Multilingual Neural Machine Translation System for 7 Turkic-Russian Language Pairs

Aidar Khusainov[1], Dzhavdet Suleymanov[1]

[1] *Tatarstan Academy of Sciences, Kazan, Russia*

Abstract

This article presents the results of experiments on the use of various methods and algorithms in creating machine translation systems for 7 Russian-Turkic language pairs. We proposed a semi-automatic procedure and created parallel corpora with a total volume of about 7.9 million sentence pairs. As a basic algorithm, we used a neural network approach based on the Transformer architecture. For the first time experiments were conducted for the Turkic languages on the use of transfer learning based on united Russian-Turkic parallel corpus. Experiments show that the systems fine-tuned on the base multilingual system are superior in quality to the basic Russian-Turkic translators. As the results of our last experiment, we have shown that a single multilingual model trained on the united Russian-Turkic corpus with additional language tags can show results comparable to fine-tuned models.

Keywords

multilingual machine translation, transfer learning, Turkic languages

1. Introduction

The field of machine translation has changed significantly in recent years with the development of neural network technologies. The development of new methods and algorithms for training models, the creation of large parallel and monolingual text corpora, as well as an increase in computational capabilities, have led to several results: the quality of machine translation for major world languages (for example, for the English-Chinese language pair for the news domain were declared to achieve the level of human translation quality [1]); machine translation systems have been developed for many low-resource languages.

Up to a certain amount of training data, the neural MT approach shows a lower quality than other statistical algorithms [2]. Therefore, the development of neural network machine translation systems for low-resource language pairs is still associated with the problem of the parallel corpora creation of the required size.

In this paper, we present the results of work aimed at using an approach to building a set of MT systems for 7 Russian-Turkic language pairs: Russian-Tatar, Russian-Kazakh, Russian-Chuvash, Russian-Bashkir, Russian-Crimean Tatar, Russian-Kirgiz, and Russian-Uzbek. Section 2 of this article presents the results of work on parallel corpora creation for the declared language pairs, as well as plans for using a rule-based approach to unify the collected multilingual parallel corpora (based on the structural-functional model of Turkic morphemes [3]). Section 3 describes the technologies used for creating neural MT systems. The "Experiments" section contains the results of evaluating the quality of the created models for each of the language pairs.

2. Turkic-Russian parallel text corpora

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL: khusainov.aidar@gmail.com (A. 1); dvdt.slt@gmail.com (A. 2)
ORCID: 0000-0002-7763-1420 (A. 1)

The data collection is the key stage in the creation of neural machine translation systems for low-resource language pairs. We have used several approaches to solve this problem. First of all, we formed a list of main sources of parallel information for all declared language pairs: news and websites of government organizations, translated books, already existing corpora. It should be noted that for different language pairs, different types of sources contain more data. For example, for the Crimean Tatar language, the main source of parallel texts at the moment is translated books, for Kazakh and Chuvash – existing corpora, for all other languages of the project (Kirgiz, Bashkir, Tatar, and Uzbek) – bilingual Internet sites.

To carry out the process of collecting data from websites, we have developed and implemented a semi-automatic data processing process. It included an initial manual search for potential Internet resources, an analysis of the site structure, a check for the presence of sitemap files (sitemap.xml), and a search for ways to automatically identify pairs of translated pages. The next step was to create a list of URLs to download text data from. The download procedure was carried out using the Trafilatura tool [4], which showed the efficiency of extracting basic text information from the web page for all analyzed websites, except for the official website of the Ministry of Justice of the Kyrgyz Republic (for this site, the basic information extraction tool was developed separately). The loaded text materials were processed using the razdel tool [5]; as a result of processing, each line of the text file contained a separate sentence. The last step at this stage was filtering and unification of auxiliary and non-printing elements ("°", "■", dashes, hyphens).

