cogent engineering

COMPUTER SCIENCE | RESEARCH ARTICLE

Morphological segmentation method for Turkic language neural machine translation

U. Tukeyev[1]*, A. Karibayeva[1] and Z h. Zhumanov[1]

Received: 24 June 2020
Accepted: 23 November 2020

*Corresponding author: U. Tukeyev, Information Systems, Al-Farabi Kazakh National University, Almaty, 050040, Kazakhstan
E-mail: ualsher.tukeyev@gmail.com

Reviewing editor:
Duc Pham, School of Mechanical Engineering, University of Birmingham, Birmingham, United Kingdom

Additional information is available at the end of the article

Abstract: Dictionaries play an important role in neural machine translation (NMT). However, a large dictionary requires a significant amount of memory, which limits the application of NMT and can cause a memory error. This limitation can be solved by segmenting each word into morphemes in parallel source corpora. Therefore, this study introduces a new morphological segmentation approach for Turkic languages based on the complete set of endings (CSE), which reduces the vocabulary volume of the source corpora. Herein, we demonstrate the proposed CSE-based morphological segmentation method for the Kazakh, Kyrgyz, and Uzbek languages and present the results of computational NMT experiments for the Kazakh language. The NMT experiment results show that in comparison with byte-pair encoding (BPE)-based segmentation, the proposed CSE-based segmentation increases the bilingual evaluation understudy score of 0.5 and 0.2 points on average for Kazakh–English and English–Kazakh pairs, respectively. Furthermore, in comparison with the BPE-based segmentation, the proposed CSE-based segmentation approach reduced the vocabulary size in NMT by more than a factor of two. This feature of the proposed segmentation approach will be crucial for NMT as the size of the source corpora is increased to improve translation quality.

Subjects: Computational Linguistic & Language Recognition; Neural Networks; Morphology

Keywords: neural machine translation; morphological segmentation; Turkic languages; Kazakh; Kyrgyz; Uzbek

ABOUT THE AUTHOR

Professor Ualsher Tukeyev runs an active research group that focusses on development and investigation in area of Natural Languages Processing (machine translation, computational linguistics, corpus linguistics) of Turkic languages. Specifically, current research of group focuses on the development of linguistically feature oriented methods for support of neural machine translation. Proposed segmentation method is used on the current project “Development and research of the Kazakh language neural machine translation system”, financed by Ministry of Education and Science of Republic Kazakhstan. Also, this segmentation method is used for the Uzbek and Kyrgyz languages investigations.

PUBLIC INTEREST STATEMENT

The article proposes a novel segmentation method for agglutinative languages to be used in neural machine translation pre-processing, which more than twice decrease the volume of the machine translation vocabulary. The proposed segmentation method is based on the construction of a complete set of language endings. A decrease in the machine translation vocabulary allows to increase a volume of the input parallel corpus for training, which leads to increase the quality of machine translation. The proposed segmentation method for agglutinative languages could be well used in the field of information retrieval for lexicon-free stemming of words, for morphological analysis of texts, for morphological tagging of language corpora.

1. Introduction

Turkic[1] languages are one of the largest language families and are spoken by more than 160 million people; the languages in this family include Turkish, Azeri, Uzbek, Kazakh, Tatar, and Kyrgyz. Approximately 13, 4.4, and 24 million people speak in Kazakh, Kyrgyz, and Uzbek languages, respectively. The Kazakh speakers live in Kazakhstan, Russia, China, Uzbekistan, Mongolia, and Turkmenistan; the Kyrgyz speakers live in Kyrgyzstan, Uzbekistan, China, and Tajikistan; and the Uzbek speakers live in Uzbekistan, Afghanistan, and Tajikistan.

Turkic languages are agglutinative, thereby making them challenging for neural machine translation (NMT) because almost all the words from the source corpus must be included in the dictionary. NMT learning is generally improved by increasing the vocabulary size; however, if the vocabulary is significantly large, the memory will overflow, thereby resulting in system error. This system error can be avoided by word segmentation.

In this study, a new morphological segmentation method is proposed by considering three languages as examples, which are Kazakh, Kyrgyz, and Uzbek. The Kazakh language was considered for experiments due to the presence of scientific developments and the presence of a parallel Kazakh-English corpus. The Kyrgyz language was considered herein because it belongs to the same Kypchak–Nogai subgroup of the Turkic languages as Kazakh, which enables the examination of the NMT complexity for the languages in a single subgroup. Furthermore, the Uzbek language was considered because it belongs to the Karluk subgroup of the Turkic languages, which enables the investigation of the NMT complexity for the languages in different Turkic language subgroups. All the three languages considered herein are low-resource languages.

When training NMT for these language pairs, the volume of the corresponding NMT dictionary rapidly increases; therefore, it requires excessive computer memory resources. The well-known approaches for text segmentation are BPE-based method (Senrich et al., 2016) and Morfessor (Creutz & Lagus, 2002), both of which are unsupervised and statistics-based methods. The advantage of these two methods lies in their universal applicability to different languages.

