First Results of the "TurkLang-7" Project: Creating Russian-Turkic Parallel Corpora and MT Systems

Aidar Khusainov[1[0000−0002−7763−1420]], Dzhavdet Suleymanov[2[0000-0003-1404-0372]], Rinat Gilmullin[3[0000-0002-8520-8921]], Alina Minsafina[4[0000-0001-6413-8429]], Lenara Kubedinova[5[0000-0002-1293-8173]], Nilufar Abdurakhmonova[6[0000-0001-9195-5723]]

[1] Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
khusainov.aidar@gmail.com
[2] Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
dvdt.slt@gmail.com
[3] Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
rinatgilmullin@gmail.com
[4] Istanbul University, Istanbul, Turkey
minsafina@yandex.com
[5] Crimean Federal University, Simferopol, Russia
kubedinova@gmail.com
[6] Tashkent State university of the Uzbek language and literature, Tashkent, Uzbekistan
abdurahmonova.1987@mail.ru

Abstract. The idea of the "TurkLang-7" project is to create datasets and neural machine translation systems for a set of Russian-Turkic low-resource language pairs. It is planned to achieve this goal through a hybrid approach to the creation of a multilingual parallel corpus between Russian and Turkic languages, studying the applicability and effectiveness of neural network learning methods (transfer learning, multi-task learning, back-translation, dual learning) in the context of the selected language pairs, as well as the development of specialized methods for the unification of parallel data in different languages, based on the agglutinative nature of the selected Turkic languages (structural and functional model of the Turkic morpheme). In this paper, we describe the main stages of work on this project and the results of the first year: we developed a semi-automatic process for creating parallel corpora, collected data from several sources on 7 Turkic languages, and conducted the first experiments to create machine translation systems.

Keywords: Neural Machine Translation, Multilingual Datasets, Data collection, Turkic Languages.

1 Introduction

The field of creating automatic machine translation systems has developed rapidly in recent years, largely due to the successful use of modern machine learning methods. However, neural network machine translation methods that allow achieving the best

results for the largest pairs of world languages (English-German, English-Chinese, and others) cannot be directly used in the case of a lack of training data.

Of particular importance are a number of subtasks related to the adaptation and refinement of existing approaches for the cases of low-resource languages. Certain success has been achieved in this area, including the transfer learning technologies and data augmentation techniques (for example, back-translation, dual learning).

This project is aimed at developing methods and software for 7 language pairs, in which one language is Russian and the other belongs to the Turkic language group. To achieve this goal and overcome the problem of lack of training data, we propose to collect parallel training data, to develop a method for the unification of the collected parallel corpora based on the structural-functional model of the Turkic morphemes [1], as well as to create software tools for training a multilingual machine translator based on transfer learning approaches and data augmentation. This should allow for the first time to create a parallel corpus for the Crimean Tatar-Russian language pair; the final machine translation system will also work with 6 more language pairs (Tatar-Russian, Bashkir-Russian, Chuvash-Russian, Kazakh-Russian, Kyrgyz-Russian and Uzbek-Russian). The total number of native speakers for specified Turkic languages is 57.93 million people [2], living predominantly on the territory of the Russian Federation and the CIS countries.

As a result of this work we will provide information on the data collection procedure, the number of parallel sentences we have collected for 7 language pairs, and results of the experiment on training Transformer-based NMT systems.

Section 2 of this article provides an overview of research in this area, section 3 contains a description of the main project's stages, section 4 – experiments and current results.

2 Related Work

The approaches applied to the development of machine translation systems have undergone major changes in recent years. Considerable efforts are directed equally towards solving machine translation problems for the cases of the world's largest languages and low-resourced languages. The amount and quality of data available for the selected languages determine the set of algorithms and approaches to create an MT system.

The problem of machine translation is solving using the so-called sequence-to-sequence models [3], built, for example, based on recurrent, convolutional neural networks, including elements of an encoder and decoder (encoder/decoder architecture). Models showing the best performance also include the attention mechanism (attention, self-attention).