The next stage in the creation of parallel corpora was the stage of documents and sentences alignment. All uploaded and processed text documents have been converted to a WARC file format, which is the standard format for web archives. Depending on the availability of machine translation systems for a particular language, we used one of two approaches to align documents and sentences:

1. In the case of the existence of a machine translation system for a specific language pair, all documents on the website were translated from one language to another, the language with a smaller amount of data on this site was selected as the source language for translation. For languages other than Crimean Tatar, Tatar, and Bashkir, Yandex.Translate and Google Translate systems were used as translation systems. For the Tatar language, we used the Tatsoft NMT [6]. For Bashkir – the Tatar translator Tatsoft with preliminary processing of Bashkir texts (processing included the conversion of Bashkir letters, which are absent in the Tatar language, to the closest Tatar counterparts). Such a sequential "Bashkir-Tatar-Russian" transformation and translation procedure showed a sufficient quality of work for the task of aligning documents and sentences.
2. There are no machine translation systems available for the Crimean Tatar language, so we prepared and used a bilingual lexicon to search for pairs of documents and segments. At this stage, some parts of the Bitextor system [7] and the Bleualign tool [8] were used.

The last step of the corpus creation process involved the creation and application of algorithms for removing duplicate sentences. The developed algorithms made it possible to carry out the data collection process for all 7 language pairs. Current results indicate that the amount of collected data can allow the construction of basic machine translation systems for most of the selected language pairs.

In total, we prepared a corpus of more than 7.9 million sentence pairs. A summary of the collected data by language is presented in Table 1.

Table 1
Amount of Collected Parallel Data

Language	Data sources	Sentences collected	Russian part	Turkic part
Kyrgiz	8 web-sites: sti.gov.kg, www.kenesh.kg, minjust.gov.kg, novosti.kg, edu.gov.kg, mineconom.gov.kg, med.kg, ru.sputnik.kg; corpus: JW300	418 000 pairs	6.4 mln words	6.1 mln words
Bashkir	7 web-sites: bash.news, ufacity.info, glavarb.ru, bashinform.ru, bashdram.ru, house.bashkortostan.ru, pravitelstvorb.ru; JW300 corpus	352 000 pairs	5.2 mln words	4.9 mln words

Tatar	3 web-sites: tatar-inform.tatar, tatarstan.ru, kiziltan.rbsmi.ru; JW300 corpus; private corpora	2 mln pairs	32.8 mln words	31.1 mln words
Uzbek	5 web-sites: kun.uz, www.uzdaily.uz, www.gazeta.uz, uza.uz, xabar.uz	404 000 pairs	7.7 mln words	7.6 mln words
Crimean-Tatar	1 web-site crimeantatars.club; 2 corpora: OPUS-GNOME и OPUS-Ubuntu; 8 translated printed books	26 000 pairs	0.17 mln words	0.16 mln words
Chuvash	Private corpus	241 000 pairs	2.9 mln words	2.8 mln words
Kazakh	WMT corpus	4.5 mln pairs	80.2 mln words	83.9 mln words

A promising direction, which in the future will increase the amount of data available for training an MT system, is to unify parallel corpora for various Turkic languages. It can be done based on the structural-functional model of the Turkic morpheme. This approach should significantly increase the amount of training data due to the use of the common features of the Turkic languages. The Turkic morpheme model includes several modules working with bilingual and multilingual dictionaries, a system for describing the morphotactics rules for different languages, as well as modules for morphological analysis and synthesis. We propose a software module for translation between Turkic languages, which will work according to the following principle: at the initial step, syntactic structures and individual words in the original sentence are analyzed and parsed, this is followed by the stage of translating individual elements into the target language, followed by the synthesis of words of the target language based on individual morphemes. The key aspect on which attention was focused was the consideration of the ambiguity that arises in the translation process. So, for a word from the original sentence, several variants of morphological analysis may be available, which in the absence of a disambiguation system for a given language will lead to ambiguous translation into the target language. The presence of several ambiguous words in a sentence will significantly increase the number of final translation alternatives. At the moment, a script for translation between Turkic languages is being developed, which will be able to generate translation options for the original sentence as follows:

- all possible translation alternatives are generated;
- for each case of ambiguity, the first option is selected;
- for each case of ambiguity, a random option is selected;
- all possible translation variants are generated, which are then ranked by a statistical language model trained on the monolingual corpus.