This study proposes a novel morphological segmentation method based on complete set of endings (CSE) (suffixes) of words in a language. The proposed CSE-based segmentation method can be applied to the agglutinative languages in the Turkic group. Furthermore, this study demonstrates the applicability of the proposed CSE-based morphological segmentation method to the Kazakh, Kyrgyz, and Uzbek languages and presents the results of computational experiments for the Kazakh language. This approach can be extended to the other languages in the Turkic language group.

The remainder of this paper is organised as follows. Section 2 provides an overview of the previous works conducted in this field. Section 3 describes the proposed CSE-based segmentation of words in the Kazakh, Kyrgyz, and Uzbek languages. Section 4 presents the experimental NMT results for the Kazakh–English language pair obtained using the proposed CSE-based segmentation method. Finally, Section 5 presents the conclusions and suggests the direction for future work.

2. Related work

The research works related to segmentation for NMT can be divided into those based on the BPE method, Morfessor, and finite-state transducers (Sennrich et al., 2016; Ataman et al.,2017; Sánchez-Cartagena & Toral, 2016).

Sennrich et al. developed methods that segment corpora into frequent sequences of characters (Sennrich et al., 2016). Specifically, the well-known byte-pair encoding (BPE) compression method was applied to English–German and English–Russian NMT systems in these methods. The authors adapted the BPE algorithm for the segmentation task to create open vocabulary. The advantage of using BPE-based method segmentation is that rare words are segmented into frequent subsequences, enabling the translations of unknown words to be built, which is one of the major goals of NMT. The BPE segmentation method is the dominant approach to subword segmentation [Tacorda et al., 2016].

The BPE-based method involves the splitting of words into different variations of word segments; however, this approach is not suitable for languages with rich morphologies, such as the Kazakh language. For example, in the learning phase, the words “жобалар”(projects), “жобасын”(project of), and “жобаның”(of project) are presented as “жоб алар</w>”(right segmentation is “жоба-лар”), “жобас ын</w>”(right segmentation is “жоба сын”), and “жобаның</w>”(without segmentation), respectively, by the BPE method. When BPE method is applied to files in the test phase, these words are not split, but are rather left as whole words. This is explained that whole words often have highest frequency than word segments, the experiments in Section 4 confirm this assumption, and therefore the vocabulary of BPE-based segmentation is more than that of morphological segmentation.

Tacorda et al. proposed the use of the controlled byte-pair encoding (CBPE) method for English–Filipino and Filipino–English translation (Tacorda et al., 2016). CBPE is used to recognise inflected words in morphologically rich languages. The authors compared the results of BPE and CBPE and concluded that both improve the bilingual evaluation understudy (BLEU) metric; however, with CBPE, the quality metric was improved slightly.

The use of BPE-based method has been considered in other researches based on Turkic languages. Ataman et al. predicted subword segments using an unsupervised morphology learning algorithm based on a prior morphology model (Ataman et al., 2017). They investigated morphological and BPE segmentation. Morphological segmentation was applied to the Turkish language; for fair comparison, only the source side was segmented. Their study presented two morphological segmentation methods, i.e. supervised and unsupervised. The supervised method maintained a full description of the morphological properties of subwords, whereas the unsupervised method was based on the Morfessor framework with category-based model averting. Experiments were performed separately using the BPE and developed methods, and the results showed that in comparison with the BPE segmentation, the developed methods improved the BLEU metric by 2.2.

Sánchez-Cartagena and Toral used a rule-based morphological analyser for Finnish language to separate words into root and inflection boundaries for vocabulary reduction for NMT (Sánchez-Cartagena & Toral, 2016). The authors combined an NMT system and phrase-based statistical machine translation (SMT) system enhanced with a neural language model. In SMT, the length of the segmented Finnish sentences is reduced by joining the most frequent sequences of morphs. BPE was used for the Finnish language because it has a more complex morphology than English. The authors concluded that combining BPE with morphological segmentation does not yield any clear improvement.

BPE performs the merging operations iteratively to find the most frequent character combination. BPE segmentation is conducted regardless of the morphology of the language. Therefore, the BPE output has no semantic meaning in languages with rich morphologies.

There are some other segmentation approaches based on Morfessor, which is an open-source software for unsupervised morphological analysis. Morfessor segments words according to their morphological structures. There are three main versions of Morfessor, which are Morfessor Baseline, Morfessor Categories-ML, and Morfessor Categories-MAP (Creutz, 2003; Creutz & Lagus, 2002, 2004, 2005; Creutz & Linden, 2004). Morfessor is a statistical morphological segmentation tool. It evaluates all possible ways by which a word can be split into two substrings, and the split with the highest probability is selected. Morfessor segments words according to their morphological structures, however, like N-gram models, it does not have a preference for infrequent words. Therefore, it suffers from the problem called out of vocabulary (OOV). This can be resolved with the use of the BPE (Banerjee & Bhattacharyya, 2018; 2007; Papli, 2017). Therefore, to avoid the appearance of several unknown tags and erroneous probabilistic segmentation, it was decided that BPE should be used for Kazakh–English and English–Kazakh language pairs. This choice was also influenced by the fact that BPE is a dominant approach in the domain of word segmentation.