There are various neural network architectures designed to speed up the learning process and improve the quality of the MT system's work: recurrent neural networks [4], convolutional neural networks [5], Transformer models [6], and Evolved Transformer. The attention mechanism was also improved: variants of multi-hop attention, self-attention, and multi-head attention were proposed [5, 7].

The choice of technology for building a machine translation system depends very much on the availability and amount of the initial training data. The presence of large mono-corpora for the source and target languages allows the use of unsupervised approaches to building MT systems. The main idea of this approach is to build a single vector space of words/phrases for both languages. At the moment, there are options for the implementation of this approach based on the statistical [8], neural network [9], and hybrid approaches [10].

Various options were also proposed for using mono corpora to improve the quality of translation in training with partial involvement of a teacher (semi-supervised approach) [11]. Another way to use monolingual data is to supplement the decoder part of the system with a language model [12]. This approach was used in the earliest works of IBM [13]. Later it was shown that an additional language model for the target language allows systems based on a statistical approach to improve the naturalness and correctness of translation [14]. A similar strategy was later also applied to neural network machine translation systems [15]. In addition to being used during decoding, neural network language and translation models can be successfully integrated internally by combining the hidden states of the models [16]. The neural network architecture allows the use of multi-task learning and parameter sharing [17].

And, finally, in [18] it was proposed to add an auxiliary autoencoding task for monolingual data, which ensures that the original sentence is obtained as a result of the consecutive translation of a sentence in both directions.

In [19], the authors showed that the quality of translation in the case of low-resource language pairs can be improved due to augmented data, where sentences in the source language are created by a simple copy of sentences in the target language.

The approach in [20] suggests a very efficient way to automatically increase data for training. The method is called back-translation (BT): first, an auxiliary translation system from the target language to the source language is trained on the available parallel data, and then this system is used to translate the monolingual corpus of the target language, thereby increasing the volume of the parallel corpus. The resulting parallel corpus is used as training data for a machine translation system.

BT is easy to use as it does not require any changes to the machine translator training algorithms. In addition to the main task of increasing the volume of training data for low-resource pairs of languages, it can also be used to use a monolingual corpus for the task of adapting an MT to a specific domain [21]. As the latest ideas for improving BT, it was proposed to abandon the generation of synthetic pairs using beam search [22] or greedy search [23]. Both of these algorithms allow searching for the posterior maximum (MAP), that is, to find the hypothesis with the maximum probability according to the model. However, the use of MAP can lead to a less diverse subcorpus of translations, since in cases of ambiguity, the algorithm will always choose the most likely option [24]. Alternatively, it is recommended to use the random sampling method [25]. This allows preserving the lexical variety of the generated sentence pairs. At the same time, additional rules can be introduced to exclude very rare translation options [26]. The important change to the approach was proposed in [27]: the authors presented an iterative process of learning / adding a synthetic part of the training corpus to improve the quality of the final systems.

In [28], it was shown how the quality of NMT can be improved if there are monolingual corpora for both languages. The use of two corpora simultaneously allows the transition from BT to the so-called dual learning: learning occurs simultaneously in both directions of translation, BT is used iteratively in both directions to gradually increase the size of the training corpus and the proportion of synthetic sentences.

The Byte-pair encoding (BPE) approach [29] deserves a separate mention; it is applied to the problem of machine translation based on basic elements less than a whole word (subword MT). The use of segmentation based on BPE [30] allows, among other things, to solve the problem of translation with an open dictionary (the system can translate any words, including those that are not present in the training corpus). BPE was created as a compression algorithm, but has been adapted for word segmentation as follows: each word from the training dictionary is represented by a sequence of characters terminated with a special end-of-word character; all symbols are added to the element dictionary; the most frequent pairs of symbols are determined - the found sequences are added to the dictionary of elements and combined into a corpus. The procedure is repeated until the specified number of merge operations is reached.

