Neural machine translation system for the Kazakh language based on synthetic corpora

Ualsher Tukeyev[1], *Aidana* Karibayeva[2], and *Balzhan* Abduali[3]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

Abstract. The lack of big parallel data is present for the Kazakh language. This problem seriously impairs the quality of machine translation from and into Kazakh. This article considers the neural machine translation of the Kazakh language on the basis of synthetic corpora. The Kazakh language belongs to the Turkic languages, which are characterised by rich morphology. Neural machine translation of natural languages requires large training data. The article will show the model for the creation of synthetic corpora, namely the generation of sentences based on complete suffixes for the Kazakh language. The novelty of this approach of the synthetic corpora generation for the Kazakh language is the generation of sentences on the basis of the complete system of suffixes of the Kazakh language. By using generated synthetic corpora we are improving the translation quality in neural machine translation of Kazakh-English and Kazakh-Russian pairs.

1 Introduction

In spite of the success of neural machine translation (NMT) in languages with sufficient data resources, the lack of large parallel corpora represents a serious practical problem for many language pairs with low resources. Every task of processing a natural language requires enormous data for a particular problem under consideration. Also, neural machine translation requires a large database to create networks, namely the need to parallel the corpora. Unfortunately, there is a serious deficiency in the amount of data - parallel corpora in the Kazakh language. Therefore, the creation of parallel corpora for languages with low resources such as the Kazakh language is very important. Parallel corpora are the main sources of statistical and neural machine translation. In most cases, they are made of parallel translations performed by professional translators. It turned out that texts with parallel translation require considerable effort. The search and preparation of such data from the Internet sources is not an easy task either.

There have been several works that propose different ways for creating synthetic corpora. There are translations with back-translation, collecting pseudo-corpora, the bilingual phrase pair, creating synthetic sentences by using recurrent neural networks etc. Employing these methods we propose a method for the generation of synthetic corpora by using the complete set of suffixes of Kazakh language. Generation of sentences was developed through the substitution of different words in a certain part of speech. The novelty is using and improving the generated synthetic parallel corpora based on complete systems of Kazakh suffixes [11].

2 Related works

Philipp Koehn and Rebecca Knowle mentioned in their work [1] that the training of NMT with low-resource settings gives worse quality compared to with high-resource settings. They addressed the key issues and the importance of using neural networks for machine translation. Work [1] compared the output of SMT and NMT respectively for different language pairs.

Imankulova et al. considered Japanese-Russian language pair as low-resource and developed the method for filtering pseudo-corpus by back-translating and filtering a monolingual corpus in the target language for low-resource language pairs [2].

Zhang et al. (2016) proposed a method, which is named as sentence reordering method and self-learning algorithm for source side language [3].

Sennrich et al. (2015) improved the NMT system by using monolingual data by back-translation of it. They found that back-translation shows more effectiveness [4].

In [5] authors trained the NMT system for low-resource by exploiting monolingual data in the target language without changing the algorithm and architecture of it with mixing common corpora. This improved the BLEU metrics to 1.2 for English-Turkish and English-Romanian systems and authors found that the copying technique is effective both alone and combined with back-translation.

In [6] researchers used the bidirectional recurrent neural network to generate parallel sentences from Wikipedia for non-English languages to SMT and NMT systems. In this work, they received good metric values to

* Corresponding author: ualsher.tukeyev@gmail.com, a.s.karibayeva@gmail.com, balzhanabdualy@gmail.com

English-Tamil and English-Hindi, which belong to low-resources.

Gu et al. (2016) proposed the universal method for lexical and sentence-level representation [7]. They concluded that using a shared source language dictionary is not suitable for generalisation between zero-resource to high-resource languages.

Wang et al. presented the method using bilingual phrase pairs and monolingual data. Testing on short and long sentences for zero-resource languages gave good and worst results respectively [8].

The paper of Gökhan Doğru et al. described the creation of domain-specific parallel corpora by automatic and semi-automatic methods. By browsing through the corpora they collected about 6500 Turkish-English abstracts for medical purposes from the Internet. Authors concentrate attention more on collecting not evaluating translation systems [9].

Zaremoodi et al. experimented with various tasks, such as named-entity recognition, syntactic parsing and semantic parsing for English-Farsi and English-Vietnamese low-resource language pairs and received different metrics. The proposed method consisted in expanding recurrent blocks with recurrent units with multiple blocks [10].