At World Machine Translation (WMT), 2019, the Kazakh language was added to the translation tasks, and the translation of Kazakh to English was considered (Briakou and Carpuat, 2019; Casas et al., 2019; Kocmi & Bojar, 2019; Littell et al., 2019; Sánchez-Cartagena et al., 2019). Briakou and Carpuat applied transfer-learning technology to Kazakh–English and English–Kazakh translations (Briakou and Carpuat, 2019). As additional data for transfer learning, parallel corpora of Turkish–English were used because Kazakh and Turkish belong to the same language group. The researchers compared different configurations of BPE and soft decoupled encoding. The texts of Kazakh corpora were Romanised, and experiments were conducted with and without Romanisation. The dictionary volume significantly increased with Romanisation. With the BPE configuration, the BLEU score was improved by 0.20 with Romanisation, whereas that of the original text (in Cyrillic) was improved by 1.24.

Casas et al. mentioned the morphological complexity of the Kazakh and Russian languages (Casas et al., 2019). Russian–Kazakh SMT was used as a pivot system for English–Kazakh NMT. The researchers employed BPE with 10,000 comparative operations for each language in NMT, producing a BLEU score of 2.32. Kocmi and Bojar and Littell et al. considered the Russian language as the pivot language as well (Kocmi & Bojar, 2019; Littell et al., 2019).

Sánchez-Cartagena et al. demonstrated the morphological segmentation using Apertium and integrated the output of rule-based machine translation (Sánchez-Cartagena et al., 2019). They segmented a source text using a rule-based morphological analyser. If a word had no valid segmentation, many segmentation variants were generated as there were known suffixes that matched the word. After morphological segmentation, the BPE was applied to all the training data. For example, “университетінің”(of her/his university) has the morphological analysis result n. px3sp.gen. The proposed morphological segmentation split this term as “университет@@ інің”, whereas BPE left the word unchanged as “университетінің(of her/his university)”. Thus, the proposed morphological segmentation divides the given words into only two parts. In contrast, our morphological segmentation based on the complete set of Kazakh endings performs splitting into more than two parts and conducts segmentation by using ending types defined exactly according to the grammar: “университет@@ і@@нің”.

Thus, to improve the quality of NMT, appropriate segmentation and satisfactory volume of parallel corpora are required. To achieve these objectives, this study proposes a morphological segmentation approach based on the CSE-model and special stemming algorithm. Furthermore, it demonstrates the usability of the proposed approach to the Turkic languages for creating a complete set of language endings considering the Kazakh, Kyrgyz, and Uzbek languages as examples and presents the results of computational experiments, wherein the proposed morphological segmentation method was applied to the Kazakh language.

3. Description of the CSE-based morphological segmentation method

Morphology refers to the structures of words in terms of minimal semantic grammatical units known as morphemes. Morphemes are usually divided into two groups, i.e. stems and affixes; stem defines the basic meaning of a word, whereas affixes define the various forms of meaning of the word. Moreover, depending on the language type, i.e. agglutinative or inflectional (fusional), affixes can have either single or multiple grammatical meanings. Thus, for agglutinative languages, each affix has a single meaning, whereas for inflective languages, an affix can have several grammatical roles, such as case, gender, and number. For agglutinative languages, several affixes may be added to a stem, so that the word as a whole carries several grammatical meanings. In an agglutinative language, such a sequence of affixes after the stem is called the ending of the word. Tukeyev et al. defined the complete system of endings for the Kazakh language (Tukeyev et al., 2016).

The proposed study is novel because it demonstrates the applicability of the proposed CSE-based morphological segmentation method for the Turkic language family. Section 3.1 briefly shows the complete set of Kazakh endings, presents the CSE-based morphological segmentation model, and demonstrates its effectiveness for the agglutinative languages of the Turkic group, if a CSE-model of morphology is created for the language. Section 3.2 describes the morphological segmentation algorithm for words in the Kazakh language and its application to other languages in the Turkic group. Further, the possibility of constructing a CSE-model morphology for language of the Turkic group is demonstrated using Kyrgyz and Uzbek as examples.

3.1. CSE-model of morphology

This section analyses the morphology of the Turkic language group, more specifically, Kazakh, Uzbek, and Kyrgyz languages, and shows that the morphological structure of these languages enables the building of a CSE-model, which is essential for application of the proposed segmentation method.

We considered the Kazakh language, wherein the endings are divided into nominal endings (nouns, adjectives, and numerals) and verbal endings (verbs, participles, gerunds, mood, and voice).

