Development of Morphological Segmentation for the Kyrgyz Language on Complete Set of Endings

Aigerim Toleush[1][0000-0002-4369-4378], Nella Israilova[2][0000-0003-3121-3765], Ualsher Tukeyev [1][0000-0001-9878-981X]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2] Kyrgyz State Technical University named after I.Razzakov
toleush.aigerim@gmail.com, inela.kstu@gmail.com, ualsher.tukeyev@gmail.com

Abstract. The problem of word segmentation of source texts in the training of neural network models is one of the actual problems of natural language processing. A new model of the morphology of the Kyrgyz language based on a complete set of endings (CSE) is developed. Based on the developed CSE-model of the morphology of the Kyrgyz language, a computational data model, algorithm and a program for morphological segmentation are developed. Experiments on morphological segmentation of Kyrgyz language texts showed 82% accuracy of text segmentation by the proposed method.

Keywords: morphological segmentation, Kyrgyz language, morphology, model

1 Introduction

The problem of the word segmentation in neural network processing of natural language texts is a very urgent problem in connection with the problems of unknown words and rare words, problems of overflowing the memory of dictionaries of the neural machine translation (NMT) system. Especially the problem of overflowing the memory of dictionaries is relevant for agglutinative languages, since when processing texts of agglutinative languages, each word form of the language is written into the system's dictionary, which leads to overflow of the system's memory. In addition, since for training systems of neural network text processing of natural languages, the larger the amount of initial data for training, the better the quality of training the system. The Kyrgyz language belongs to agglutinative languages and is a low-resource language since there are low language resources for natural language processing (NLP).

This paper proposes a solution to the problem of morphological segmentation for the Kyrgyz language based on a new morphology CSE (Complete Set of Endings) – model [1]. The advantage of computational data model of word segmentation based on morphology's CSE-model is its tabular form, which makes it possible to obtain a solution in one-step. Moreover, computational solutions for CSE-model are the universal. Necessary for computational processing of segmentation of a particular language is the construction of a tabular data model of word segmentation of this language.

2 Related works

In recent years, the NMT systems output quality is significantly improved. The main problem of morphologically rich language translation is a fixed-size vocabulary of NMT consisting high frequency words to train the neural network. The using a fixed number of words in language pairs leads to out-of-vocabulary problems. Therefore, neural systems are often mistaken in the translation of rare and unknown words [2].

In addition, the large-size vocabularies are difficult to process because they affect the computational complexity of the model, taking up a large amount of memory. These lead to difficulties when translating the morphological-rich agglutinative languages that need a large vocabulary size.

The agglutinative languages in Turkic languages group have a complex structure of words and many ordering the affixes, which can produce a large vocabulary. An example of such languages with complex morphology and low-resourced is the Kyrgyz language. Because of the inflectional and derivational morphology, it has a large volume of word list. For the Kyrgyz language, the above presented issues are significant because of the inability of the NMT to determine the complex structure of words in the training corpus.

The morphologically rich languages need a deep analysis at the word level. In order to reduce the vocabulary size and improve the translation of morphologically rich languages, morphological word segmentation is used. The morphological segmentation is a process of dividing words in sub-units. As the result of morphological segmentation, the words in the agglutinative languages splits to the word stem and affixes that reduces the vocabulary size for model training. Previous studies had shown that a morpheme-based modeling of the language can provide a better performance [3].

Two methods for unsupervised segmentation of words into small units in Finnish and English languages were introduced in [4]. They used Minimum Description Length (MDL) principle with recursive segmentation in the first method. The linear segmentation and Maximum Likelihood (ML) optimization are used in the second method. For the morphologically rich Finnish language, the recursive segmentation with the MDL principle had shown better results than sequential segmentation.

The BPE (byte pair encoding) method was proposed as rare word segmentation and translation of them by NMT [2]. In this paper was showed that the NMT was able to translate rare words by their subunits. To increase the quality of word segmentation they had used list of stop words, words that cannot be divided into segments. At the result, they demonstrated that NMT models with open vocabulary could better translate rare words than models with large-size vocabularies.

