SOLUTION OF THE PROBLEM OF UNKNOWN WORDS UNDER NEURAL MACHINE TRANSLATION OF THE KAZAKH LANGUAGE

Aliya Turganbayeva[1][0000-0001-9660-6928] and Ualsher Tukeyev[2][0000-0001-9878-981X]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2] Al-Farabi Kazakh National University, Almaty, Kazakhstan
turganbayeva16@gmail.com, ualsher.tukeyev@gmail.com

Abstract. The paper proposes a solution to the problem of unknown words for neural machine translation. The proposed solution is shown by the example of a neural machine translation of a Kazakh-English language pairs. The novelty of the proposed method is the search for unknown words in the dictionary of a trained model of neural machine translation. A dictionary of synonyms is used to search for words that are similar in meaning to the unknown words, that was found. Moreover, the found synonyms are checked for the presence in the dictionary of a trained model of neural machine translation. After that, a new translation of the generated sentence of the source language is performed. The base of words-synonyms of the Kazakh language, consisting of different parts of speech, is collected. The total number of synonymous words collected is 1995. Software solutions to the unknown word problem have been developed in the python programming language.

Keywords: Neural machine translation, unknown words, Kazakh language.

1 Introduction

The quality of a neural machine translation substantially depends on solving the problem of unknown words. This problem is associated with the concepts of “in-domain” (in the field) and “out-of-domain” (outside the field). By “in-domain” domain is meant a selection of source data on which neural machine translation is trained. If during testing or during a real translation, words that did not appear in the "in-domain" come across, then these will be unknown words. Some machine translation systems leave these unknown words untranslated, either replace them with the abbreviation "UNK", or translate them with words that are close in meaning. Accordingly, the last decision, namely, finding a word that is close in meaning, is also a difficult task.

2 Related works

To solve the problem of unknown words in the literature, several approaches have been proposed that can be divided into three categories. The first category of approaches focuses on improving the speed of calculating the output of softmax so that it can support a very large vocabulary. The second category uses information from the context. In particular, in relation to the problem of machine translation in [1], the system learns to indicate some words in the original sentence and copy them to the target sentence. In [2], when setting up the answer to a question in context, placeholders for named objects were used. The third category of approaches changes the input / output unit itself from words to a lower resolution, such as characters [3] or byte codes [4]. Although this approach has the main advantage that it can suffer less from the problem of unknown words, learning usually becomes much more difficult as the length of the sequences increases significantly.

In traditional machine translation, many off-vocabulary words still remain during testing, and they greatly reduce translation performance. In [5], when solving the problem of extra-vocabulary, attention is paid to how to correctly translate extra-vocabulary words. For this, additional resources such as comparable data and thesaurus of synonyms are used. One notable exception is the work [6; 7], which also focuses on the syntactic and semantic role of off-vocabulary words and suggest replacing off-vocabulary words with similar words during testing.

An effective method for solving the problem of unknown words is proposed and implemented in [1]. The authors trained the NMT system on data that was supplemented by the output of the word alignment algorithm, which allowed the NMT system to display for each out-of-dictionary word in the target sentence the position of its corresponding word in the original sentence. This information was later used in the post-processing phase, which translates each out-of-dictionary word using a dictionary.

In [8], a method is proposed for processing rare and unknown words for models of neural networks using the attention mechanism. Their model uses two softmax layers to predict the next word in conditional language models: one predicts the location of the word in the original sentence, and the other predicts the word in the short list dictionary. At each time step, the decision about which softmax layer to use is adaptively taken by the multilayer perceptron, which is context-specific.

To solve the problem of unknown words, in [9] a replacement-translation-recovery method is proposed. At the substitution stage, rare words in the test sentence are replaced by similar dictionary words based on the similarity model obtained from monolingual data. At the stages of translation and restoration, the sentence will be translated with a model trained in new bilingual data with the replacement of rare words, and finally, the translations of the replaced words will be replaced by the translation of the original words.

In [10], a method for processing unknown words in the NMT is proposed, based on the semantic concept of the source language. First, the authors used the semantic concept of the semantic dictionary of the source language to find candidates for dictionary words. Secondly, they proposed a method for calculating semantic similarities by

integrating the source language model and the semantic concept of the network to get a better word replacement.

3 Obtained results

Models, algorithms and software solutions have been developed for the task of unknown words in the neural machine translation of the Kazakh language. A technology (method) for solving the problem of unknown words in the neural machine translation of the Kazakh language has been developed, which consists of the following steps: 1. Segmentation of the source text of the Kazakh language. 2. An algorithm for searching for unknown words in the dictionary of a trained model of neural machine translation for the Kazakh-English language pairs has been developed. 3. For each unknown word in the source text of the test corpus, a search is made for its synonyms in the dictionary of synonyms. 4. The found unknown words are replaced with synonymous words. 5. The next step is the machine translation of the modified source text. The base of words-synonyms of the Kazakh language, consisting of different parts of speech, is collected. The total number of synonymous words collected is 1995. Each word contains at least one synonym word, maximum 35 synonyms. An algorithm is developed for searching unknown words in the dictionary of a trained model of neural machine translation for a Kazakh-English language pairs. Software solutions to the unknown word problem have been developed in the Python3 programming language.