The nominal endings in the Kazakh language have four types of base affixes, i.e. plural affixes (K), possessive affixes (T), case affixes (C), and personal affixes (J). These endings can occur in sequences of one, two, three, or four types of affixes, in order, as prescribed by the morphotactic of the language. All Turkic languages have these four types of base affixes.[2] Any ending comprising a single affix is semantically valid. The valid two-, three- and four-affix combinations are KT, TC, CJ, KC, TJ, and KJ; KTC, KTJ, TCJ, and KCJ; and KTCJ, respectively. Thus, the total number of ending combinations for words with nominal bases is 15 (= 4 + 6 + 4 + 1).

The system of endings for verbal bases in Kazakh includes endings type of verbs, participles, moods, and voices. The system of verb endings include the following affixes types: tenses (eight), persons (three), and negation. Thus, the total number of possible types of verb endings is 25 (= 8 × 3 + 1). The system of participle endings includes participle endings (R), plural endings (K), possessive endings (T), case endings (C), and personal endings (J). Possible semantically acceptable variants of participle endings types, verbal participles, moods, and voices are 11, 1, 6, and 8, respectively. Therefore, the total number of ending types for words with verbal bases will be 51 (= 25 + 11 + 1 + 6 + 8), whereas the total number of types of endings with nominal bases and types of endings of words with verbal bases is 66 (= 15 + 51).

According to these ending types, finite sets of endings were constructed for all the main parts of speech in the Kazakh language. The number of endings for parts of speech with nominal bases

Table 1. Total numbers of endings for verbal bases

Type of verbal base	Number
Verb	516
Participle	1,960
Adverbial	10
Mood	70
Voice	126
Derivative adverb or adjective	47
Total:	2,729

Table 2. Ending types

Designation in Kazakh	Ending type in Kazakh	Number	Ending type in Kyrgyz	Number	Ending type in Uzbek	Number
K	-lar, -dar, -tar, -ler, -der, -ter	6	-lar, -dar, -tar, -ler, -der, -ter, -lor, -dor, -tor, -lór, -tór, -dór.	12	-lar	1
T	-m, -ym, -im, -ń, -yń, -iń -ńyz, -ńiz, -yńыз, -ińiz -sy, si, -y, i	14	-m, -ym, -ım, -um, -úm, -ń, -yń, -ıń, -uń, -úń, -ńyz, -ńız, -ńyz, -ńúz, -ıńız, -uńuz, -úńúz, -uńyz, -byz, -bız, -buz, -búz, -ybyz, -ıbız, -ybúz, -úbúz, -ńap, -yńap, -ńyzdar, -yńyzdar, -sy, -y	32	-im, m,-ing,-ng -i, -si, -imiz, -miz, ingiz, -ngiz	10
C	-nyń, -niń, -dyń, -diń, -tyń, -tiń, -ǵa, -ge, -qa, -ke, -a,-e. -ny, -ni, -dy, -di, -ty, -ti, -da, -de, -ta, -te, -nda,-nde, -nan, -nen, -tan, -ten, -dan, -den, -men, -ben, -pen, -menen, -benen, -penen	36	-nyń, -dyń, -tyn, -ga, -ka, -ny, -dy, -ty, -da, -ta, -dan, -tan	12	-ning, -ga, -ni, -dan, -da	5

(*Continued*)

Table 2. (*Continued*)

Designation in Kazakh	Ending type in Kazakh	Number	Ending type in Kyrgyz	Number	Ending type in Uzbek	Number
J	-myn, -min, -byn, -bin, -pyn, -pin, -syn, -sin, -syndar, -sinder, -syz, -siz, -sizder, -syzdar, -byz, -biz, -pyz, -piz	20	-myn, -syn, -syz, -byz, -synar, -syzdar	6	-man, -san, -miz, -siz,	4

Table 3. Examples of types of endings in Kazakh, Kyrgyz, and Uzbek

Type	Example in Kazakh	Example in Kyrgyz	Example in Uzbek
CJ	qala+dan+myn	śaary+dan+myn	shahar+dan+man
TJ	mekteb+i+ min	mekteb+ı+ myn	maktab+i+ man
TC	dápter+im+nen	depter+ım+dan	daftar+im+dan
KJ	oquśy+lar+myz	okuchuu+lar+byz	talaba+lar+miz
KC	tereze+ler+den	tereze+ler+dan	deraza+lar+dan
KT	dápter +ler+im	depter+ler+ım	daftar+lar+im
KTJ	oquśy+lar+y+ myz	okuchuu+lar+y+ byz	talaba+lar+i+ miz
KTC	oquśy+lar+ymyz+dan	okuchuu+lar+ybyz+dan	talaba+lar+i+ miz+dan
KCJ	oıynśy+lar+dan+byz	oıunchu+lar+dan+byz	o'yinchi+lar+dan+miz
TCJ	kóśe+ńiz+den+min	kuchuu+nguz+dan+myn	ko'cha+ngiz+dan+man
KTCJ	oquśy+lar+ymyz+dan +syzdar	okuchuu+lar+ybyz+dan +syzdar	talaba+lar+imiz+dan+siz