A group of scientists represented word segmentation method that reduces the vocabulary of agglutinative languages and maintains the linguistically word properties [5]. They called this method Linguistically Motivated Vocabulary Reduction (LMVR). To develop the LMVR they used unsupervised segmentation framework Morfessor Flatcat and can be used for any language pairs where a source language is morphologically rich. The LMVR outputs better results than the byte pair encoding technique [6].

It should be noted that the task of word segmentation differs from the task of morphological analysis. Therefore, we do not provide an analysis of works on morphological analysis. Although, of course, the problem of word segmentation can be solved using adapted methods and algorithms for morphological analysis [7, 8]. However, the

authors believe that this way of solving the word segmentation problem is computationally expensive.

In this paper, the word segmentation of the Kyrgyz language texts through creation of morphology's CSE-model is realized.

3 General characteristics of the Kyrgyz language

The Kyrgyz language is a one of the Turkic languages and the national language of the indigenous population of the Kyrgyz republic. Native speakers of Kyrgyz language are ethnic Kyrgyz living in the bordering regions of Kazakhstan, in the Namangan, Andijan and Ferghana regions of Uzbekistan, in some mountain regions of Tajikistan and the Xinjiang Uygur Autonomous Region of the China and some other countries [9].

According to the morphological classification, the Kyrgyz language refers to agglutinative languages. They are characterized by system of word-formation and inflectional affixation, a single type of declension and conjunction, grammatical uniqueness of affixes and the absence of significant alternations [8, 9].

The part of speech that indicates an action, state of the person and object is called a verb. The verb has conjugated and non-conjugated forms. The non-conjugated forms include the participles, the action names. Conjugate forms in a sentence act as predicate, non-conjugate forms as any member of a sentence. The verb has grammatical categories of time, person, number, mood, voice.

The affixes may be used in combination as: plural-case (ата-лар-да (fathers have), stem-plural affix-locative case affix); plural-possessive (ата-лар-ы (their fathers), stem-plural affix-possessive 3rd person affix); plural-person (ата-лар-быз (we are fathers), stem-plural affix-1 person plural affix); plural-possessive-case (ата-лар-ыбыз-дын (of our fathers), stem-plural affix-plural affix-possessive affix 1st person plural form); plural-possessive-person (ата-лар-ы-быз (we are their fathers), stem-plural affix-possessive affix 3rd person-1st person plural type affix); person-case; possessive-case (бала-м-ды (my child), stem-possessive 1st person type affix-accusative affix), possessive-person (бала-ң-мын (I am your child)).

In the morphological segmentation, we do not consider derivational affixes, because they change the meaning of words and the grammatical word class by producing new word.

4 Development of Kyrgyz morphology's CSE-model

For developing the complete system of endings of the Kyrgyz language, the approach proposed by Tukeyev U. in [1] is used.

Consider the Kyrgyz language ending system of two classes: nominal parts of speech affixes (nouns, adjectives, numerals) and endings of verb base words (verbs, participles, moods, voices). In this section, we will consider the complete set of endings in Kyrgyz language.

The Kyrgyz language has the four types of endings: plural endings (denoted by K); possessive endings (denoted by T); case endings (denoted by C); personal endings (denoted by J); a stem denoted by S.

There are four variants of placing the types of endings: from one type, from two types, from three types and from four types. A number of placements defines by following formula:

$A_n^k = n!/(n-k)!.$ (1)

Then, the number of placements will be determined as follows:

$A_4^1 = 4!/(4-1)!=4,$
$A_4^2 = 4!/(4-2)!=12,$ (2)
$A_4^3 = 4!/(4-3)!=24,$
$A_4^4 = 4!/(4-4)!=24.$

Total possible placements are 64.

Consider which of them are semantically valid.

Placements of one type of ending (K, T, C, J) are all semantically valid by definition.

Placements of two types of endings can be as follows: **KT, TC, CJ**, JK **KC, TJ**, CT, JT **KJ**, TK, CK, and JC.

An analysis of the placement semantics by two types of endings shows that the bold placements are valid (**KT, TC, CJ, KC, TJ, KJ**) and other placements are invalid. For instance, JK – there is no use of plural endings after personal endings. Therefore, the number of permissible placements of two types of endings will be 6.

