МРНТИ 20.23.19
УДК 004.021

DOI: https://doi.org/10.51889/2021-1.1728-7901.24

STEMMING OF KAZAKH LANGUAGE

Bogdanchikov A.V. [1], Baimuratov O.A. [1] , Ayazbayev D.A. [1]*

[1]Suleyman Demirel University, Kaskelen, Kazakhstan
⁎e-mail: dauren.ayazbayev@sdu.edu.kz

Abstract

Nowadays natural language processing is widely used. For instance, it can be used to translate text, in search engines systems, text topic identification. Such applications require preprocessing of text. It should be done, because preprocessing of text can influence on system accuracy. Text preprocessing can be done by several ways. One approach is identifying root of word. Advantage of identifying root of word is that it can save memory of computer, because repeated roots will be saved one time. This paper describes stemming systems, which can identify root of word. In literature review part authors reviewed to stemming algorithms, which can identify roots of words of Russian, Uzbek, Turkish languages. Then authors proposed stemming system, which can identify root of word of Kazakh language. In current paper authors describe how their system works. To test the system words from various parts of speech were entered. Proposed system can identify roots of noun, verb, adjective, numeral words. The system response can be seen in table 1. Pictures below show what kinds of suffixes, endings can be concatenated with root of word of Kazakh language. However not all combinations are shown in pictures. In conclusion part advices for how to develop stemming system are written.

Keywords: stemming, morphology, suffix, ending, parts of speech, algorithms.

Аңдатпа

А.В. Богданчиков[1], О.А. Баймуратов[1], Д.А. Аязбаев[1]
[1]Сулейман Демирель атындағы университет, Қаскелең қ., Қазақстан

ҚАЗАҚ ТІЛІНІҢ СӨЗІНІҢ ТҮБІРІН АНЫҚТАЙТЫН ЖҮЙЕ

Бүгінгі таңда табиғи тілді өңдеу кең пайдалынылады. Мысалы: бір мәтінді басқа тілге аудару үшін, іздеу жүйелерінде, мәтіннің тақырыбын анықтағанда қолданады. Бұндай қолданыстардың алғашқы сатыларының бірі – мәтінді өңдеу. Өйткені ол жүйенің дәлдігіне әсер ете алады. Мәтінді өңдеудің әртүрлі әдістер бар. Мәтінді өңдеудің бір әдісі – сөздің түбірін анықтау. Сөздің түбірін анықтау арқылы компьютердің жадысын үнемдеуге болады. Өйткені қайталанатын түбірлер бір рет қана сақталынады. Бұл мақалада сөздің түбірін табатын жүйелер туралы талқыланады. Авторлар әдебиетке шолу бөлімінде орыс, өзбек, түрік тілдерінің сөздерінің түбірін анықтайтын алгоритміне шолу жасай отырып, қазақ тілінің сөздерінің түбірін анықтайтын жүйеге ұсыныс береді. Мақалада авторлардың жүйелері қалай жұмыс істейтіні жайлы жазылған. Жүйенің қаншалықты дәл істейтінін тексеру үшін авторлар жүйеге әртүрлі сөз таптарындағы сөздерді енгізді. Жүйе зат есімнің, етістіктің, сын есімнің, сан есімнің түбірлерін таба алады. Жүйенің қайтарған жауабын 1-кестеден көруге болады. Мақаланың суреттерінен қазақ тілінің сөзінің түбіріне қандай жұрнақтарды, жалғауларды жалғауға болатынын көруге болады. Дегенмен, суреттерде барлық қиыстырулар көрсетілмеген. Авторлар мақаланың қорытынды бөлімінде сөздің түбірін анықтайтын жүйені қалай жасау керектігіне кеңес береді.

Түйін сөздер: сөздің түбірі, морфология, жұрнақ, жалғау, сөз таптары, алгоритмдер.

Аннотация

А.В. Богданчиков[1], О.А. Баймуратов[1], Д.А. Аязбаев[1]
[1]Университет имени Сулеймана Демиреля, г. Каскелен, Казахстан