Table 4. Kazakh endings with segmented suffixes (fragment)

Word ending	Sequence of suffixes
darymyzbenbiz	dar y myz ben biz
darymyzbenmin	dar y myz ben min
darymyzbensiz	dar y myz ben siz
darymyzbensiń	dar y myz ben siń
lermenbiz	ler men biz
lermenmin	ler men min
lermensiz	ler men siz
lermensiń	ler men siń
uim	u im
uiń	u iń
ge	ge
ǵa	ǵa
m	m
ń	ń
y	y

(nouns, adjectives, and numerals) is 1,998 and that with verbal bases is 2,729 (Table 1). Hence, there are a total of 4,727 endings for all parts of speech in the Kazakh language.

Israilova and Bakasova considered the formation of Kyrgyz morphology (2018). The Kyrgyz language has ending types similar to those in the Kazakh language. Kyrgyz has ending types E1, E2, E3, and E4, which correspond to K, T, C, and J, respectively, in Kazakh. The ending types in Kazakh, Kyrgyz, and Uzbek are listed in Table 2.

The numbers of base affixes of each type are different in Kazakh, Kyrgyz, and Uzbek; therefore, the number of possible endings in each of these languages will be different. Table 3 lists the examples of Kazakh, Kyrgyz and Uzbek endings for some ending types.

All agglutinative languages have strict systems of word formation and rules for affix conjunction. Kazakh, Uzbek, and Kyrgyz, like other Turkic languages, are grammatically similar in terms of the types of endings. Having studied the types of endings in Kyrgyz and Uzbek, the CSE-based method created for either of these languages could be applied to the segmentation algorithm based on the CSE-model of the Kazakh language. The morphological segmentation algorithms and models based on the CSE-model for the Turkic languages are discussed in the next section.

Table 5. Example of CSE-based morphological segmentation for the Kyrgyz word "śaarydanmyn" (I am from town)

Iteration	Splitting of the word on each iteration into stem and ending	Comments
1	śa-arydanmyn	Did not find any matches "arydanmyn" from the endings list on the first column of the table containing Kyrgyz endings
2	śaa-rydanmyn	Did not find any matches "rydanmyn" from the endings list on the first column of the table containing Kyrgyz endings
3	śaar-ydanmyn	Did not find any matches "ydanmyn" from the endings list on the first column of the table containing Kyrgyz endings
4	śaary-danmyn	Found a match with "danmyn" and split the word into stem and ending
Result: Receive word ending that are affixes segmented from the second column of the table containing Kyrgyz endings		Śaary@@ dan@@ myn

Table 6. Example of CSE-based morphological segmentation for the Uzbek word "daftarlarim" (My exercise books)

Iteration	Splitting of the word on each iteration into stem and ending	Comments
1	da-ftarlarim	Did not find any matches with "ftarlarim" from the endings list on the first column of the table containing Uzbek endings
2	daf-tarlarim	Did not find any matches with "tarlarim" from the endings list on the first column of the table containing Uzbek endings
3	daft-arlarim	Did not find any matches with "arlarim" from the endings list on the first column of the table containing Uzbek endings
4	dafta-rlarim	Did not find any matches with "rlarim" from the endings list on the first column of the table containing Uzbek endings
5	daftar-larim	Found a match with 'larim'and split the word into stem and ending
Result: Receive word endings that are affixes segmented from the second column of the table containing Uzbek endings		daftar@@ lar@@ im

3.2. CSE-based morphological segmentation algorithm

It is possible to create a CSE-based model for each agglutinative language in the Turkic group, as shown in the previous section. Therefore, the algorithm for the morphological segmentation of words will be the same for all languages in the Turkic group. This algorithm includes two stages: 1) splitting of stems and word endings and 2) segmentation of word endings into component suffixes.

1) The stem and ending of a word can be split using a stemming algorithm, which is also based on the use of the CSE-model of the agglutinative languages in the Turkic group. The

Table 7. Example of CSE-based morphological segmentation for the Kazakh word “qaladanmyn” (I am from town)

Iteration	Splitting of the word on each iteration into stem and ending	Comments
1	qa-ladanmyn	Did not find any matches with “ladanmyn” from the endings list on the first column of the table containing Kazakh endings
2	qal-adanmyn	Did not find any matches with “adanmyn” from the endings list on the first column of the table containing Kazakh endings
3	qala-danmyn	Found a match with “danmyn” and split the word into stem and ending
Result: Receive word ending that are affixes segmented from the second column of the table containing Kazakh endings		qala@@ dan@@ myn

Table 8. Number of sentences in the Kazakh–English parallel corpus from websites

Corpus name	Number of sentences
Akorda	40,661
Primeminister	6,680
mfa.gov	9,895
economy.gov	6,550
strategy2050	45,986
Total	109,772