3 Project Description

The project aims to create machine translation systems for such language pairs, for most of which there are not enough (or not at all) parallel data to train modern neural models. Therefore, the essential stage is the data collection procedure. To solve this task, we proposed several approaches. First of all, we tried to combine all the data that already exist. There are several main sources for parallel information: news and government organization web-sites, translated books, already existing corpora. So the first stage of the project was to gather information about existing data sources. Must 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 is books, for Kazakh and Chuvash – existing corpora, for all other languages of the project (Kyrgyz, Bashkir, Tatar, and Uzbek) – bilingual web-sites.

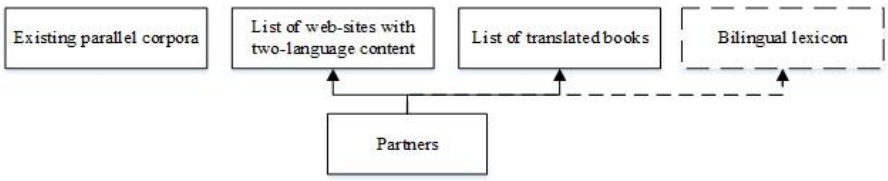

Fig. 1. First stage: collecting data sources

As for website data collection, we established a semi-automatic process of data processing, Fig. 2. We first manually analyzed the site structure, the existence of sitemap files, looked for ways to automatically connect pairs of translated pages. The next step was to create a list of URLs we need to download, having in mind not to harm the work of the sites and trying to do all auxiliary downloads (for example, downloading

the main news page to collect each news page URL from it) as slow as practically possible. The download procedure was conducted by Trafilatura tool [31] that showed great performance extracting only main text data from the page for all analyzed web-sites except only one of the Ministry of Justice of the Kyrgyz Republic official web-site. The downloaded text materials were then further processed by the razdel tool [32] resulting in files split by sentences. The last step in this stage is text filtering that removes all characters like '°', '■', etc.

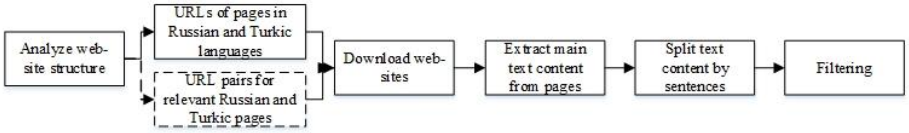

Fig. 2. Second stage: web-site data processing

The next stage is a document and segment alignment process, Fig. 3.

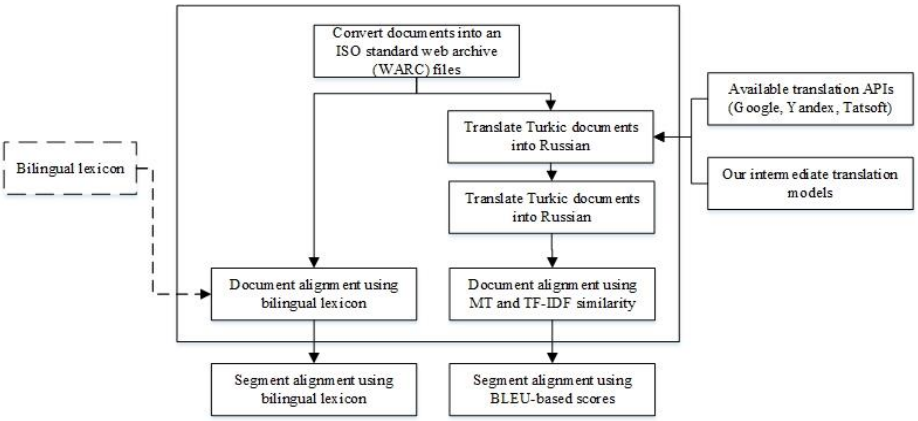

Fig. 3. Stage three: document and segment alignment

We first convert all downloaded and processed text documents into a WARC-format file which is a standard format for web archives. Documents in different languages are saved in separate site folders. Depending on the availability of machine translation systems we can use one of two approaches for document and segment alignment. If there is an MT system, we translate all the documents from the website in one language to another where the source language is the language with less amount of data. We used Yandex and Google MT systems for all languages except Crimean-Tatar, Tatar, and Bashkir. For Tatar, we used Tatsoft NMT [33], and for Bashkir also Tatsoft Tatar-Russian NMT system with several preprocessing of Bashkir texts (converting some specific Bashkir symbols into closest Tatar characters). This "Bashkir-Tatar-Russian" translation procedure was good enough for the task of alignment.