СИСТЕМА ОПРЕДЕЛЯЮЩЯЯ КОРЕНЬ СЛОВА КАЗАХСКОГО ЯЗЫКА

На сегодняшний день обработка естественного языка широко применяется. Например, ее можно применить: при переводе текста, в поисковых системах, при определении темы текста. В таких применениях сначала следует предварительно обработать текст, так как она может повлиять на конечный результат. Существуют разные методы предварительной обработки текста. Одним из них является: определение корня слова. Данный метод может сберечь память компьютера. Потому что, повторяющиеся корни будут сохранены один раз. В данной статье рассматриваются разные системы, которые могут определять корни слов. В обзоре литератур рассматриваются алгоритмы для русского, узбекского, турецкого языков. Далее авторы предлагают свою систему, которая может определять корень слова казахского языка. В статье авторы описывают как работает их система. Чтобы проверить точность работы системы, было введено слова из разных частей речи. Предложенная система может определять корни имен существительных, прилагательных, числительных, глаголов. Ответ системы можно увидеть в таблице №1. Из рисунков статьи можно увидеть какие суффиксы, окончания можно соединять с корнями слов казахского языка. И тем не менее, в рисунках приведены не все комбинации. В заключении авторы дают советы по разработке системы, которая может определять корни слов.

Ключевые слова: корень слова, морфология, суффикс, окончание, части речи, алгоритмы.

1 Introduction

Nowadays application of natural language processing is wide. For instance, it can be used to explore people's opinions regarding some topic [1], in search engine system to recommend query [2], in machine translation systems and topic identification [3]. Sometimes text should be preprocessed. It is necessary to get better result. Preprocessing can be done by several ways. One approach is preprocessing by stemming. Stemming is the process of identifying root of word by removing the last characters from it [4]. The last characters can be suffixes, endings. For instance, to identify root of Kazakh word "қасқырдан", the last characters will be deleted until left part will be found in dictionary. In this case when ending "дан" will be deleted, word "қасқыр" will be found in dictionary. Hence "қасқыр" will be root. For languages with a lot of suffixes, endings stemming can be useful. Because there is no need to store all word forms, hence computer memory can be saved. In addition, stemming can be applied in spell checking systems. Currently there are many stemming algorithms, but all of them cannot be applied to the same language. It is because of languages can have different rules. In this paper we propose our stemming system for the Kazakh language.

2 Literature Review

Stemming can be preprocessing step of some analysis. Hence it can influence on the result of experiment. Nowadays there are many stemming algorithms for various languages. For example, in works [5-8] authors described algorithms respectively for Kazakh, Russian, Uzbek, Turkish languages. Languages can have different rules. Therefore one algorithm designed for specific language cannot be applied to other languages. In [6] stemming algorithm is written in Snowball string processing language. Snowball has compiler, which can translate Snowball program into source code of programming languages like C#, Go, Java, Javascript, Object Pascal, Python and Rust. In [7] authors discussed about Lovins stemmer, Porters stemmer, Dawson stemmer, Paice/Husk stemmer and Krovetz stemmer. In [7] authors tested Lovins and Paice/Husk stemmers by Uzbek words. After evaluating existing stemmers authors proposed their Uzbek stemmer. In [8] authors discussed about "A-F" , "L-M" algorithms to find out root of word . Authors of [9] paper developed stemming system for Arabic language. To identify root of word authors used FARASA text processing toolkit and Lucene library. Peculiarity of their stemming system is that it can identify plurality form of word.

3 Methodology

Kazakh language is an agglutinative language. According to [10] it has 10 parts of speech. In this paper following parts of speech were analyzed: noun, verb, adjective, numeral. [11] was used as a stemming dictionary. [11] was chosen because it contains part of speech. Pictures 1, 2, 3, 4 respectively show what kinds of suffixes and endings can be concatenated with noun, verb, adjective, numeral of Kazakh language. However they don't show all concatenation combinations.

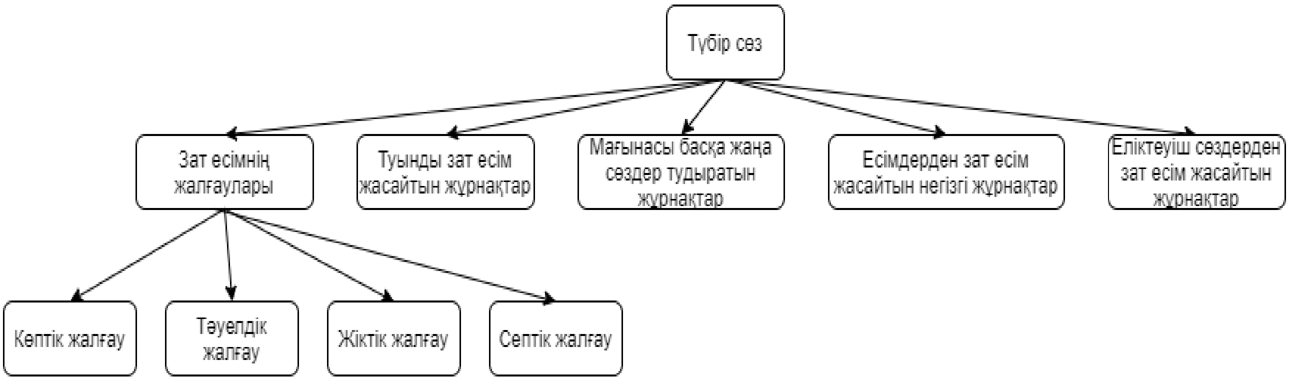
Түбір сөз
Зат есімнің жалғаулары
Туынды зат есім жасайтын жұрнақтар
Мағынасы басқа жаңа сөздер тудыратын жұрнақтар
Есімдерден зат есім жасайтын негізгі жұрнақтар
Еліктеуіш сөздерден зат есім жасайтын жұрнақтар
Көптік жалғау
Тәуелдік жалғау
Жіктік жалғау
Септік жалғау