3 Generation of synthetic corpora for the Kazakh language

3.1 Algorithm of generation of synthetic corpora for Kazakh

The main idea of the proposed method is that the generation of various variants of words is carried out by parts of speech and by suffixes for each word of a simple sentence of Kazakh. For other language pairs, the special rules for transformation of Kazakh to English and of Kazakh to Russian were employed [11]. For each type of suffix (template of morphological structure types of suffixes) of the Kazakh language, an equivalent grammatical structure logical template (pattern) in the target language (Russian and English) was constructed. Based on the grammatical structure of the logical template we built a template of programme structure for transfer of the morphological structure of words' suffixes into the equivalent grammatical structure of the target language.

Table 1. Logical templates for verbs Kazakh-English and Kazakh-Russian.

Languages	The tense of language and transliteration	Grammar structure for Kazakh, English, Russian
Kaz:	Жай нақ осы шақ	[PRN] [V+A(PresSm)+(Sg,Pl)+(P1,P2,P2v,P3)]
Eng:	Present Simple	[PRN] [V]
Rus:	Настоящее время	[PRN] [V+(Sg,Pl) +(P1,P2,P2v,P3)]
Kaz:	Жедел өткен шақ	[PRN] [V+A(PastOper)+(Sg,Pl)+(P1,P2,P2v,P3)]
Eng:	Past Simple	[PRN] [V + ed]
Rus:	Прошедшее время	[PRN] [V(Sg,Pl) +(P1,P2,P2v,P3)]

For example for “Мен келдім”, in Kazakh, there are two tenses. First one is Present Simple, the second one is Past Simple. Table 1 presents the logical templates for verbs, in addition showing what kind of suffixes for each language is selected.

The algorithm of generated synthetic corpora of Kazakh is presented below in a sequential.

- Step 1. Get the complete set of Kazakh suffixes
- Step 2. Select one of the Kazakh suffixes words with nominal bases.
- Step 3. Create the structure of simple sentences for source language.
- Step 4. Find with a long matching structure for simple sentences.
- Step 5. Select appropriate context for the chosen sentences.
- Step 6. Put the selected context in the files of each part of speech.
- Step 7. Create the same file for each part of speech with translated selected context for the target language.
- Step 8. Get parallel files with context.
- Step 9. Start automatic generation for sentences of different structure.
- Step 10. Collect parallel sentences from generated files.
- Step 11. Give the received parallel sentences to train NMT.

The detailed description of the algorithm on example will be considered below.

3.2 Detailed description of generation synthetic corpora based on complete set of suffixes

Generation of structure starts with determining of suffixes name. These structures strictly retain their structure.

The simple sentence structure in Kazakh language is the following:

pronoun verb
pronoun adverb verb
pronoun noun verb
pronoun adjective noun verb
pronoun noun noun verb
pronoun noun adjective noun verb
pronoun noun adverb adverb verb
pronoun noun adverb verb
adverb noun verb, pronoun noun verb
adverb noun verb, pronoun verb
pronoun noun adjective, adjective noun verb
pronoun adverb adjective, adjective noun verb
pronoun adverb adjective noun verb

This approach is based on a complete set of suffixes of Kazakh language. The creation of the complete set developed with Tukeyev and et al. (2016) and described in [11].

Kazakh words have nominal suffixes and verbal suffixes. The nominal suffixes consist of nouns, adjectives, numerals etc. They have four types of base affixes: plural affixes (K), possessive affixes (T), case affixes (C) and personal affixes (J), therefore, considering all types of base affixes placements variants, we get 64 possible variants. For example KT, TC, KTJ, TCJ, KTCJ, etc. The verbal stems include the following types of suffixes: verb suffixes, participle suffixes, verbal adverb suffixes, mood suffixes and voice suffixes, so the total number of verbal base suffixes is 59. The total number of the suffixes of words with nominal and verbal bases is 74. Based on this method we began with personal affixes for our experiment. The personal affixes have 8 types of suffixes.

Due to the need for big parallel corpora, the generated synthetic corpora have been created. This is created with the help of Kazakh structure sentences and Kazakh suffixes. In the creation of generated synthetic corpora, we firstly considered suffixes, which would then be added to form different words. When generating a sentence, not only the suffix is changed but the words change as well. Big corpora were compiled with the same structure, for example, Kazakh language has eight types of pronouns and then suffixes corresponding to pronouns are added to the verbs. Finally, the type of verb - negative or non-negative must be determined.