Acceptable placements of three types of endings will be 4 from 24 possible variants: **KTC, KTJ, TCJ**, **and KCJ**.

Placement of four types of endings will be 1: **KTCJ.**

In total, there are 4 allowed placements from one type, 6 from two types, 4 from three types and 1 from four types of endings. The total number of all acceptable placements is 15.

The set of the endings of verbs in Kyrgyz language contains following types:

- set of verbs endings,
- set of participle endings (атоочтук),
- set of gerund endings (чакчылдар),
- set of mood endings,
- set of voice endings.

The set of endings to verb stems includes following types: tenses (9 types); person (3 types); negative type. Then the number of possible types of verb endings is 28.

The set of endings to verb stems of participle includes endings of participle (denoted by R), plural endings (denoted by K), possessive endings (denoted by T), case endings (denoted by C), personal endings (denoted by J).

Then, the possible variants of ending types will be:

1) with one type of endings: **RK, RT, RC, RJ,**

2) with two types of endings: **RKT, RTC,** RCJ, RJK, **RKC**, RTJ, RCT, RJT, **RKJ**, RTK, RCK, RJC,

3) with three types of endings: **RKTC**, RTCJ, RCJK, RJKT, RKTJ, RTCK, RCJT, RJKC, RKCJ, RTJK, RCTK, RJTK, RKCT, RTJC, RCTJ, RJTC, RKJT, RTKC, RCKT, RJCK, RKJC, RTKJ, RCKJ, RJCT,

4) with four types of endings: RKTJC, RTKJC, RCKTJ, RJKTC, RKTCJ, RTKCJ, RCKJT, RJKCT, RKJTC, RTJKC, RCTKJ, RJTKC, RKJCT, RTJCK, RCTJK, RJTCK, RKCTJ, RTCJK, RCJKT, RJCKT, RKCJT, RTCKJ, RCJTK, RJCTK.

In total, the number of acceptable ending types is 9.

Consider the ending types of gerund. They contain endings followed by personal endings: PJ, where P is a base of gerund, J is the personal ending. Therefore, there is 1 type of gerund ending.

There are five forms of voice in Kyrgyz language, however we will consider four of them: reflexive, passive, joint, compulsory. The basic voice has no special endings, and it is the initial form for the formation of other voice forms using affixes. Respectively, there are 4 types of voice endings.

Moods in Kyrgyz language have five forms: imperative, conditional, desirable, subjunctive and indicative. The formal indicators of indicative mood are the affixes of the verb tense. Therefore, indicative mood affixes will not be considered. Types of mood endings are also determined according to the previous scheme: the basic endings of moods followed by personal endings. In total, there are 6 types of mood endings.
As a result, the total number of types of endings of verb stems will be 48. According to these calculations, the total number of types of endings with noun stems and with verb stems will be 63.

The next task is to determine the forms of endings and their number from the received types of endings. In this direction, finite sets of endings for all the main parts of the Kyrgyz language were constructed. In first, it needs to define the number of endings of single types K, T, C, J (Tables 1-3).

Table 1 – Number of endings type K and T.

Suffixes type K	Suffixes type T	
	Singular	Plural
-лар- -лер- -лор- -лөр- -дар- -дер- -дор- -дөр- -тар- -тер- -тор- -төр-	м, -ым (-им, -ум, -үм),	-быз (-биз, -буз, -бүз), ыбыз (-ибиз, -убуз, -үбүз),
	-ң, -ың (-иң, -уң, -үң) ,	-ңыз (-ңиз, -ңуз, -ңүз), -ыңыз (-иңиз, -уңуз, -үңүз),
	-ыңыз, -иңиз, -уңуз, -үңүз	-ңар (-ңер, -ңор, -ңөр), -ыңар (-иңер, -ңар, -үңөр), -ңыздар (-ңиздер, -ңуздар, -ңүздөр), -ыңыздар (-иңиздер, -уңуздар, -үңүздөр)
	-сы (-си, -су, -сү), -ы (-и, -у, -ү) -ныкы (-ники, -нуку, -нүкү); -дыкы (-дики, -дуку, -дүкү); -тыкы (-тики, -туку, -түкү).	-сы (-си, -су, -сү), -ы (-и, -у, -ү), -ныкы (-ники, -нуку, -нүкү); -дыкы (-дики, -дуку, -дүкү); -тыкы (-тики, -туку, -түкү).
12	62 (different)	

Table 2 – Number of endings type C.