Figure 1. Concatenation of noun with suffixes, endings

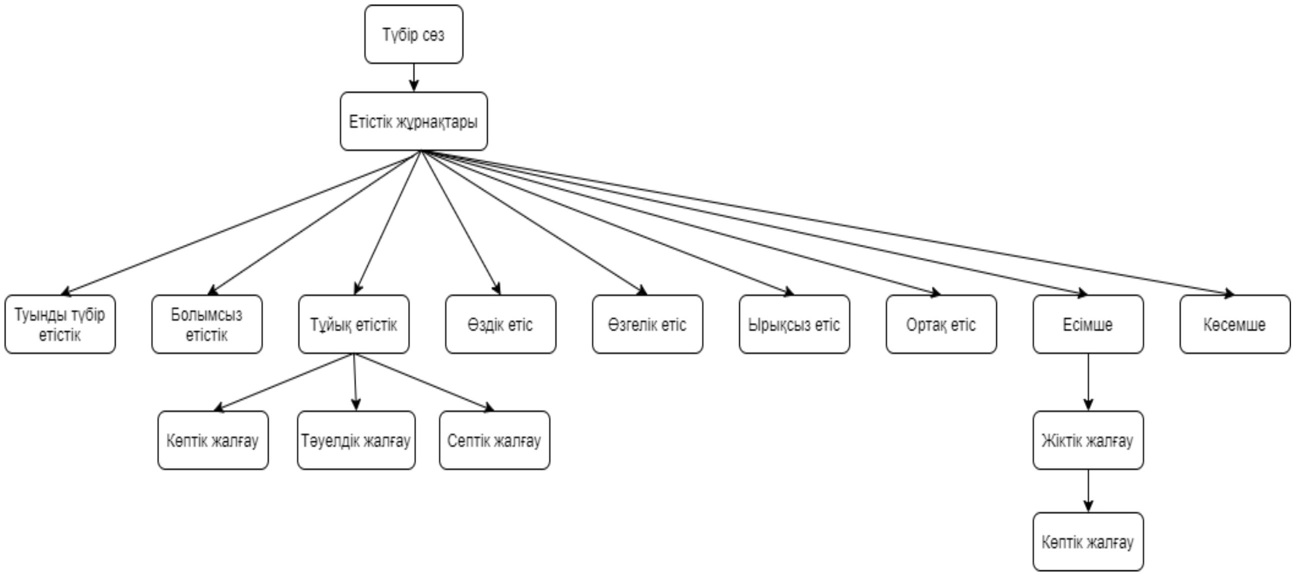

Figure 2. Concatenation of verb with suffixes, endings

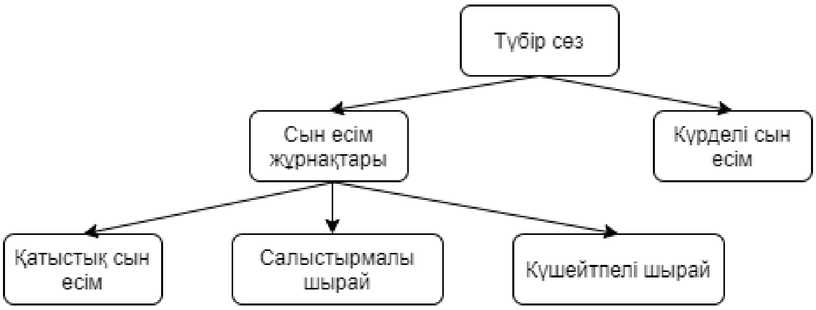

Figure 3. Concatenation of adjective with suffixes

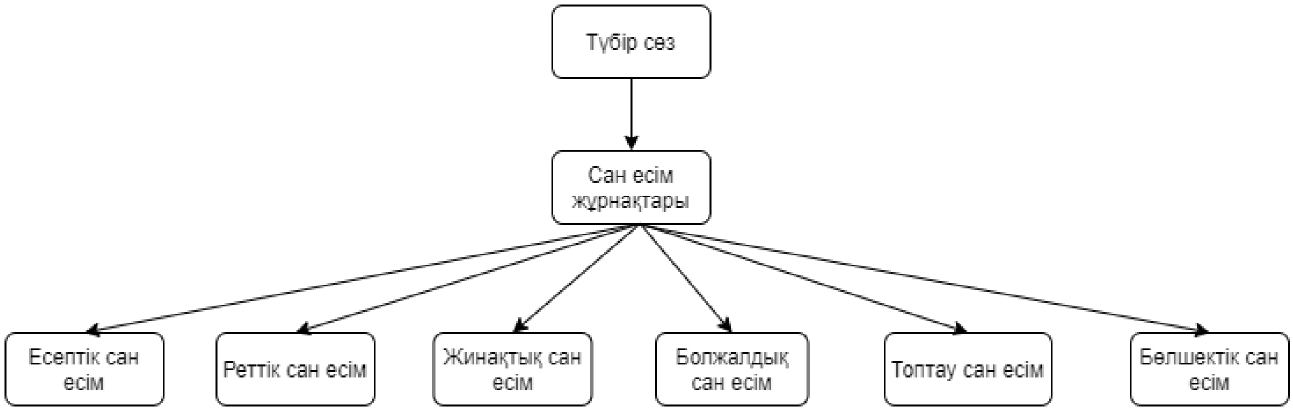

Figure 4. Concatenation of numeral with suffixes, endings

Stemming system of current paper works as following:

1) User enters some word.

2) Entered word will be searched in dictionary. If dictionary contains entered word, that word of dictionary will be considered as root. Entered word in dictionary can be met several times, hence several roots can be defined.

3) After that as were shown in pictures 1-4 various suffixes, endings will be concatenated to the root. For instance, user entered word “ұстаздың”. In [11] dictionary there is word “ұстаз”. “ұстаз” is subpart of word “ұстаздың”, hence “ұстаз” will be considered as root. Then the left part “дың” should be classified. The word “ұстаз” is noun, therefore the left part “дың” will be compared with suffixes, endings of noun. “дың” is a kind of ending(gen case), therefore final result will be “ҰСТАЗ”+GEN.

