive ending (P1P2);
5. the end of the plural + case end (P1P3);
6. the ending of the plural + personal ending (P1P4);
7. plural ending + possessive ending + case ending (P1P2P3);
8. plural ending + possessive ending + personal ending (P1P2P4);
9. possessive ending + case ending (P2P3);
10. possessive ending + personal ending (P2P4);
11. case ending + personal ending (P3P4).

For verbs, there are the following types of endings:

1. P1 - the end of the negation;
12. P2 - the end of time;
13. P3 is a personal ending.

The following combinations of verb endings are possible:

1. the end of time (P2);
14. end of time + personal ending (P1P3);
15. end of negation + end of time (P1P2);
16. end of negation + end of time + personal ending (P1P2P3) [6].

Based on the analysis of this work, you can see that in this algorithm the system of ending the Kazakh language is not fully covered and the order of adding suffixes to the basis is not fully defined.

On the Internet [7], the algorithm of stemming for the Kazakh language was published. This algorithm is implemented based on Porter's stemmer. The algorithm lists the end of the words of the Kazakh language. The above list of graduation covers only the smallest part of the end of the Kazakh language.

III. Complete set of Kazakh endings

The system of the endings of words of the Kazakh language consists of two classes: the endings to nominal bases (nouns, adjectives, numerals) and the endings to the verb stems (verbs, participles, gerunds, mood and voice). [11]

The system of base affixes to the nominal bases of words of the Kazakh language has four types [12]:

- affixes of the plural (denoted by K),
- possessive affixes (denoted by T),
- case affixes (denoted by C),
- personal affixes (denoted by J),
- the stem (stem) is denoted by S.

All possible variants of placement of types of affixes: from one type, from two types, from three types and from four types. The number of placements is defined by the formula: Ank = n!/(n-k)!.

Then, the number of placements will be determined as follows: A41 = 4!/(4-1)!=4, A42 = 4!/(4-2)!=12, A43 = 4!/(4-3)!=24, A44 = 4!/(4-4)!=24. Total number of possible placements 64.

In [11], by considering semantically admissible allocations, a complete set of semantically admissible types of the endings of the Kazakh language is defined.

So, for words with nominal bases, the semantically acceptable number of types of Kazakh language endings is 15. The semantically acceptable number of types of Kazakh language endings for participles is 11, for verbs is 25, for verbal participles is 1, for inclinations is equal to 6 and for pledges is 8. So, the total number of types of word endings with verb stems is 51.

Total, the total number of endings with nominal bases plus the total number of types of endings of words with verb stems will be equal to 66.

According to these types of endings, finite sets of endings are constructed for all main parts of the speech of the Kazakh language. Thus, for parts of speech with nominal bases, the number of endings is 1213 (all plural variants are taken into account), and the number of endings of parts of speech with verbal bases is: verbs - 432, parti-ciples-1582, gerunds-48, inclination-240, liens-80. Total, 3565 total of the endings. [13]

Total, the total number of endings with nominal bases plus the total number of types of endings of words with a verb stem will be equal to 74.

IV. Stemming algorithm for Kazakh language

The principle of the proposed algorithm of stemming words of the Kazakh language, based on the complete system of endings is as follows:

In the system of the endings of the Kazakh language all endings are divided into classes according to the length of the endings. At first, in the word is searched the ending of the maximum length for a given word: it is into two symbols shorter than words (assuming that the length of the base cannot less than 2). Intended ending of length L is searched in the corresponding class of endings of length L. If the ending is not in this class, then length of the intended ending is decremented by one and looked up in the corresponding class of endings, etc. as long as, there will be no ending, or word will be without end.

Designations:

L(e)max – the maximum length of endings in the system of language ending; w – the analyzed word:

e(w) – the ending of the analyzed word; L(w) – the length of the analyzed word;

L[e(w)] – the supposed length of the ending of the given word. L[e(w)]max – the maximum length of the ending of the given word.

The steps of the algorithm:

It is determined the length of the analyzed word L (w).

1. It is determined the maximum length of the ending of the analyzed word:
   $L [e (w)]^{max} = L (w) – 2.$
   Where 2 is the minimum length of the basic of word.
2. If $L (w) \leq L (e)^{max}$ (if the ending of the word w less than or equal to the maxi- mum length of ending in the system of language ending), then the presumed length of the ending of the given word L [e (w)] assign the maximum length of the ending of the analyzed word:
   $L [e (w)] = L [e (w)]^{max}$. Next, go to 5.
3. Otherwise: the presumed length of the ending of the given word L [e (w)] assign $L(e)^{max}$:
   $L [e (w)] = L (e)^{max}$.
4. Make a sample of ending e (w) of length L [e (w)] from the given word w.