Suffixes types C	
1. Атооч (nominative)	-
2. Илик (genitive)	- нын, нин, нун, нүн, дын, дин, дун, дүн, тын, тин, тун, түн;
3. Барыш (dative)	- га, ге, го, гө, ка, ке, ко, кө, а, е, о, ө, на, не, но, нө;

4. Табыш (accusative) 5. Жатыш (locative) 6. Чыгыш (ablative)	- ны, ни,ну, нү, ды, ди, ду, дү, ты, ти, ту, тү, н;-да, де, до, дө, та, те, то, тө -да, де, до, дө, та, те, то, тө, нда, нде, ндо, ндө; -дан, ден, дон, дөн, тан, тен, тон, төн, нан, нен, нон, нөн.
	65 (different)

Table 3 – Number of endings type J.

Suffixes type J	
Singular	Plural
J1)-мын, мин, мун, мүн; J2)-сың, сиң, суң, сүң; J3)-сыз, сиз, суз, сүз; J4) -	-быз, (-пыз), -биз, (-пиз), -буз, (-пуз), -бүз, (-пүз), -сыңар, сиңер, суңар, сүңер; - сыздар, сиздер, суздар, сүздер; -
28	

The sum of all four types of endings, specifically K, T, C, J is 167 (different).

Combinations in placement KT: K*T= (12 suffixes K) *(5 suffixes different T) = 60 endings KT. For choosing of suffix in T for each suffix of K is working harmony rules of the Kyrgyz language. In addition, a similar suffix of T not take to account, therefore are 5 suffixes T. The Table 4 shows the number of possible endings for a pair of plural and possessive endings denoted as KT.

Table 4 – Number of endings of KT

Example	Suffixes type K	Suffixes type T		Number of endings
		Singular	Plural	
	-лар- -лер- -лор- -лөр- -дар- -дер- -дор- -дөр- -тар- -тер- -тор- -төр-	-ым, -им, -ум, -үм	-ыбыз, -ибиз, -убуз, -үбүз	12*5=60
		-ың, -иң, -уң, -үң	-ыңыз, -иңиз, -уңуз, -үңүз	
		-ыңыз, -иңиз, -уңуз, -үңүз	-ыңыз, -иңиз, -уңуз, -үңүз	
		-ы, -и, -у, -ү	-ы, -и, -у, -ү	
ата (grandfa-ther)	-лар-	-ым, -ың, -ыңыз, -ы	-ыбыз	5

In the Table 4 the example has five endings: ата-лар-ым (my fathers), ата-лар-ың (your fathers), ата-лар-ыңыз (your fathers polite), ата-лар-ы (their fathers), ата-лар-ыбыз (our fathers).
Here the law of vowel harmony: hard vowels with hard vowels, and soft vowels with soft vowels. The vowel in a root defines a tone of affix. This means that vowel of affixes should be in harmony with the last syllable of the word.

Calculating the remaining types of endings, we will get that type placement KC has 60 endings and type placement KJ has 36 endings.

In pair of CJs: the genitive, dative and accusative cases are not used, thus, we consider only locative and ablative cases. We cannot consider the nominative case because it does not have any suffixes and represents the initial form of the word. The locative and ablative cases have 8 endings for each of them. So, there are 96 endings of CJ combination.

The possessive T endings have singular and plural types: -м, -ң, -ңыз, -сы, -быз, -ңар, -ңыздар. Therefore, for singular and plural types the number of endings is considered separately. There are 14 types of singular T endings and 16 types of plural T endings. As the 5 types case suffixes can be used with possessive endings the total number of TC possible endings will be 150 (14*5+16*5).