For Crimean-Tatar there are no available MT systems yet, so we used a bilingual lexicon to find pairs of documents and segments. Some parts of the Bitextor [34] system and the bleualign [35] tool were used in this stage.

And the last step of the corpora creation process includes executing deduplication and rule-based data augmentation algorithms, Fig. 4. The main idea is to use a rule-based inter-Turkic machine translation system to convert all collected Turkic-Russian corpora into one target Turkic-Russian corpus. For instance, when building the Crimean-Tatar-Russian MT system we can lie on not only a few thousand sentences for this language pair but also translate all Turkic sentences in Tatar-Russian, Bashkir-Russian, etc. corpora to create an augmented Crimean-Tatar corpus. The key here is the usage of the Turkic morpheme model [1], which first analyzes source Turkic sentence (splitting it into stem morphemes and affixal chains) and then synthesize target Turkic language sentence using linguistic databases. We plan to use this system in the second year experiments.

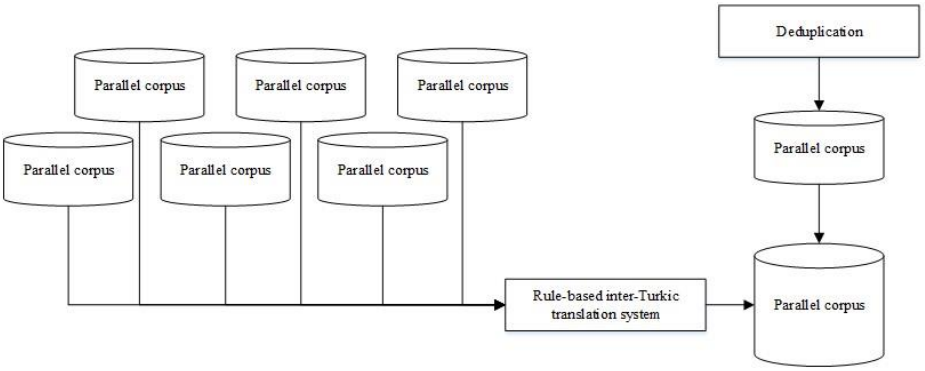

Fig. 4. Fourth stage: using rule-based data augmentation

The key task of building a machine translation model is solved based on a neural network approach. At the initial stage, we use the Transformer neural network architecture, a key feature of which is the use of the multi-head attention mechanism and the absence of convolutional and recurrent layers.

We plan to use a complex approach to build multilingual MT systems:

- using various approaches to transfer knowledge from one language pair to another (for example, a fine-tuning neural network pre-trained on a more resource-rich language pair/pairs; introducing a token to represent the source language at the level of the embedding layer of a neural network and learning a single multilingual neural network);
- developing a common representation of word parts for all declared languages (for example, common byte-pair encoding elements for all languages);
- using methods for training data augmentation, for example, back-translation algorithm (using intermediate versions of the translator to increase the volume of parallel data based on the translation of monolingual text corpora) and its modifications

that use the random sampling method instead of beam search to obtain translations with richer vocabulary;
- development of methods for unification of collected parallel corpora for different language pairs, which will allow to use corpora available for other language pairs. It is proposed to use the structural and functional model of the Turkic morpheme [1] to include information on the relationship of grammatical roles performed by affixes in various Turkic languages into the training data. This information will allow at the initial stage of the corpus preparation to form uniform elements for morphemes in different languages.