3. Training neural models for machine translation

Training MT models was done based on a neural network approach. We used the Transformer neural network architecture, the key feature of which is the use of the multi-head attention mechanism.

Three series of experiments were carried out:

- basic models for each of the language pairs;
- an experiment on the use of parallel data for all related languages for training a "general" model, followed by additional training to the target language NMT model;
- training of a unified Russian-Turkic multilingual model with the addition of a language tag to the original sentence.

For all experiments, the collected parallel corpora were randomly divided into training, test, and validation sets; the minimum size of test and validation samples is 1000 sentence pairs (for Crimean Tatar, Kirgiz, Uzbek, Bashkir languages), 2000 pairs – for Kazakh, 2500 pairs – for Tatar, and 2900 pairs – for Chuvash languages.

As part of the first experiment we trained ensembles of neural network models for all translation directions (14 ensembles for 7 language pairs). Each ensemble consisted of 8 independent neural network models, 4 of which were “left-to-right” and 4 – “right-to-left” models. To control the training

process, 3 criteria were used: ce-mean-words, perplexity, BLEU. As the exit criteria, we set the maximum value of training epochs equal to 300 (3000 – for the Crimean Tatar), and the maximum number of iterations, during which the improvement of the target criterion ce-mean-words, equal to 5, was not achieved.

The second experiment was to test the hypothesis that low-resource language pairs can benefit from the use of a pre-trained multilingual neural network. This "common" neural network is trained on the corpora for all 7 languages. At the initial stage, a single Russian-Turkic corpus was formed (a random split of sentences for validation and test samples was not repeated, the division was obtained by combining already formed splits for each of the language). Based on this corpus, one neural network model was trained from left to right and one from right to left. Then this model was used as a baseline for additional fine-tuning for each specific language pair. Fine-tuning was carried out with the same settings as in the first experiment, except that the maximum values for the number of training epochs were increased and the current best values of the criteria were reset.

For the third experiment, the target Turkic language tag was added to all the original Russian sentences in the format "<language_code> Original sentence". Another difference of this experiment was the use of the SentencePiece [9] algorithm to split words into parts (in the first two experiments, the division was carried out using the BPE algorithm [10]).

4. Experiments

To assess the quality of the built machine translation systems, we used the BLEU metric.

Based on the results of the first experiment, we can note the correlation between the quality of translation and the size of used training corpus. At the same time, the rather high BLEU values for some languages were influenced by a small number of different sources used to create parallel corpora. This led to the fact that the test subcorpus had very similar distribution with the training subcorpus.

Table 1 presents the translation quality for the base models from the first experiment. BLEU values were calculated for an ensemble of 8 neural networks, as well as for 4 independent "left-to-right" models.

Table 2

First experiment's result: BLEU scores for base models

Translation direction	Training data	8-NN ensemble, BLEU	4 separate NN, BLEU
Russian-Kazakh	4 513 000	48,2	45,6; 47,8; 43,9; 46,3
Kazakh-Russian		64,3	62,6; 62,3; 62,2; 62,2
Russian-Tatar	1 994 000	34,6	32,6; 32,6; 32,4; 32,6
Tatar-Russian		37,5	35,3; 34,8; 35,4; 34,8
Russian-Kirgiz	416 000	19,7	16,7; 17,6; 17,7; 18,4
Kirgiz-Russian		21,6	18,6; 18,0; 17,8; 18,5
Russian-Uzbek	404 000	32,8	30,4; 30,2; 30,1; 30,2
Uzbek-Russian		34,2	31,7; 31,1; 31,2; 31,9
Russian-Bashkir	351 000	45,7	42,7; 43,8; 43,1; 42,6
Bashkir-Russian		45,4	40,8; 40,2; 40,3; 40,0
Russian-Chuvash	236 000	21,9	18,1; 18,3; 18,1; 18,2
Chuvash-Russian		24,8	20,6; 20,9; 20,6; 20,6
Russian-Crimean Tatar	26 000	13,5	12,3; 12,6; 12,5; 12,4
Crimean Tatar-Russian		15,7	13,9; 13,5; 13,9; 14,3

Table 3 shows the translation quality for the models from the second experiment, obtained by fine-tuning of the general Turkic model; and for the the multilingual model from the third experiment, obtained during training with the additional language tags (marked in the table as multilingual).