The novelty of the proposed technology for solving the problem of unknown words in the neural machine translation of the Kazakh language is the proposed algorithm for searching for unknown words in the dictionary of the trained model of neural machine translation for the Kazakh-English language pairs. To find words that are close in meaning to an unknown word, a dictionary of synonyms is used. In this case, an additional check is made for the presence of this synonym word in the dictionary of the trained model. These steps of the proposed technology for solving the unknown word problem are essentially actions that convert the extra-dictionary words of the source text into dictionary words, i.e. out-of-domain words are converted to in-domain words.

4 Experimental part

4.1 Training Data

The model was trained on parallel buildings with a volume of 132,983 (files: origtrain.kaz, origtrain.eng) and 135,000 sentences (files: train.kaz, train.eng). The test set is: for a case with a volume of 132,983 sentences - 7,868 parallel aligned sentences (files: kazen-test1.kaz, kazen-test1.eng, kazen-test2.kaz, kazen-test2.eng), and for a case with a volume 135,000 offers - 5,000 parallel aligned offers (files: test1.kaz, test1.eng, test2.kaz, test2.eng). The dictionary is created from frequently used words in cases (occurring more than 3 times). The dictionary for the origtrain corpus is 16 878 words in Kazakh and 19 124 words in English (files: origvocab.kaz,

origvocab.eng), and for the train corps 48 154 words in Kazakh and 20 957 words in English (files: vocab.kaz, vocab.eng).

To evaluate the results of the translation, the BLEU score was used.

4.2 Training Details

Firstly, the source text of the Kazakh language is segmented. As a model for text segmentation, the model of the complete set of Kazakh endings are taken. A program for segmenting the words of the Kazakh language based on the complete set of endings of the Kazakh language has been developed. The developed segmentation program differs from analogs in that it divides the word, proceeding not from statistical calculations, but from the grammatical features of the word. This allows you to take into account the meaning of morphemes during segmentation. According to the results of segmentation, the words are divided into morphemes with a unique set of characters "@@". Desegmentation is done by replacing a unique character set with an empty character. Then, unknown words are searched in the dictionary of the trained model of neural machine translation for the Kazakh-English language pairs. At the next stage, for each unknown word in the source text of the test corpus, a search is made for its synonyms in the dictionary of synonyms. The found unknown words are replaced with synonyms. To do this, the following sequential scheme for finding a substitute synonym is made: firstly, the first word is taken from the list of synonyms, then it is checked for presence in the trained dictionary. If this synonym word is present in the trained dictionary, then the unknown word is replaced by this synonym, whereas if this synonym word is not present in the trained dictionary, then the next synonym word is taken and the procedure is repeated until a suitable synonym is found in the trained dictionary. In case, no synonym word is present in the trained dictionary, so unknown word does not change.

At the next step, the modified source text translated by machine translation.

To do this, firstly, you should activate the TensorFlow virtual environment:

```
$ source ~/tensorflow/venv/bin/activate
```

Next, start the model training process:

```
$ python -m nmt.nmt --attention=scaled_luong --
subword_option=bpe --src=kaz --tgt=eng --
vocab_prefix=/media/gpu2/data/exp10-unkword/mo-kaz-
eng/origvocab --train_prefix=/media/gpu2/data/exp10-
unkword/mo-kaz-eng/origtrain2 --
dev_prefix=/media/gpu2/data/exp10-unkword/mo-kaz-
eng/kazen-test1 --test_prefix=/media/gpu2/data/exp10-
unkword/mo-kaz-eng/kazen-test2 --
out_dir=/media/gpu2/data/exp10-unkword/unk-
models/model02-kaz-eng --num_train_steps=100000 --
steps_per_stats=100 --num_layers=2 --num_units=128 --
dropout=0.2 --metrics=bleu | tee /media/gpu2/data/exp10-
unkword/unk-models/logmodel02-kaz-eng.txt
```

Listing 1. Script to start the model learning process

Here: origvocab is a dictionary in Kazakh (English), origtrain2 is a segmented corpus in which replacements with synonyms are made, kazen-test1 and kazen-test2 are a test set of sentences, model02-kaz-eng is a trained model.

4.3 Learning outcomes using the proposed technology (method) for solving the problem of unknown words in a neural machine translation of a Kazakh-English languages pairs

In Table 1 below provides a description of the source data for the training and testing of the NMT of the Kazakh-English language pairs using the proposed technology (method) for solving the unknown word problem. Table 2 presents estimates of the machine translation of the Kazakh-English language pairs of the basic version without using the version using the proposed method for solving the unknown word problem.

Table 1. Initial data for training and testing of the NMT of the Kazakh-English language pairs without using and using the proposed technology (method) for solving the unknown word problem.