4 Experiments and Results

4.1 Data Collection

The developed algorithms allowed us to run the process of data collection for all of the 7 language pairs. The current results show that the amount of data collected is substantial and can allow to build basic MT systems for most of the language pairs, Table 1–5.

Table 1. Kyrgyz corpus statistics

Source	Initial Ru docs	Initial Ky docs	# of parallel sentences
https://sti.gov.kg/	467	467	4 178
http://www.kenesh.kg/	7257	7265	36 389
http://minjust.gov.kg/	1789	1789	17 092*
http://novosti.kg/	50643	39952	36 517
https://edu.gov.kg/	1207	1207	21 140
http://mineconom.gov.kg/	509	509	5 856
http://med.kg/	480	408	1 370
https://ru.sputnik.kg/news/	42240	41834	43 874
JW300	-	-	276 866
	Total		426 190

Table 2. Bashkir corpus statistics

Source	Initial Ru docs	Initial Ba docs	# of parallel sentences
bash.news	58 267	24 753	1 006
https://ufacity.info/	681	681	5 219
https://glavarb.ru/	1886	2 369	3 366
http://www.bashinform.ru/*			254 972
http://bashdram.ru/	1202	1 181	2 122
https://house.bashkortostan.ru/	1264	432	2 164

https://pravitelstvorb.ru/	12 384	6 570	11 714
JW300	-	-	47 658
Encyclopedia	-	-	24 692
	Total		352 913

Table 3. Tatar corpus statistics

Source	Initial Ru docs	Initial Tt docs	# of parallel sentences
tatar-inform.tatar	675 400	201 927	708 665
https://tatarstan.ru/	15 211	151 995	232 075
https://kiziltan.rbsmi.ru/	5 582	36 372	12 166
JW300	-	-	207 100
Existing Ru-Tt corpus	-	-	982 537
	Total		2 142 543

Table 4. Uzbek corpus statistics

Source	Initial Ru docs	Initial Uz docs	# of parallel sentences
https://kun.uz/	28 797	192 444	In progress
www.uzdaily.uz	45 515	6 661	52 117
https://www.gazeta.uz/	34 840	17 834	In progress
http://uza.uz/	8 709	33 465	In progress
http://xabar.uz/	1 629	1 629	14 532
	Total		66 649+

Table 5. Crimean-Tatar corpus statistics

Source	Initial Ru docs	Initial Crh docs	# of parallel sentences
https://www.crimeantatars.club/	5 753	1 832	In progress
OPUS-GNOME (Latin)	-	-	105 000
OPUS-Ubuntu (Latin)	-	-	19 000
Literature (9 books)	-	-	Scanning
	Total		124 000+

For Chuvash-Russian we used private parallel corpus of 205 000 sentence pairs and for Kazakh-Russian – 5 million sentence pairs from WMT competition.

4.2 Machine Translation Systems

For the first experiment setup we chose the Transformer-Base architecture without usage of monolingual data. Four independent NN were trained with different seed

values, together with four right-to-left models used for rescoring giving us 8 NN ensemble.

The example of changing perplexity and BLEU during training process for one of the Russian-Bashkir NN presented in Fig. 5.