Table 3
Second and third experiments' result: BLEU scores for multilingual and fine-tuned models

Translation direction	BLEU	Difference from the base models, % BLEU
Russian-Kazakh	47,8	+0%
Kazakh-Russian	61,9	-1,1%
Russian-Kazakh (multilingual)	48,4	+1,3%
Russian-Tatar	33,6	+3,1%
Tatar-Russian	36,4	+3,1%
Russian-Tatar (multilingual)	33,2	+1,8%
Russian-Kirgiz	22,2	+20,7%
Kirgiz-Russian	25,0	+34,4%
Russian-Kirgiz (multilingual)	22,5	+22,3
Russian-Uzbek	33,4	+9,9%
Uzbek-Russian	35,5	+11,3%
Russian-Uzbek (multilingual)	31,1	+2,3%
Russian-Bashkir	45,9	+4,8%
Bashkir-Russian	47,3	+15,9%
Russian-Bashkir (multilingual)	47,3	+8,0%
Russian- Chuvash	28,0	+53%
Chuvash-Russian	30,4	+45,4%
Russian- Chuvash (multilingual)	25,8	+41%
Russian-Crimean-Tatar	22,7	+80,2%
Crimean-Tatar-Russian	24,4	+70,6%
Russian-Crimean-Tatar (multilingual)	15,0	+19%

The results of the second experiment demonstrate that for language pairs with the smallest amount of training data, the use of a pre-trained "common Turkic" model can significantly improve the quality of translation. For instance, for Russian-Chuvash translation the BLEU increased from 18.3 to 28.0, and for Russian-Crimean Tatar – from 12.6 to 22.7. The quality of the translation of the unified multilingual model from Russian into 7 Turkic languages should be separately noted. For all 7 languages, the quality of work of one single model surpassed the quality of work of the base models, and for 3 languages – also of the fine-tuned versions. At the same time, this unified multilingual model has advantage at inference stage in terms of memory efficiency.

5. Conclusions

This paper presents the results of the creation of parallel corpora for 7 Turkic languages and the development of software for training neural machine translation systems. A list of sources of parallel data was formed, auxiliary software tools were developed, which made it possible to form a parallel corpus with a total volume of more than 7 million sentence pairs. For some languages, for example, Crimean Tatar, a parallel corpus was built for the first time; for other languages, the volumes of the existing corpora were significantly increased. For the first time, the necessary software was developed and complex experiments were carried out to build modern neural machine translation models for a group of 7 Turkic languages.

6. Acknowledgements

The reported study was funded by RFBR according to the research project № 20-07-00823.

7. References

[1] Hassan H. et al, Achieving Human Parity on Automatic Chinese to English News Translation, 2018. URL: http://arxiv.org/abs/1803.05567.

[2] Koehn P., Knowles R., Six Challenges for Neural Machine Translation, 2017. URL: http:arxiv.org/abs/1706.03872.

[3] Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B., About Turkic Morpheme Portal, in: Proc. of 2nd Workshop CMLS'20, Kazan, November 12-13, 2020, pp. 226-243.

[4] trafilatura: Web scraping tool for text discovery and retrieval. URL: https://trafilatura.readthedocs.io/.

[5] razdel: rule-based system for Russian sentence and word tokenization. URL: https://github.com/natasha/razdel.

[6] Russian-Tatar machine translation system Tatsoft. URL: https://translate.tatar.

[7] Bitextor: tool to automatically harvest bitexts from multilingual websites. URL: https://github.com/bitextor/bitextor.

[8] Bleualign: an MT-based sentence alignment tool. URL: https://github.com/rsennrich/Bleualign.

[9] Kudo T., Richardson J SentencePiece, A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, 2018. URL: http://arxiv.org/abs/1808.06226.

[10] Sennrich, R., Haddow, B., Birch, A Neural Machine Translation of Rare Words with Subword Units, volume 1 of Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1715-1725.