In the TJ placement the T1 cannot be used with the first side J ending*: "ата-м-мын"* is wrong, because we cannot say "me is my grandfather". Also, T2 and T3 connects only with J1 because there is no use of T2 with J2 as: "*ата-ң-сың" – "you are your grandfather"*. The possible combination of T and J endings and the total number 62 of TJ is shown in Table 5 below.

Table 5 – Number of endings of TJ

Suffixes type T				Suffixes type J		Number of endings
Vowels		Consonants		Singular	Plural	
Hard	Soft	Hard	Soft			
Т1-м Т2-ң Т3-ңыз Т4-сы	-м, -ң, -ңи, -си	-ым -ы -ыңыз -ы	-им -иң -иңиз -и	J1)-мын, мин, мун, мүн; J2)-сың, сиң, суң, сүң J3)-сыз, сиз,суз, сүз; J4) -	-быз, биз, буз, бүз, пуз, пүз; -сыңар, сиңер, суңар, сүңер; -сыздар, сиздер, суздар, сүздер; -	Т1-J2:8; Т1-J3:8; Т2-J1:10; Т3-J1:10; Т4-J1:10; Т4-J2:8; Т4-J3:8; Total: 62

Considering the endings of type placement KTC we will get next: there are 12 types of plural, 4 types of possessive and 5 types of case endings. As a result, the possible number of endings of KTC is 240.

In the Kyrgyz language, there is possible only one combination of endings from four types: KTCJ. There can be a combination of endings from 12 affixes type K, 4 affixes type T, 2 affixes type C and 6 affixes type J. At the result, we will get 576 endings.

The verbs can be used in three tenses as future, present and past. There are four types of past tense definite, accidental, general, habit. The accidental past tense has 172 possible affixes in total with its negative affixes. The habit past tense is formed by attaching affix -чу, -чү to the word stem. So, there are 20 possible combination of habit past tense and possessive, however, it does not have a negative type.

The all types of past tense on 3rd person plural form are made by attaching affix -ыш (-иш, - уш, -үш, -ш) to the word stem following tense affixes. For example: *бар-ыш-ты (went), ише-ш-ти (worked).*

The Table 6 below shows the definite past tense suffixes with personal endings.

Table 6 – Definite past tense affixes

Number	Person	Suffixes				Number of endings
Singulars	1 person	-ды, ты-м	-ди, ти-м	-ду, ту-м	-дү, тү-м	8
	2 person	-ды, ты-ң	-ди, ти-ң	-ду, ту-ң	-дү, тү-ң	8
	2 person polite	-ды, ты-ңыз	-ди, ти-ңиз	-ду, ту-ңуз	-дү, тү-ңүз	8
	3 person	-ды, ты	-ди, ти	-ду, ту	-дү, тү	8
Plural	1 person	-ды, ты-к	-ди, ти-к	-ду, ту-к	-дү, тү-к	8
	2 person	-ды, ты-ңар	-ди, ти-ңер	-ду, ту-ңар	-дү,тү-ңөр	8
	2 person polite	-ды, ты-ңыздар	-ди, ти-ңиздер	-ду, ту-ңуздар	-дү, тү-ңүздөр	8
	3 person	ыш, ш-ты	-иш, ш-ти	-уш, ш-ту	-үш, ш-тү	8
Total						64

The negative form of definite past tense is formed by adding after the stem of the verb the affix -ба (-бе, -бо, -бө), if the stem ends with a vowel or voiced consonant. If the stem ends to a deaf consonant then the negative form of definite past tense is made by attaching the affix -па (-пе, -по, -пө). For example, *мен бар-ба-ды-м (I did not go), сиздер кел-бе-ди-ңиздер.* So, the number of endings for negative form of definite past tense is 64 too.

An accidental past tense can be formed through attaching affix -ып (-ип, -уп, -үп, -п) or affix -ыптыр (-иптир, -уптур, -үптүр, -птыр, -птир, -птур, -птүр) and possessive endings after the verb stem. Number of endings of accidental past tense is 112. Moreover, other types of past tense have 64 affixes each.