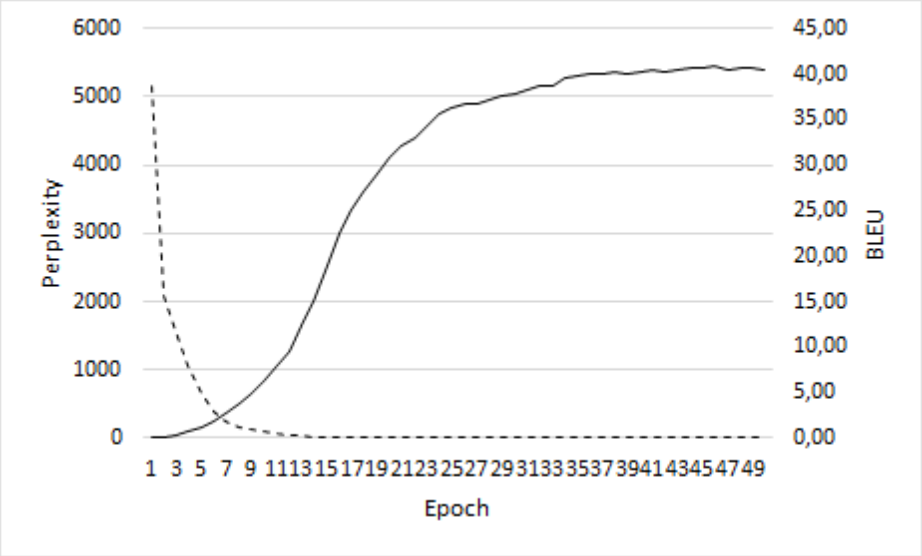

Fig. 5. Perplexity and BLEU values during Russian-Bashkir NN training

Experiments were conducted on DGX-1 workstation with eight 32GB V-100 GPUs. Marian toolkit [36] were used. The obtained results for all 8 NN ensemble, 4 left-to-right NN ensemble, and for each of the left-to-right models presented in Table 6.

Table 6. The quality of built MT systems for Test subcorpora

Translation direction	Full 8 NN ensemble	4 NN ensemble	Model-1	Model-2	Model-3	Model-4
Russian-Bashkir	45.7	45.3	42.7	43.8	43.1	42.6
Bashkir-Russian	45.4	43.1	40.8	40.2	40.3	40.0
Russian-Kyrgyz	19.7	19.2	16.7	17.6	17.7	18.4
Kyrgyz-Russian	21.6	20.4	18.6	18.0	17.8	18.5
Russian-Chuvash	21.9	21.3	18.1	18.3	18.1	18.2
Chuvash-Russian	24.8	24.2	20.6	20.9	20.6	20.6
Russian-Kazakh	48.2	49.0	45.6	47.8	43.9	46.3
Kazakh-Russian	51.4	51.4	51.1	58.3	54.0	45.2

The results of the first experiment proved the dependency between the size of training corpus and the quality of translation, but also we obtained very high results for systems, where we used small number of different data sources. This fact lead us to the impossibility of separating train and test subcorpus in terms of sources, so we just randomly divided sentence pairs. Therefore, there are no identical sentences in train-

ing and testing parts, but they can be very similar and have almost identical lexical and grammatical features. Based on that we planned a new task for the second year of the project: to create a set of rules which can be then used to manually form testing corpora for Turkic-Russian languages. These corpora give us the possibility to objectively compare different MT systems.

5 Conclusion

In this paper we presented the main ideas and first results of TurkLang-7 project: developed software tools, collected parallel data and NMT systems for Turkic-Russian languages. We plan to continue data collecting process and to conduct fine-tuning and data augmentation experiments using rule-based inter-Turkic translation system.

6 Acknowledgments

The reported study was funded by RFBR, project number 20-07-00823.

References

1. Gatiatullin, A.: Mnogofunkcionalnij Internet servis kak instrument dlya formirovaniya I ispolzovaniya leksikograficheskoj bazy tyurkskih yazykov. In: Sohranenie yazykov narodov mira i razvitie yazykovogo raznoobraziya v kiberprostranstve: kontekst, politika, praktika 2019, 117–125 (2019).
2. Ethnologue: Languages of the World, http://www.ethnologue.com, last accessed 2020/05/17.
3. Sutskever, I, Oriol, V., Quoc, V.: Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 3104–3112 (2014).
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 (2014).
5. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, N.: Convolutional sequence to sequence learning. In: International Conference of Machine Learning, 1243–1252 (2017).
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, 30, 5998–6008(2017).
7. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv:1705.04304 (2017).
8. Artetxe, M., Labaka, G., Agirre, E.: Unsupervised Statistical Machine Translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3632–3642 (2018).
9. Artetxe, M., Labaka, G., Agirre, E., Cho, K.: Unsupervised Neural Machine Translation. arXiv:1710.11041 (2017).
10. lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-Based & Neural Unsupervised Machine Translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 5039–5049 (2018).