Based on the sentence structure such as “pronoun noun adverb adverb verb” Kazakh sentences were created, e.g. “Мен университетке бүгін ерте келдім”. Here a considerable attention is given to Kazakh suffixes, which in this situation took personal affixes. Then through the automatic generation the following structure of sentences with changing context of words was produced:

- Мен университетке бүгін ерте келдім
- Мен университетке бүгін ерте келмедім
- Мен университетке бүгін келдім.
- Мен университетке бүгін келмедім.
- Мен университетке ерте келдім.
- Мен университетке ерте келмедім.
- Мен бүгін ерте келдім.
- Мен бүгін ерте келмедім.
- Мен университетке келдім.
- Мен университетке келмедім.
- Мен ерте келдім
- Мен ерте келмедім.
- Мен бүгін келдім.
- Мен бүгін келмедім.
- Мен келдім.
- Мен келмедім.

As it was seen in similar cases, the obtained sentence is structurally different: the pronouns and ”verbs + Kazakh suffixes” remain the main ones, while other parts of speech may change or not be included. One structure has produced 16,128 synthetic sentences. A selection of possible contexts and suffixes are presented in the following tables.

Table 2. Appropriate context for creating structural sentences like «Мен университетке бүгін ерте келдім» in Kazakh

Pronoun	Noun	Adverb	Adverb2	Verb	Suffixes
Мен Сен Сіз Ол Біз Сендер Сіздер Олар	университетке мектепке жұмысқа сабаққа үйге колледжге балабақшаға дүкенге супермаркетке паркке стадионға	бүгін кеше таңертен кешке азанда түсте	ерте кеш жүгіріп кешігіп асығып ойнап асықпай кешікпей тез баяу жылдам	кел келме	-дім -дің -діңіз -ді -дік -діңдер -діңіздер -ді

Table 3. Appropriate context for creating structural sentences like «I come to university today early» in English

Pronoun	Noun	Adverb	Adverb 2	Verb	Suffixes
I You You He We You You They	to university to school to work to the lesson home to college to the kindergarten to the store to the supermarket to the park to the stadium	today yesterday in the morning in the evening in the morning at lunch	early late running late hurry playing slowly without delay fast slowly fast	come did not come	

Table 4. Appropriate context for creating structural sentences like «Я пришёл в университет сегодня рано» in Russian

Pro_noun	Noun	Adverb	Adverb2	Verb	Suffixes

Я Ты Вы Он Мы Вы Вы Они	в университет в школу на работу на занятие домой в колледж в детский сад в супермаркет в парк в стадион	сегодня вчера утром вечером утром в полдень	рано поздно бегая опоздав поспешно играя спокойно не опаздывая быстро медленно быстро	приш не приш	-ёл -ёл -ли -ёл -ли -ли -ли -ли

Table 2 presents the appropriate context of words for the structure “pronoun noun adverb adverb verb” in the Kazakh language. The first column is a pronoun. The pronoun in Kazakh language has 8 types and respectively the last column has 8 types for those suffixes. The personal suffixes have 4 types: -дім, -дым, -тым, -тім, in this case, we have “-дім”, because of the Kazakh language has the law of harmony, also imposed on the verb column. The other columns are nouns, adverbs and different adverbs. This part of speech fills many contexts of words. It is independent of the pronoun, and in the end, these columns contain empty lines. The empty lines result from the fact that the generating sentences do not take nouns or adverbs. Therefore the produced sentence structure consists of a pronoun and verb+Kazakh suffixes. Table 3 and Table 4 show the same data as Table 1, for English and Russian languages respectively.

In the case of Russian and English exactly the same files as Kazakh are created, in Kazakh language files – if the file contains verbs, another file will contain a different part of speech, for example nouns, adverbs and etc. Subsequently, these files will provide the base for forming sentences through the automatically generated system. Three tables show one situation for each language, and there should be parallel data in the files, then by changing the code inside the program parallel corpora are obtained. For instance, the combination in the programme for the sentence “Мен университетке бүгін ерте келдім” will be w "Pronoun + noun + adverb + adverb2 + [verb + suffix]", in the case of Russian the program will give the combination as "Pronoun + [verb + suffix] + noun + adverb + adverb2". As a result, the “verb+ suffixes” goes to the second position in the Russian language. The suffixes for Russian are also written separately in the file, and these suffixes, like in the Kazakh part, are added to the verbs.

Exactly the same procedure is followed in the case of English. However, in the English part, the file with suffixes is empty, due to the fact that English does not have the suffixes.

The same files will be created for other sentences in order to automatically generate multiple sentences, the sentences like: “Бүгін қар жауған соң, мен университетке бармадым” that are translated as “Because of the snow, I didn’t go to university”, “Я не пошел в университет, потому что сегодня был снег,” the former being the translation to English, the latter to Russian. By changing the structure and the context we obtain 439,176 generated synthetic parallel corpora.