Present simple and complex tenses have 12 affixes in total. The present simple forms by -ууда (-оодо, -өөдө, -үүдө) affix and affixes of definite future tense. While the complex present tense is formed by participle 2 affix -а (-е, -й) and -ып (-ип, -уп, -үп, -п) attached to the base of the verb and auxiliary verbs *"жат" – "lie", "тур" – "stand)", "отур" – "sit", "жүр" – "go"* in the present simple tense. For example: ачыл-ууда; жаз-а-м; кел-е жат-а-мын; күт-үп жүр-бүз.

In the Kyrgyz language, verbs have three types of future tense. The definite future tense *(Айкын келер чак)* is formed by joining the participial affix -а with phonetic variants and personal affixes to the base (root). So, there are 69 positive endings of definite future tense. In the 1st person, the full (а-мын) and incomplete (а-м) forms are possible: *бар-а-мын, бар-а-м -I will go.* If the last syllable of the stem ends with a vowel, then the participle affix -й added: *оку-й-мун – I will read.* If the last syllable of the stem ends with a consonant, then the phonetic variants of the participle affix -а (-е, -о, -ө) are added.

The negative type of verb in future tense is formed using affix -ба (-бе, -бо, -бө, -па, -пе, -по, -пө) after the word stem. There are 72 forms of endings of definite future tense negative type.

Indefinite future tense is formed by joining an affix -ар with phonetic variants followed by personal affixes to the word stem. Also, its negative type is formed using an affix -бас with phonetic variants. As the result of calculation, there are 61 possible affixes of indefinite future tense and 64 affixes of its negative type. The complex future tense has 8 types of affixes.

Thus, the total number of affixes of the verb in Kyrgyz language is 721.

The participles have nine acceptable placements of affixes. The participle 1 has 19 affixes (-ган, -ген, -гон, -гөн, -кан, -кен, -кон, -көн,-оочу, -уучу, -өөчү, -үүчү, -чу, -чү-, ар, -ер, -ор, -өр, -р), but only first 14 of them are used with plural affixes. The example of the word "*бар" – "go"* in the form of RKC is "*бар-ган-дар-га" – "to people who is went".*

The affixation of word *"айт"(say)* with RJ endings will be as following: *айт-кан-мын (I said); айт-кан-сың (You said); айт-кан-сыз (You said (polite)); айт-кан-быз (We said); айт-кан-сыңар (You said (plural)); айт-кан-сыздар (You said (plural polite)).*

In RKJ endings, the participle has 14 forms of suffixes because of the suffixes (-ар, -ер, -ор, -өр, -р) do not use with plural suffixes. The Table 7 shows the analysis of RKJ endings.

Table 7 – Number endings of RKJ

Example	Suffixes type R	Suffixes type K	Suffixes type J	Number of endings
	-ган, -ген, -гон, -гөн, -	-лар- -лер- -лор-	-мын, -мин, -мун, -мүн, -быз, -биз, -буз, -бүз, -пыз, -пиз, -пуз, -пүз	14*3=42

	кан, -кен, -кон, -көн- -оочу, -уучу, -өөчү, -үүчү, -чу, -чү-	-лөр- -дар- -дер- -дор- -дөр- -тар- -тер- -тор- -төр-	-сың, -сиң, -суң, -сүң, -сыңар, -сиңер, -суңар, -сүңер	
			-сыз, -сиз, -суз, -сүз, -сыздар, -сиздер, -суздар, -сүздер	
Бар-	-ган-	-дар-	-быз, -сыңар, -сыздар	

In the Kyrgyz language, the following moods of the verb are distinguished: Imperative mood; Conditional mood; Desirable mood; Subjunctive mood.
The imperative mood is made from verbal stem through attaching the affixes -гын, -кын, -ңыз, -ыңыз, -сын, -гыла, -кыла, -ңыздар, -ыңыздар and their phonetic variants, also, negative type is formed by attaching the -ба, (-бе, -бо, -бө, -па, -пе, -по, -пө) affix after the stem. As a result, there are 78 forms of imperative mood suffixes. There are 84 of conditional mood suffixes that are formed by using affix -са and its phonetic variants with possessive affixes. For example: *жаан жаада – if it rains, барсам – if I go, барбаса – if he/she do not go, жазсаңыз – if you write.*
The desirable mood has 31 suffixes and subjunctive mood affixes consists of 8 forms of affixes. The positive type suffixes of conditional mood have 69 forms of suffixes.