11. Munteanu, D., Fraser, A., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In: ACL, 265–272 (2004).
12. Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., Bengio, Y.: On using monolingual corpora in neural machine translation. arXiv:1503.03535 (2015).
13. Brown, P., Cocke, J., Pietra, S., Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Roossin, P.: A statistical approach to machine translation. Computational Linguistics, 16, 79–85 (1990).
14. Koehn, Ph., Och, F., Marcu, D.: Statistical phrase-based translation. In: Conference of the North American Chapter of the Association for Computational Linguistics, 127–133 (2003).
15. He, W., He, Z., Wu, H., Wang, H.: Improved neural machine translation with smt features. In: Conference of the Association for the Advancement of Artificial Intelligence, 151–157 (2016).
16. Gulcehre, C., Firat, O., Xu, K., Cho, K., Bengio, Y.: On integrating a language model into neural machine translation. Computer Speech & Language, 45, 137–148 (2017).
17. Domhan, T., Hieber, F.: Using target-side monolingual data for neural machine translation through multi-task learning. In: Conference on Empirical Methods in Natural Language Processing, 1500–1505 (2017).
18. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Semi-supervised learning for neural machine translation. arXiv:1606.04596 (2016).
19. Currey, A., Barone, A., Heafield, K.: Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In: Proc. of WMT, 148–156 (2017).
20. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. arXiv:1511.06709 (2015).
21. Bertoldi, N., Federico, M.: Domain adaptation for statistical machine translation with monolingual resources. In: Workshop on Statistical Machine Translation, 182–189 (2009).
22. Sennrich, R., Haddow, B., Birchm A.: Improving neural machine translation models with monolingual data. In: Conference of the Association for Computational Linguistics, 1, 86–96 (2016).
23. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. arXiv:1711.00043 (2017).
24. Ott, M., Auli, M., Grangier, D., Ranzato, M.: Analyzing uncertainty in neural machine translation. In: Proceedings of the 35th International Conference on Machine Learning, 80, 3956–3965 (2018).
25. Imamura, K., Fujita, A., Sumita, E.: Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, 63, 55 (2018).
26. Graves, A.: Generating sequences with recurrent neural networks. arXiv:1308.0850 (2013).
27. Hoang, V., Koehn, P., Haffari, G., Cohn, T.: Iterative back-translation for neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, 18–24 (2018).
28. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Semi-supervised learning for neural machine translation. In: Conference of the Association for Computational Linguistics, 1965–1974 (2016).
29. Gade, F.: A New Algorithm for Data Compression. C Users J., 12(2), 23–38 (1994).

30. 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, 1715–1725 (2016).
31. Barbaresi, A.: Generic Web Content Extraction with Open-Source Software. In: Proceedings of KONVENS, Kaleidoscope Abstracts, 267–268 (2019).
32. Rule-based system for Russian sentence and word tokenization, https://github.com/natasha/razdel, last accessed 2020/04/07.
33. Khusainov, A., Suleymanov, D., Gilmullin, R.: The Influence of Different Methods on the Quality of the Russian-Tatar Neural Machine Translation. In: Kuznetsov S.O., Panov A.I., Yakovlev K.S. (eds) RCAI 2020, LNCS, 12412, 251–261, Springer, Cham (2020).
34. Espla-Gomis, M., Forcada, M.: Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. The Prague Bulletin of Mathematical Linguistics, 93, 77–86 (2010).
35. Sennrich R., Volk, M.: Iterative, MT-based sentence alignment of parallel texts. In: Nordic Conference of Computational Linguistics, 175–182 (2011).
36. Fast Neural Machine Translation in C++, https://marian-nmt.github.io/, last accessed 2020/09/17.