5. Checking e (w) on coincidence with the ending from the list of endings length L [e (w)].

If coincide, then define the basic of the given word:

St (w) = w – e (w), i.e. from the given word stands out basic.

Otherwise

6. Reduce the presumed length of the end of the given word into unit:

L [e (w)] = L [e (w)] -1.

7. If L [e (w)] < 1, then word w without endings. Go to 9. Otherwise – go to 6

8. The end.

The proposed algorithm does not use the base database, and this ensures the speed of the algorithm.

V. EXPERIMENTAL RESULTS

Checking of algorithm work with the help of program - various texts in the Kazakh language. The volume of texts in total was about 17 000 words. The appendix of the words in these texts has been removed and the result is summarized (shown in figure 1 below).

The following table gives an overview of the correctness and the correctness of the findings on the basis of the algorithm presented in the table.

TABLE I. THE RESULT OBTAINED ON THE BASIS OF TEXTS.

Texts	**Number of words**	**Correct found stems in percentage (%)**
on economics	6800	70-80
on business	3100	72-83
on technology	4700	73-80
on industry	2400	70-77
simple literary	2000	85-90

Often, errors are also encountered here, although the key is correct. Here are some reasons why the word base is not found correctly:

- The question of disconnection of words in another word, based on a word phrase;
- The presence of one base in the same base.

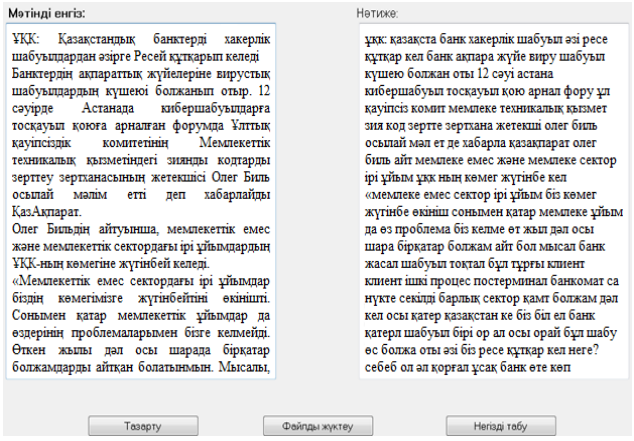

Fig 1. Result of receiving the word basics in the text

TABLE II. THE MAIN REASONS FOR FINDING THE WRONG STEMS

Errors	**Reasons**
Basics can be founded a wrong if due to the adoption of the last letters (such as -м(-m), -ым(-ym), -ім(-im))	Letters –м(-m), -ым(-ym), -ім(-im) of the basics as the possessive ending
Removal of a verbally-based word from a legitimate word, and vice versa	Undefineable that the basis of the words is a name or a verbal statement

This algorithm performs a comparison only when the length of the affixes is in the range 1-6, and in other cases the comparison not performed. This is due to the fact that often there is an error when the length of affixes is from 1 to 6.

This algorithm in general on the basis of justification of applications from the base, with the basis of the base being comparable only if it has a length between 1 and 6, and the rest of the baseline, if the additional length is greater than 6, is not fulfilled. This is because the error often occurs when an additional length is between 1 and 6.

In general, there are two types of errors in the algorithms of stemming: overstemming and under stemming. Over stemming is an error of the first kind, when inflectional words are mistakenly attributed to a single lemma. Understemming is an error of the second kind, when the morphological forms of one word are referred to different lemmas. Stemming algorithms try to minimize both of these errors, although reducing one type of error can lead to an increase in the other. [14]

This research performed and financed by the grant Project IRN AP05132950 "Development of an information-analytical search system of data in the Kazakh language", awarded to The Republican State Enterprise (RGP) on the right of economic management (PVC) “Institute of Information and Computational Technologies”.

VI. RESULTS AND FUTURE WORKS

In the paper, the work of the lemmatization and stamping algorithms is analyzed. In the wide use of algorithms are implemented for English and Russian languages. For the Kazakh language, it is necessary to improve and implement the algorithms of stemming and lemmatization, which will produce results quickly and correctly.