Table 9. Segmented Kazakh–English parallel corpora vocabulary volume

Vocabulary type	Volume of vocabulary
BPE-based segmentation	27,533
CSE-based segmentation	12,794

proposed algorithm is a lexicon-free stemming algorithm based on the CSE of Kazakh language (Tukeyev & Turganbaeva, 2016). Herein, this algorithm is proposed for all Turkic language group. All the endings in the set of endings of the agglutinative languages in the Turkic group are divided into classes according to their length. The algorithm first looks for an ending of maximum length for the given word, which will be two symbols less than the length of the word; it is assumed that the stem cannot contain less than two symbols. The assumed ending of length (L) is searched for in an appropriate class of endings of L. If the ending is not in this class; then, the length of the assumed ending is decreased by one (accordingly, the assumed ending of the word is decreased by one symbol on the left side, and this symbol is added to the assumed stem of the word), and the received ending is searched for in the appropriate ending class until the stemming procedure is complete or the word has no ending.

In the following, $L(e)^{max}$ is the maximum length of endings in the set of endings for the language, e(w) is the ending of analysed word w, st(w) is the stem of w, L(w) is the length of w, L[e(w)] is the calculated length of the ending of w, and $L[e(w)]^{max}$ is the maximum length of the ending of analysed word w.

Table 10. Results of NMT training with different segmentation options for the Kazakh–English parallel corpus

Language pair	Training options: left side–right side of language pair	Dev, BLEU	Test, BLEU
English–Kazakh	No segmentation–CSE segmentation	18.2	17.9
English–Kazakh	No segmentation–BPE segmentation	18.2	17.7
Kazakh–English	CSE segmentation–No segmentation	25.4	25.3
Kazakh–English	BPE segmentation–No segmentation	25.4	24.8

The steps of the algorithm for splitting the stem and ending are as follows.

1. Determine L(w).

2. Determine the maximum length of an ending of the analysed word: L[e(w)]^{max} = L(w)—2, where 2 is the minimum length of the word stem.

3. If L(w) ≤ L(e)^{max}; then, assign to L[e(w)] the value of L[e(w)]^{max}: L[e(w)] = L[e(w)]^{max}.

4. Otherwise, assign to L[e(w)] the value of L(e)^{max}: L[e(w)] = L(e)^{max}.

5. Select ending e(w) of length L[e(w)] for analysed word w.

6. Check e(w) for matching with the endings from the list of endings of length L[e(w)]. If it matches, then the stem of the word is determined: st(w) = w—e(w). Go to step 9.

7. Otherwise, the calculated length of the ending of the analysed word is decreased by one: L [e(w)] = L[e(w)]—1.

8. If L[e(w)] < 1, then word w is without ending. Go to step 9. Otherwise, go to step 5.

9. End.

2) The word ending is segmented into its component suffixes using a single state transducer, presented as a table of endings with segmented suffixes (Table 4). The columns in this table list the endings of words of the agglutinative languages in the Turkic group and suffixes corresponding to each ending. Note that Table 4 is only a fragment of the common table of the Kazakh endings with segmented suffixes.

The algorithm for segmenting the ending of a word into its component suffixes involves two steps, i.e. finding the ending of the current word in the endings table of the agglutinative language and reading the sequence of suffixes corresponding to the ending of the word. Tables 5–7 present examples of morphological segmentation based on CSE for Uzbek, Kyrgyz and Kazakh. The corresponding table of endings must be used for each language, as presented in Table 4.

The algorithm described above involves separation of the stem and ending of a word without using a dictionary of stems of agglutinative languages in the Turkic group, which is known as lexicon-free algorithm.

4. Experiments and results

The proposed CSE-based segmentation method was applied to Kazakh–English NMT in a pre-processing phase. This section presents the results of the experiments comparing the proposed CSE-based segmentation and BPE-based segmentation. The choice of BPE is justified by the fact that it is the de facto standard for word segmentation in the domain of neural machine translation

(2019; Provilkov et al.). In addition, Morfessor requires the lexicon of morphemes for each language, which incurs additional expenses, while the BPE does not require such additional data.

The Kazakh–English parallel corpora were collected from the news sections of government agency websites (Table 8). The parallel corpora contained one sentence per line, which is tokenised with spaces. The collected corpora were assembled, cleaned, and aligned. The resulting Kazakh–English parallel corpus was pre-processed through tokenisation, normalisation, and shuffled.

The resulting parallel corpora of Kazakh–English comprised 109,772 sentences, where 80,000 sentences were utilised for training, and the remaining were divided into two sets, i.e. test and dev. The test and dev file included 15,000 and 14,772 sentences, respectively.