At the result of the constructed complete sets of endings, for nominal stems is 2 096, and the number of endings for verb stems is: verbs – 721, participles 1 – 1 325, gerund – 25, moods – 201, voices – 189, action name – 8, derivative adverbs – 36, and number of base suffixes – 167 (K-12, T-62, C-65, J-28). Total, there are 4 768 endings in Kyrgyz.

The developed CSE model of the Kyrgyz language is a computational morphology model based on the use of the complete set of language endings [10]. This computational model belongs to the "Item and Arrangement" (IA) morphology model [11, 12]. IA-model is convenient for agglutinative languages. The CSE computational model is an alternative to the TWOL computational model of morphology [13], which is more focused on describing the morphology of inflected languages. TWOL computational model is useful for description of the dynamic nature of allomorphs, uses two levels of word forms representation and refers to the morphology model "Item & Process" (IP-model).

5 Developing a computational data models and an algorithm of morphological segmentation

The principle of the Kyrgyz language word segmentation is next. The morphological segmentation using the CSE model is that a tabular data structure is built, which is a decision table. This decision table has two columns: the first column consists the word's endings; the second column is the segmented endings. The morphological segmentation algorithm of the current word will consist of:

1) finding the ending of the word,
2) finding the segmentation of the found word's ending.
To finding the ending of word, stemming algorithms are used without a stem dictionary and with a stem dictionary [14, 15]. To find the segmentation of the endings of words, a decision table of segmentation of word endings is used: the word's ending found during stemming of a word is searched for in the first column of the decision table, then the selected value of the second column in the found row of the table is the segmentation of this word's ending. For experiments are used the universal program[1] of stemming and segmentation based on CSE model.

6 Results of the experiments

This works main target was to develop morphological segmentation for the Kyrgyz language through a vocabulary of endings, created by technology of CSE-model, and to use a list of stop words for improving of result of segmentation. The segmentation program was developed taking into account the above-mentioned algorithms, schemes and all the rules. Pronouns, conjunctions, postpositions, interjections, particles, modal words and auxiliary verbs were taken as stop words. These are the words that have no endings. Also, stop word list contains names of people and place names. The total number of stop words is 600.
In order to verify the operation of the algorithm and the program, the following testing was done. For segmentation, a texts were taken from the official web pages of the government, ministries, news agencies and the grammar books. This was done to mix different genres of publications. Common volume of text is 6450 words. The part sentences of them with segmentation is provided below:

- *“Атам келгенге чейин күтө турамын”(Wait till my father comes)*, segmented as *“Ата-м кел-ген@@ге чейин күт-ө тур-а@@мын.”*
- *“Алар жолго камынып жатышат”(They are getting ready for the road)*, segmented as *“Ал-ар жол-го камын-ып жат-ыш@@а@@т.”*
- *“Биздин үйгө коноктор келишти”(Guests came to our house)* , segmented as *“Биз-дин үй-гө конок-тор кел-иш@@ти.”*
- *“Апамдан китептер кайда турганын сурадык”(We asked my mother where books)* , segmented as *“Апам-дан китеп-тер кайда тур-ган@@ы@@н сура-ды@@к.”*
- *“Мындай дарылар дарыканаларда рецептсиз сатылбайт” (These drugs are not sold without prescription in pharmacies)*, segmented as *“Мындай дары-лар дарыкана-лар@@да рецепт-сиз сатыл-ба@@й@@т.”*
- *“Мен компьютерде отурганды анча жактырбайм” (I do not like sitting at the computer), segmented as “Мен компьютер-де отур-ган@@ды анча жактыр-ба@@й@@м.”*
- *“Биздин көчөнүн балдарынын көбүрөөгү мектепте окушат” (Most of the children on our street go to school), segmented as “Биз-дин көчө-*

[1] https://github.com/NLP-KazNU

нүн бал-дар@@ы@@нын көбүрөө-гү мектеп-те оку@@ш@@а@@т."