4 Data and results

To test our system we entered words from various parts of speech like noun, verb, adjective, numeral. Table 1 shows stemming results for words: тастардың, тастармен, толқын, тау.

Table 1. Entered words to test noun

Entered word	*System response*
тастардың	*ТАСТА+Verb+Pos+Қатыстық сын есім+Өзгелік етіс+Есімше+Pnon+Gen+A3Sg+Sg* *ТАС+Noun+Verb+A3pl+Pnon+Gen*
тастармен	*ТАСТА+Verb+Neg+Қатыстық сын есім+Өзгелік етіс+Есімше+P1+Dat+A3Sg+Sg* *ТАС+Noun+Verb+A3pl+P1pl+Dat*
толқын	*ТОЛҚЫ+Verb+Pos+Қатыстық сын есім+Өздік етіс+Ырықсыз етіс+Pnon+Sg* *ТОЛҚЫН+Verb+Pos+Қатыстық сын есім+Pnon+Sg* *ТОЛҚЫН+Noun+A3sg+Pnon+Nom*
тау	*ТАУ+Noun+A3sg+Pnon+Nom*

Where
Sg – singular
Pl – plural
A1, A2, A3 – type of ending (жіктік жалғау)
P1, P2, P3 – type of ending (тәуелдік жалғау).
Nom, Gen, Dat, Acc, Loc, Abl, Ins – respectively seven cases of Kazakh language.

Pnon – non possessive. Stemming was done correctly for some words, but some words were classified incorrectly. In table 1 words тастардың, тастармен, толқын have several roots, but actually each of them has only one root. Hence our system also returned wrong answers.