4 Training neural machine translation of Kazakh language

TensorFlow was used as a tool for training. Prior to training, the following data must be prepared:

- train file (for Kazakh and English/Russian)
- testdev file (for Kazakh and English/Russian)
- test file (for Kazakh and English/Russian)
- vocabulary (for Kazakh and English/Russian)

Before preparing datasets for training, the sentences were shuffled to avoid the duplication of similar suffixes. The tests were carried out on the total of 20,000 and 10,000 sentences respectively for two types of testing.

The testing sets were divided into two parts: for dev and test. The first part of testing incorporated approx. 10% of training data, while the second part took 10% of training data. The training file consisted of the remaining 80% of the data. The vocabulary consisted of the determined context words that are used in the generated synthetic corpora. After preparing the data, the training was started.

5 Results

The experiments considered 4 language pairs, namely from English to Kazakh and Russian, in the case of Kazakh, from Kazakh to English and Russian. The trained set consisted of 0.4M parallel sentences BLEU metrics. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. The quality is assessed by the correlating the machine and human outputs [12].

Table 5. Received BLEU metrics by using generated synthetic corpora.

Language systems	BLEU
Russian-Kazakh	15.3
Kazakh-Russian	14.4
English-Kazakh	15.7
Kazakh-English	16.4

The examples of translation for each pair and direction are shown below.

for Russian-Kazakh NMT

src: Он не пришёл в школу утром.
ref: Ол мектепке азанда келмеді.
nmt: Ол университетке таңертең келмеді.

for Kazakh-Russian NMT

src: Біз университетке бүгін асықпай келдік.
ref: Мы пришли в университет сегодня спокойно.
nmt: Мы пришли в парк сегодня спокойно.

for English-Kazakh NMT

src: You came to university in the morning slowly.
ref: Сіздер университетке таңертең баяу келдіңіздер.
nmt: Сіз университетке таңертең баяу келдіңіз.

for Kazakh-English NMT

src: Мен университетке таңертең баяу келдім.
ref: I came to university in the morning slowly.
nmt: I came to university in the morning slowly.

As seen from the results shown above the main part of sentences is translated correctly, but in some case the system gave wrong result in finding a pronoun in the second person and in parsing noun as seen from translation result obtained from NMT in line 3. Further works plan to reduce this problem by increasing the number of sentence structures in the training set.

Conclusion

The approach proposed in this work solves the problem of small volumes of parallel corpora for a low-resource Kazakh language. The novelty of the proposed approach is based on the generation of synthetic corpora for pairs of Kazakh-English and Kazakh-Russian languages using the complete system of Kazakh language suffixes. In the future, a further increase in the volumes of parallel corpora for the Kazakh-English and Kazakh-Russian pairs are planned, and the expansion of this technology for generating synthetic corpora for other low-resource languages.

References

1. P.Koehn, R.Knowles, Proc. First Workshop on NMT - (2017), p. 28–39 (available at http://www.aclweb.org/anthology/ W17-3204)
2. A.Imankulova, T.Sato, M.Komachi, proc. WAT2017, (2017), p. 70–78
3. J.Zhang, Ch.Zong, Proc. 2016 Conference on Empirical Methods in Natural Language Processing (2016), p. 1535–1545
4. R.Sennrich, B. Haddow, A. Birch. (available at https://arxiv.org/abs/1511.06709)
5. A. Currey, A. Valerio Miceli Barone, K. Heafield, Proc. 2nd Conference on Machine Translation, Association for Computational Linguistics. (2017), p. 148–156, (available at http://www.aclweb.org/anthology/W17-4715)
6. S. Harsha Ramesh, K. Prasad Sankaranarayana, (available at https://arxiv.org/abs/1806.09652)
7. J. Gu, H. Hassan, J. Devlin, Victor O.K. Li. (available at https://arxiv.org/abs/1802.05368)
8. Y. Wang, Y. Zhao, J. Zhang, Ch. Zong, Zh. Xue. Towards Neural Machine Translation with Partially Aligned Corpora National Laboratory of Pattern Recognition, available at https://arxiv.org/abs/1711.01006)
9. G. Doğru, A. Martín-Mor, A. Aguilar-Amat, Proc. LREC 2018, (2018)
10. P. Zaremoodi Wray Buntine Gholamreza Haffari, Monash University, (available at http://aclweb.org/anthology/P18-2104)
11. U. Tukeyev, A. Sundetova, B. Abduali, Zh. Akhmadiyeva, N. Zhanbussunov, Proc 8th International conference ICCCI 2016. – (2016), p. 563-574.
12. Wikipedia. Bleu metrics, (available at https://en.wikipedia.org/wiki/BLEU)