


IRSTI 20.23.25

DOI: <https://doi.org/10.26577/JMMCS.2022.v113.i1.12>

D.R. Rakhimova , **A.Z. Zhunussova*** 
Al-Farabi Kazakh National University, Kazakhstan, Almaty
*e-mail: alia_94-22@mail.ru

POST-EDITING FOR THE KAZAKH LANGUAGE USING OPENNMT

The modern world and our immediate future depend on applied intelligent systems, as new technologies develop every day. One of the tasks of intelligent systems is machine (automated) translation from one natural language to another. Machine translation (MT) allows people to communicate regardless of language differences, as it removes the language barrier and opens up new languages for communication. Machine translation is a new technology, a special step in human development. This type of translation can help when you need to quickly understand what your interlocutor wrote or said in a letter.

The work of online translators used to translate into Kazakh and vice versa. Translation errors are identified, general advantages and disadvantages of online machine translation systems in Kazakh are given. A model for the development of a post-editing machine translation system for the Kazakh language is presented.

OpenNMT (Open Neural Machine Translation) is an open source system for neural machine translation and neural sequence training. To learn languages in OpenNMT, you need parallel corpora for language pairs. The advantage of OpenNMT is that it can be applied to all languages and can handle large corpora. Experimental data were obtained for the English-Kazakh language pair. Experimental data were obtained for the English-Kazakh language pair.

Key words: Opennmt, neural machine translation, turkic languages.

Д.Р. Рахимова, А.Ж. Жунусова*
Әл-Фараби атындағы Қазақ ұлттық Университеті, Қазақстан, Алматы қ.
*e-mail: alia_94-22@mail.ru

OpenNMT көмегімен қазақ тіліне постредактрлеу

Қазіргі әлем және біздің жақын болашағымыз қолданбалы интеллектуалды жүйелерге байланысты, өйткені жаңа технологиялар күн сайын дамып келеді. Интеллектуалды жүйелердің міндеттерінің бірі - бір табиғи тілден екіншісіне машиналық (автоматтандырылған) аударманы қолданып аудару. Машиналық аударма (тілдік аударма) адамдарға тілдік айырмашылықтарға қарамастан байланыс жасауға мүмкіндік береді, өйткені ол тілдік тосқауылды жойып, қарым-қатынас үшін жаңа тілдерді ашады. Машиналық аударма - бұл жаңа технология, адам дамуындағы ерекше қадам. Аударманың бұл түрі сізге әңгімелесушінің хатта не жазғанын немесе не айтқанын тез түсіну қажет болғанда көмектесе алады.

Онлайн аудармашылардың жұмысы бұрын қазақшаға және керісінше аударылатын. Аударма қателері анықталды, онлайн-машиналық аударма жүйесінің қазақ тіліндегі жалпы артықшылықтары мен кемшіліктері келтірілді. Қазақ тіліне арналған өңдеуден кейінгі машиналық аударма жүйесін әзірлеу моделі ұсынылған.

OpenNMT (Open Neural Machine Translation) – нейрон машинасын аудару және нейрон рет-тілігін оқытуға арналған ашық бастапқы жүйе. OpenNMT-де тілдерді үйрену үшін сізге тілдік жұптарға параллель корпустар қажет. OpenNMT-тің артықшылығы - ол барлық тілдерге қолданыла алады және үлкен корпустарды басқара алады. Эксперименттік мәліметтер ағылшын-қазақ тілі жұбы үшін алынды.

Түйін сөздер: Opennmt, нейрон машиналық аударма, түркі тілдері.

Д.Р. Рахимова, А.Ж. Жунусова*
Казахский национальный университет имени аль-Фараби, Казахстан, г.Алматы
*e-mail: alia_94-22@mail.ru

Постредактирование для казахского языка с использованием openNMT

Современный мир и наше ближайшее будущее зависят от прикладных интеллектуальных систем, так как новые технологии развиваются с каждым днем. Одной из задач интеллектуальных систем является машинный (автоматизированный) перевод с одного естественного языка на другой. Машинный перевод (МП) позволяет людям общаться независимо от различия языков, поскольку это устраняет языковой барьер и открывает новые языки общения. Машинный перевод - это новая технология, особый шаг в развитии человека. Этот тип перевода может помочь, когда нужно быстро понять, что ваш собеседник написал или сказал в письме.

Работа онлайн-переводчиков, используемых для перевода на казахский язык и обратно. Выявлены ошибки перевода, даны общие преимущества и недостатки онлайн систем машинного перевода на казахском языке. Представлена модель разработки системы пост-редактирования машинного перевода для казахского языка.

OpenNMT (Open Neural Machine Translation) – это система с открытым исходным кодом для нейронного машинного перевода и обучения нейронной последовательности. Для обучения языки в OpenNMT нужны параллельные корпуса для языковых пар. Преимуществом OpenNMT является применение ко всем языкам и может работать с большими корпусами. В статье рассматривается обучения тюркские языки в OpenNMT. Было получено экспериментальные данные для англо-казахского языковой пары.