5 Discussion

To test our system we entered words from various parts of speech and examined results ourselves. Particularly, we entered words from [11] dictionary adding them endings, suffixes. Our stemming system is written in Python programming language. Python is convenient because it has short syntax. In addition, it has framework Django by which web version of stemming system can be developed. When stemming system is tested, various kinds of words like compound words, loan words from other languages should be given to the system. It should be done, because that words can have exceptional rules for root identification. For example, words of Kazakh language follow vowel harmony rules. However some loan words have exceptions (мысалы: корабльден, гастролінде). Our stemming system doesn't check presence of harmony rules. It concatenates to the root various kinds of suffixes and endings. Hence from loan words roots, suffixes, endings can be properly identified. In addition, in stemming there are two types of errors: under stemming, over stemming. Under stemming occupies when closely related words not stemmed to the same stem [3]. If words "дос", "достық", "достарым" have different stems, it will be under stemming. Over stemming occupies when words with different stemming grouped to the same stem. If words "аралар", "арақашықтық" have the same stem "ара". According to table 1 when words "тастардың", "тастармен" were entered our stemming system properly identified their roots as "тас". As shown in table 1 some words had several candidate roots. It happened because from [11] dictionary several times matchings were occupied. However our system classified the last parts with wrong roots incorrectly. Our system has other limitation. If user enters word, which is not in dictionary, our system will return nothing. Therefore when design stemming system, its dictionary should have a lot of words. In Kazakh language some words have several meanings. For instance, word "таста" has meanings to throw and on stone. To properly identify its root surrounding words should be also analyzed. Hence sometimes it is better to use lemmatization, which does it.

6 Conclusion

In current paper we developed stemming system of Kazakh language. In list below is couple of advices for designing stemming system.

1) When design stemming system its dictionary should be rich.

2) To add suffixes, endings exist special rules. Therefore, to design stemming system developer should know that rules.

3) Words can have several meanings, hence sometimes root cannot be identified only by one word. In this case surrounding words can help to define in which context that word was used. Therefore surrounding words should be also analyzed.

References

*1 1 Mamykova Zh.D., Mutanov G.M., Sundetova Zh.T., Torekul S.M. (2018) Podhody k razrabotke informacionnoj sistemy monitoringa mnenij i ocenki social'nogo samochuvstvija [*Approaches to the development of an information system for monitoring opinions and assessing social well-being*]. Vestnik KazNU. Serija matematika, mehanika, informatika – № 4(100), 63-77.*

2 Vidinli I. B., Ozcan R. New query suggestion framework and algorithms: A case study for an educational search engine// Information Processing and Management. – Accepted 1 February 2016. – V. 52. Iss. 5, - pages 733-752.

3 Nathani B., Joshi N., Purohit G.N. Design and Development of Unsupervised Stemmer for Sindhi Language //Procedia Computer Science, V. 167, - pages 1920-1927.

4 Margaret Rouse, Matthew Haughn. Stemming [Electronic resource]. –URL: https://searchenterpriseai.techtarget.com/definition/stemming (accessed date:05.01.2021)

5 Fedetov A.M., Tussupov J.A., Sambetbayeva M.A., Fedetova O.A., Sagnayeva S.K., Bapanov A.A., Tazhibaeva S.Zh. Classification model and morphological analysis in multilingual scientific and educational information systems. //Journal of Theoretical and Applied Information Technology. – 2016 – V.86. Iss 1, - pages 96-111.

6 Russian stemming algorithm [Electronic resource]. – URL: https://snowballstem.org/algorithms/ russian/stemmer.html (accessed date: 05.01.2021)

7 Abdul Jalil M., Ismailov A., Abd Rahim N.H., Abdullah Z. The Development of the Uzbek Stemming Algorithm // Advanced Science Letters. May 2017. – V. 23. Iss. 5.

8 Hayri S., Yıltan B. FindStem: Analysis and Evaluation of A Turkish stemming algorithm // String Processing and Information Retrieval, 10th International Symposium, SPIRE 2003, Conference: Manaus, Brazil, October 8-10, 2003, Proceedings

9 Alnaied A., Elbendak M., Bulbul A. An intelligent use of stemmer and morphology analysis for Arabic information retrieval //Egyptian Informatics Journal. – Accepted 18 February 2020. – V.21. Iss.4.

10 S.Arpabeko (2004) Kazak tili: Anyktamaly [Kazakh language: Directory] .Anyktamalykty kyrastyrushy – Kyrastyruga katyskandar: Abdikylova R.M., Kyzekova Z.S. – Almaty, 120. (In Kazakh)

11 Malbakov M., Esenova Ķ., Hinajat B., Shojbekov R., Yderbaev A., Zholshaeva M., Kyderinova Ķ., Zhybaeva O., Fazylzhanova A., Əshimbaeva N., Zhanabekova A. (2011) Kazak ədebi tilinin sөzdigi. [Dictionary of the Kazakh literary language] On bes tomdyk. 14-tom, Almaty, T. 14: – T–Ұ. – 800. (In Kazakh)