At present, the affix database is complemented by suffixes, since in the Kazakh language the number of suffixes is many, and they are subdivided into word-forming and form-forming suffixes. And form-forming suffixes are added to the base of affixes.

The frequency of occurrence of suffixes after the end: Suffixes can be added only after the case endings. For example, қала+да-ғы [qala+da+gy], here suffix –ғы[gy] added after case ending –да[da]. Let's show an example of how endings are added after the suffix can arrive [15].

TABLE III. SEQUENCE OF ADDING SUFFIXES AFTER THE ENDINGS.

Stem	**C Case endings**	**S Suffix**	**K Plural Endings**	**T Possessive endings**	**C Case endings**	**S Suffix**

қала (qala)	-да (-da)	-ғы (-gy)	-лар (-lar)			
қала (qala)	-да (-da)	-ғы (-gy)	-лар (-lar)	-	-да (-da)	
қала (qala)	-да (-da)	-ғы (-gy)	-лар (-lar)	-ым (-ym)		
қала (qala)	-да (-da)	-ғы (-gy)	-лар (-lar)	-ым (-ym)	-да (-da)	
қала (qala)	-да (-da)	-ғы (-gy)	-лар (-lar)	-ым (-ym)	-да (-da)	-ғы (-gy)

REFERENCES

[1] Hiemstra, D. Using language models for information retrieval. Enschede. The Netherlands. 2001.

[2] Porter Stemming for Russian language. https://eigenein.xyz/snowball/Russian stemming algorithm.

[3] Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search. 2003. 4-5 pp.

[4] M.F. Porter, “An algorithm for suffix stripping”, Program, 14(3) 1980, pp. 130−137.

[5] Willett P. The Porter stemming algorithm: then and now // Program: Electronic Library and Information Systems.2006. B. 3, Vol.40, C. 219–223.

[6] Barakhnin, V.B., Fedotov, A.M., Bakiyeva, A.M., Bakiyev, M.N., Tazhibayeva, S.Zh., Batura, T.V., Kozhemyakina, O.Yu., Tussupov, D.A., Sambetbaiyeva, M.A., Lukpanova, L.Kh. The algorithms of the generation and of the stemming of the word forms of the kazakh language // Cloud of Science. Series: Applied Research. 2017. T. 4, № 3. 434–449 pp.

[7] Basic snowball stemming algorithm for kazakh language. https://github.com/iborodikhin/stemmer-kaz.

[8] Svetlov, A.V., Komendantov. A.S. Automation of the process of obtaining linguistic information: modern possibilities. Herald of Volgograd State University. Series 2, Linguistics. 2017. T.16. №2, 39-46pp.

[9] Fedotov, A.M., Tussupov, D.A., Sambetbaiyeva, M.A., Erimbetova, A.S., Bakiyeva, A.M., Idrisova, I.A. Model of determination of normal forms of word for kazakh language. Herald of the Novosibirsk State University. Novosibirsk national research state university. 2015. №1. 107-116pp.

[10] SegalovichI. V., Maslov M. A., Russian Morphological Analysis and Synthesis With Automatic Generation of Inflection Models For Unknown Words. Moscow:Dialog, 1998, vol. 2, p.547–552.

[11] Tukeyev U., Sundetova A., Abduali B., Akhmadiyeva Z., Zhanbussunov N. (2016) Inferring of the Morphological Chunk Transfer Rules on the Base of Complete Set of Kazakh Endings. In: Nguyen N., Iliadis L., Manolopoulos Y., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2016. Lecture Notes in Computer Science, vol 9876. Springer, Cham.

[12] Bektayev, K. Big Kazakh-Russian and Russian-Kazakh dictionary. Almaty, 1995, 703pp.

[13] Tukeyev, U., Automaton models of the morphology analysis and the completeness of the endings of the kazakh language. Proceedings of the international conference “Turkic languages processing” TURKLANG-2015 September 17–19, Kazan, Tatarstan, Russia, 2015. – P. 91-100.

[14] Manning, C.D., Raghavan, P., Schutze, H. Introduction to information retrieval. Cambridge University Press. Publishing house – Williams. Moscow. 2011. 528pp.

[15] Directory of short suffixes and endings, http://u-s.kz/publ/3913-ysasha-zhrnatar-men-zhalaular-anytamalyy.html