We used TensorFlow[3] “sequence to sequence” model in all the experiments and applied the following settings for the hyperparameters:

- 2-layer LSTM seq2seq model
- 1,024 dim hidden units
- 0.2 dropout
- bidirectional encoder (i.e. one bidirectional layer for the encoder)
- subword-option
- Adam optimiser
- 1.0 learning rate
- 100,000 training steps
- 128 epochs
- 50 max sequence length

We experimented with the standard hyper parameters by calibrating the number of units and concluded that training with 1,024 dim hidden units leads to an improvement in the quality of translation. During training, a model checkpoint was saved every 1,000 iterations. The duration of the training was 100,000 epochs. The training corpus on the Kazakh language side was segmented into stems and affixes for each word using the proposed CSE-based segmentation method. The stems and affixes of the Kazakh–English parallel corpora were separated by symbols @@, similar to that in the BPE-based segmentation. The NMT vocabulary was created based on the frequencies of occurrence of the words in the training file, wherein words that occurred only three or more times were included. The corresponding Kazakh vocabulary volumes are listed in Table 9.

The increase in the vocabulary size of the baseline NMT can be explained as follows. In the NMT with BPE, some words of the Kazakh text are not segmented, whereas in the NMT with CSE-segmentation, all the words with endings are segmented. For example, in the NMT with BPE, the word “Қазақстанның (of Kazakhstan)” was left as a whole without any segmentation, whereas, in the proposed CSE-based segmentation method, it was segmented as “Қазақстан@@ның”(of Kazakhstan). Therefore, the volume of the vocabulary in NMT with the proposed CSE-based segmentation is less than that in the NMT with BPE-based segmentation.

In the experiments, different segmentation options were used for training, as follows:

- English side: no segmentation; Kazakh side: CSE segmentation
- English side: no segmentation; Kazakh side: BPE segmentation
- Kazakh side: CSE segmentation; English side: no segmentation
- Kazakh side: BPE segmentation; English side: no segmentation

Table 10 lists experiment results.

The experimental results were evaluated using the BLEU metric. These values were not sufficient to indicate the good quality of the NMT. The main reason for this result is the unavailability of sufficient data for neural network training of this language pair. In actual practice, it is recommended that large parallel corpora be used for NMT training for adequate machine translation accuracy (Koehn & Knowles, 2017; Poncelas et al., 2018). However, for the Kazakh language, similar to the other languages in the Turkic family except Turkish, there are no sufficiently large parallel corpora.

In comparison with byte-pair encoding (BPE)-based segmentation, the proposed CSE-based segmentation increases the BLEU score of 0.5 and 0.2 points on average for Kazakh–English and English–Kazakh pairs, respectively. Furthermore, the proposed CSE-based segmentation reduces the vocabulary volume by a factor of more than two, i.e. from 28,000 to 13,000, which will be crucial when a larger volume of source corpora is available.

5. Conclusion

In this study, we developed CSE-based segmentation method and investigated its applicability to the Kazakh, Kyrgyz, and Uzbek languages. These languages, similar to all the Turkic languages, have four types of affixes for forming endings. Consequently, the proposed CSE-based segmentation approach could easily be applied to other languages in the Turkic language family. Computational experiments were conducted using the proposed CSE-based segmentation for NMT of the Kazakh language. In comparison with the BPE-based segmentation method, the proposed CSE-based segmentation method reduced the NMT vocabulary volume by more than twice and increased the BLEU score of 0.5 and 0.2 points on average for Kazakh–English and English–Kazakh pairs, respectively. When the size of the source parallel corpora was increased to improve the quality of NMT learning, the NMT vocabulary size reduction was significant. However, the small size of the available corpora in Turkic languages, other than Turkish, significantly limits the supplication of the proposed method.

In the future, corpora for other Turkic languages should be collected, and the CSE-model of morphology for other Turkic languages should be used in segmentation task for NMT of these languages and NMT transfer-learning experiments should be conducted for languages from other subgroups of the Turkic languages. Furthermore, the possibility of using CSE-model for lexicon-free stemming for informational retrieval in Turkic languages should be investigated, and CSE-model and lexicon-free stemming algorithm should be used for morphological analysis of Turkic languages and for pre-processing of corpora of the Turkic languages for tagging of corpora texts. In the future, the proposed segmentation method will be investigated and applied for processing the morphology of other agglutinative languages, such as Tatar, Karakalpak.

Acknowledgements
The authors would also like to express gratitude to Mikel L. Forcada for assisting us in writing this manuscript.

Funding
This work was supported by [Ministry of Education and Science of the Republic of Kazakhstan], followed by the grant number [AP05131415].

Author details
U. Tukeyev[1]
E-mail: ualsher.tukeyev@gmail.com
A. Karibayeva[1]
E-mail: a.s.karibayeva@gmail.com
Z h. Zhumanov[1]
E-mail: z.zhake@gmail.com
[1] Information Techology Faculty, Information Systems department, Al-Farabi Kazakh National University, Almaty, 050040, Kazakhstan.