The results show that 5 264 words out of 6 450 are segmented correctly. This means that the word segmentation accuracy is 82%. The analysis of incorrect segmentations show that causes of incorrect segmentation are: 1) for single letter endings - accepting the last letter of stem as an ending of participle, gerund and possessive ending; 2) for endings from 2-d and 3-d letters - accepting the last letters of stem as a possessive, past and future tense ending.

7 Conclusion and future works

In this paper, we propose a solution to the word segmentation of the Kyrgyz language by using the CSE - morphology model. A computational model on base of CSE - model of the morphology of the Kyrgyz language has been built, a computational data model for morphological segmentation as decision table has been developed, and experiments have been carried out with texts of the Kyrgyz language. The results obtained show 82% accuracy of morphological text segmentation. To improve the quality of morphological segmentation of the text, it is planned to conduct research on the improving of the dictionary's volume of Kyrgyz language stems in the proposed scheme of morphological segmentation. In future works is planned to use received results for preprocessing stage for an investigation of neural machine translation for the Kyrgyz language pair with different languages.

References

1. Tukeyev U.: Automation models of the morphological analysis and the completeness of the endings of the Kazakh language. Proceedings of the International Conference "Turkic Languages Processing: TurkLang-2015". – Kazan: Publishing house Academy of Sciences of the Republic of Tatarstan, pp. 91-100, 2015. (in Russian)
2. Sennrich R., Haddow B., Birch A.: Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1715-1725, 2016.
3. Pan Y., Li X., Yang Y., Dong R.: Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation. ArXiv: 2001.01589. (2020).
4. Creutz M., Lagus K.: Unsupervised discovery of morphemes. Proceedings of the ACL-02 workshop on Morphological and phonological learning, vol. 6, pp. 21-30, 2002.
5. Ataman D., Negri M., Turchi M., Federico M.: Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. The Prague Bulletin of Mathematical Linguistics, Vol. 108, pp. 331–342, 2017.
6. Ataman D., Federico M.: An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation. Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, Vol. 1, pp. 97-110, 2018.

7. Washington, J. N., Ipasov, M. and Tyers, F. M.: A finite-state morphological analyser for Kyrgyz.// Proceedings of the 8th Conference on Language Resources and Evaluation, LREC2012- Istanbul, Turkey, pp. 934-940, 2012
8. Israilova N.A., Bakasova P.S.: Morphological analyzer of the Kyrgyz language. Proceedings of the V International Conference on Computer Processing of Turkic Languages Turklang 2017. - Conference proceedings. In 2 volumes. T - Kazan: Publishing house Academy of Sciences of the Republic of Tatarstan, pp. 100-116, 2017. (in Russian)
9. Biyaliev K.A.: Guide to the grammar of the Kyrgyz language. / Kyrgyz-Russian Slavic University. Bishkek, 128 p. (in Russian)
10. Tukeyev U.: Computational models of Turkic languages morphology on complete sets of endings. QS Subject Focus Summit 'Modern Languages and Linguistics', section 'Linguistics and Artificial Intelligence', report, 2020. https://qssubjectfocus.com/moscow-2020/
11. Spencer A.: Morphological theory. An Introduction to Word Structure in Generative Grammar. Blackwell Publishers. pp.512, 1991.
12. Plungyan V.A.: Common morphology: Introduction to problematics: Educational manual. 2-d edition, edited.-M.: Editorial UPCC, pp. 384, 2003. (in Russian)
13. Koskenniemi K.: Two-level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki, pp.160, 1983.
14. Tukeyev U.A., Turganbayeva A.: Lexicon - free stemming for the Kazakh language // Proceedings of the international scientific conference "Computer science and Applied Mathematics" dedicated to the 25th anniversary of Independence Of the Republic of Kazakhstan and the 25th anniversary of the Institute of Information and Computing Technologies. - Almaty, pp. 84-88, 2016. (in Russian)
15. Tukeyev U., Karibayeva A., Zhumanov Zh.: Morphological Segmentation Method for Turkic Language Neural Machine Translation. Cogent Engineering, Volume 7, 2020 - Issue 1 https://doi.org/10.1080/23311916.2020.1856500.