Parallel corpora with a volume of 132,983 sentences (Kazakh-English)		Parallel corpora with a capacity of 135,000 sentences (Kazakh-English)	
origtrain.kaz, origtrain.eng	132 983	train.kaz, train.eng	135 000
Test data 1: kazen-test1.kaz, kazen-test1.eng	3 000	Test data 1: test.kaz, test.eng	3 500
Test data 2: kazen-test2.kaz, kazen-test2.eng	4 868	Test data 2: test2.kaz, test2.eng	1 500
Vocabulary of Kazakh (words): origvocab.kaz	16 878	Vocabulary of Kazakh (words): vocab.kaz	48 154
Vocabulary of English (words): origvocab.eng	19 124	Vocabulary of English (words): vocab.eng	20 957

The model was trained on parallel buildings with a volume of 132,983 (files: origtrain.kaz, origtrain.eng) and 135,000 sentences (files: train.kaz, train.eng). The test set is: for a case with a volume of 132,983 sentences - 7,868 parallel aligned sentences (files: kazen-test1.kaz, kazen-test1.eng, kazen-test2.kaz, kazen-test2.eng), and for a case with a volume 135,000 offers - 5,000 parallel aligned offers (files: test1.kaz, test1.eng, test2.kaz, test2.eng). The dictionary is created from frequently used words in cases (occurring more than 3 times). The dictionary for the origtrain corpus is 16 878 words in Kazakh and 19 124 words in English (files: origvocab.kaz,

origvocab.eng), and for the train corps 48 154 words in Kazakh and 20 957 words in English (files: vocab.kaz, vocab.eng).

Table 2. Estimates of machine translation of the Kazakh-English language pairs of the basic version without using and version using the proposed method for solving the unknown word problem.

Corpora volume	BLEU without preprocessing	BLEU after preprocessing
Parallel corpora with a volume of 132,983 sentences (Kazakh-English)	8.9	8.9
Parallel corpora with a capacity of 135,000 sentences (Kazakh-English)	13.2	13.6

By calling inference seq2seq tenserflow for small texts, you can get translation results. We used inference to check how our approach reduces the number of unknown words. The command to call inference seq2seq tenserflow:

```
python -m nmt.nmt \
--out_dir=/media/gpu2/data/Mikelscorpora/wmtexperiment/
model1-kaz-eng \
--inference_input_file=/media/gpu2/data/exp-06-
2019/matin41.kaz \
--inference_output_file=/media/gpu2/data/exp-june-
2019/infer-results/output_infer4
```

Table 3. Results through inference of machine translation of the Kazakh-English language pairs of the basic version without using and version using the proposed method for solving the problem of unknown words.

Text volume	Number of unknown words in the text without preprocessing	Number of unknown words in the text after preprocessing
Text with volume 13 sentences	20	14
Text with volume 25 sentences	35	34
Text with volume 36 sentences	23	19
Text with volume 275 sentences	344	336

The application of this technology does not provide such a significant improvement, since only synonyms are used for rare (unknown) words and it may be that synonyms themselves are rare words or rare (unknown) words do not have synonyms.

5 Future work

In the future, it is planned to apply statistical methods to determine the position of unknown words in the source text in the direction of research , research on statistical methods to determine related words for unknown words. It is also planned to use the word2vec model to replace rare words with their frequently occurring words that are close to meaning.

6 Acknowledgments

This work was carried out under grant No. AP05131415 “Development and research of the neural machine translation system of Kazakh language”, funded by the Ministry of Education and Science of the Republic of Kazakhstan for 2018-2020.

References

1. Luong, M.T., Sutskever, I., Le, Q.V., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 11-19 (2015).
2. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, Ph.: Teaching machines to read and comprehend. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 1693-1701 (2015).
3. Generating sequences with recurrent neural networks, https://arxiv.org/pdf/1308.0850.pdf, last accessed 2019/08/27.
4. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1715-1725 (2016).
5. Marton, Y., Callison-Burch, Ch., Resnik, Ph.: Improved statistical machine translation using monolingually-derived paraphrases, In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language, pp. 381-390 (2009).
6. Zhang, J., Zhai, F., Zong, Ch.: Handling unknown words in statistical machine translation from a new perspective. In: Proceedings of the First CCF Conference Natural Language Processing and Chinese Computing, pp. 176-187 (2012).
7. Zhang, J., Zhai, F., Zong, Ch.: A substitution-translation-restoration framework for handling unknown words in statistical machine translation. In: Journal of Computer Science and Technology 28(5), 907–918 (2013).
8. Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., Bengio, Y.: Pointing the unknown words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 140-149 (2016).
9. Li, X., Zhang, J., Zong, C.: Towards zero unknown word in neural machine translation. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2852-2858. AAAI Press (2016).
10. Li, Sh., Xu, J., Miao, G., Zhang, Y., Chen, Y.: A Semantic Concept Based Unknown Words Processing Method in Neural Machine Translation. In: Proceedings of the 6th CCF International Conference on Natural Language Processing, pp. 233-242. NLPCC (2017).