Citation information
Cite this article as: Morphological segmentation method for Turkic language neural machine translation, U. Tukeyev, A. Karibayeva & Z h. Zhumanov, *Cogent Engineering* (2020), 7: 1856500.

Notes

1. http://www.languagesgulper.com/eng/Turkic.html
2. http://www.languagesgulper.com/eng/Turkic.html
3. https://github.com/tensorflow/nmt

References

Ataman, D., Negri, M., Turchi, M., & Federico, M. (2017). Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 331–342. https://doi.org/10.1515/pralin-2017-0031

Banerjee, T., & Bhattacharyya, P. (2018). Meaningless yet meaningful: Morphology grounded subword-level

NMT. *Proceedings of the second workshop on subword/character level models* (pp. 55–60

Briakou, E., & Carpuat, M. (2019). The University of Maryland's Kazakh–English neural machine translation system at WMT19. *Proceedings of the fourth conference on machine translation* (pp. 134–140), August 1-2.

Casas, N., Fonollosa, J. A. R., Escolano, C., Basta, C., & Costa-Jussà, M. R. (2019). The TALP-UPC machine translation systems for WMT19 news translation task: Pivoting techniques for low resource MT. *Proceedings of the fourth conference on machine translation (WMT19)*, August 1-2 (pp. 155–162).

Creutz, M. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. *Proceedings of the 41st annual meeting on association for computational linguistics*, 1, 280–287.

Creutz, M., & Lagus, K. 2002. Unsupervised discovery of morphemes. *Proceedings of the ACL-02 workshop on morphological and phonological learning*, 6, 21–30.

Creutz, M., & Lagus, K. 2004. Induction of a simple morphology for highly-inflecting languages. *Proceedings of the seventh meeting of the ACL special interest group in computational phonology: current themes in computational phonology and morphology* (pp. 43–51).

Creutz, M., & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, *1*(106–113), 51–59. https://tuhat.helsinki.fi/ws/portalfiles/portal/77193625/Creutz05akrr.pdf

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1), 1–34. DOI: 10.1145/1187415.1187418

Creutz, M., & Linden, K. 2004. *Morpheme segmentation gold standards for Finnish and English* (Tech. rep. A77). Publications in Computer and Information Science, Helsinki University of Technology.

Gallé, M. (2019). Investigating the effectiveness of BPE: The power of shorter sequences. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 1375–1381).

Israilova, N., & Bakasova, P. (2018). Morphological analyzer of the Kyrgyz language. *International conference on computer processing of Turkic languages Turklang-2018*, 2, 100–116.

Kocmi, T., & Bojar, O. (2019). CUNI submission for low-resource languages in WMT news 2019. *Proceedings of the fourth conference on machine translation (WMT19)*, August 1-2, pp. 234–240.

Koehn, P., & Knowles, R. Six challenges for neural machine translation. *Proceedings of the first workshop on neural machine translation*, 2017, Vancouver, pp. 28–39

Littell, P., Lo, C., Larkin, S., & Stewart, D. (2019). Multi-source transformer for Kazakh-Russian-English neural machine translation. *Proceedings of the fourth conference on machine translation (WMT19)*, August 1-2, pp. 267–274.

Papli, K. (2017). *Morpheme-aware subword segmentation for neural machine translation* [bachelor thesis]. University of Tartu Institute of Computer Science. p. 27.

Poncelas, A., Shterionov, D., Way, A., de Buy Wenniger, G. M., & Passban, P. (2018). Investigating backtranslation in neural machine translation. *21st annual conference of the european association for machine translation*, Association for Computational Linguistics, pp. 249–258

Provilkov, I., Emelianenko, D., & Voita, E. BPE-dropout: Simple and effective subword regularization. *The 58th annual meeting of the association for computational linguistics (ACL 2020)*, pp. 1882–1892.

Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., & Sánchez-Martínez, F. (2019). *The Universitat d'Alacant submissions to the English-to-Kazakh news translation task at WMT*.

Sánchez-Cartagena, V. M., & Toral, A. Abu-MaTran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. *Proceedings of the first conference on machine translation. ACL*, 2016.

Sennrich, R., Haddow, B., & Birch, A. 2016. Neural machine translation of rare words with subword units. *Proceedings of the 54th annual meeting of the association for computational linguistics*, 1, 1715–1725.

Tacorda, A. J., Ignacio, M. J., Oco, N., & Roxas, R. E. 2017. Controlling byte pair encoding for neural machine translation. *2017 international conference on Asian language processing*, 168–171.

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 N. Nguyen, L. Iliadis, Y. Manolopoulos, & B. Trawiński (Eds.), *Computational collective intelligence. ICCCI 2016. Lecture notes in computer science* (Vol. 9876, pp. 563–574). Springer.

Tukeyev, U. A., & Turganbaeva, A. (2016). Lexicon - free stemming for the Kazakh language. *Materials 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 technology*, Almaty, pp. 84–88. (In Russian)

***Cogent Engineering* (ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**