Ключевые слова: Opennmt, нейронный машинный перевод, тюркские языки.

1 Introduction

Türkic languages, a language family, spread over the territory from Turkey in the west to Xinjiang in the east and from the coast of the East Siberian Sea in the north to Khorasan in the south. The speakers of these languages live compactly in the CIS countries (Azerbaijanis – in Azerbaijan, Turkmens – in Turkmenistan, Kazakhs – in Kazakhstan, Kyrgyz – in Kyrgyzstan, Uzbeks – in Uzbekistan; Kumyks, Karachais, Balkars, Chuvashs, Tatars, Bashkirs, Nogais, Yakuts, Tuvans, Khakass, Mountain Altai – in Russia; Gagauz – in the Transnistrian Republic) and beyond its borders – in Turkey (Turks) and China (Uighurs). Currently, the total number of speakers of the Türkic languages is about 120 million.

The Türkic languages are similar in structure and meaning. This can be seen in the following table:

Table 1. Words of Türkic languages

Ancient Türkic	Turkish	Turkmen	Turkmen	In Kazakh	Kyrgyz	In uzbek	In Uyghur	Tyvan
Ana	ana/anne	ene	ana	ана /ana/	Ene	Ona	Ana	Ава
Burun	Burun	burun	borin	мұрын /murin/	murun	Burun	burun	Думчук
Qol	Kol	qol	qul	қол /qol/	Qol	qo'l	kol	Хол
Yol	Yol	ýol	yul	жол /jol/	Jol	yo'l	yol	орук (чол)
Semiz	Semiz	semiz	simez	семіз /semiz/	semiz	Semiz	semiz	Семис

But languages are different from other languages, they have special characteristics:

- proximity of the lexical structure;
- the law of harmony;
- agglutination – a series of affixes;
- lack of a category;
- lack of auxiliary words (prepositions);
- special word order.

Therefore, when translating, the Turkic languages give out morphological, lexical, and semantic errors.

To evaluate errors, we will use well-known translation machines, such as: Google, Yandex:
Table 2. Online transfers completed in February 2021 y.

Source text for translation into Türkic language	Name of machine translation systems and translation results		disadvantages
	Yandex	Google	
Таныш булығыз, бу минем гаиләм (Tatar language)	Meet my family	Meet me, this is my family	Google Translate pays attention to punctuation. If there is an exclamation mark at the end of a sentence, this is correctly translated as "meet", otherwise it is incorrectly translated as "meet".
Сизни тәбрикләшкә ижәзәт беринц (Uigur language)	No translation	Let me congratulate you	There is no Uyghur translation in Yandex. It accepts both Tatar language.
Мені жерге қаратпа (Kazakh language)	don't put me on the ground	don't put me on the ground	Translate phraseological units into a straight line
Бу тугрида гап хам булиши мумкин емас (Uzbek language)	It's all in tugrida and can not be found	This is out of the question	Yandex translation could not translate the word "tugrida" and completely lost the meaning of the sentence
Birsey icmek istiyorum (Turkish language)	I want to drink birsey	I want to drink something	In the translation of Yandex, the word "something" replaced by the word "birsey".

This article discusses training parallel corpus in Opennmt. The advantage of Opennmt is its universal applicability to different languages, including the Turkic languages.

We also get the results of computational experiments for the Kazakh language.

The rest of the paper is organised as follows:

- Section 2 provides an overview of previous work carried out in this area.

- Section 3 presents Opennmt for Kazakh-English, English-Kazakh language pairs.
- Section 4 presents experimental NMT results for Kazakh-English, English-Kazakh language pairs.
- Section 5 presents conclusions and suggests directions for future work.

2 Previous scientific work

In our country, in Kazakhstan, MT (machine translation) of the Kazakh language has been developing since 2000. Professor U. A. Tukeyev was one of the first to study machine translation. He managed to create a scientific school that is actively engaged in research in the field of MT. Among domestic students, one can note the study of models and methods of semantics of machine translation from Russian into Kazakh language [1], a statistical model of the alignment of English-Kazakh words using the machine translation algorithm [2].

To improve the morphologies of the Turkic languages, vocabulary training on a parallel corpus was used [3-6]. Learning on the parallel corpus of the English-German language pair in Opennmt, studied in foreign works. In this work, 50k vocabulary was learned for each pair and it was shown that in Opennmt Bleu was 17.60 [7]. Opennmt showed a better result than in the Nematus Bleu system by 0.5. Also, the architecture and applicability of Opennmt in other areas were considered.

3 Opennmt for Turkic languages

To build a neural network, the project uses the capabilities of the Torch deep machine learning library [8].

Opennmt in the model contains the following parameters [9-10]:

- encoder_type: transformer;
- decoder_type: transformer;
- position_encoding: true;
- enc_layers: 6;
- dec_layers: 6;
- heads: 8;
- rnn_size: 512;
- word_vec_size: 512;
- transformer_ff: 2048;
- dropout_steps: [0];
- dropout: [0.1];

- attention_dropout: [0.1];
- train_steps: 100000.

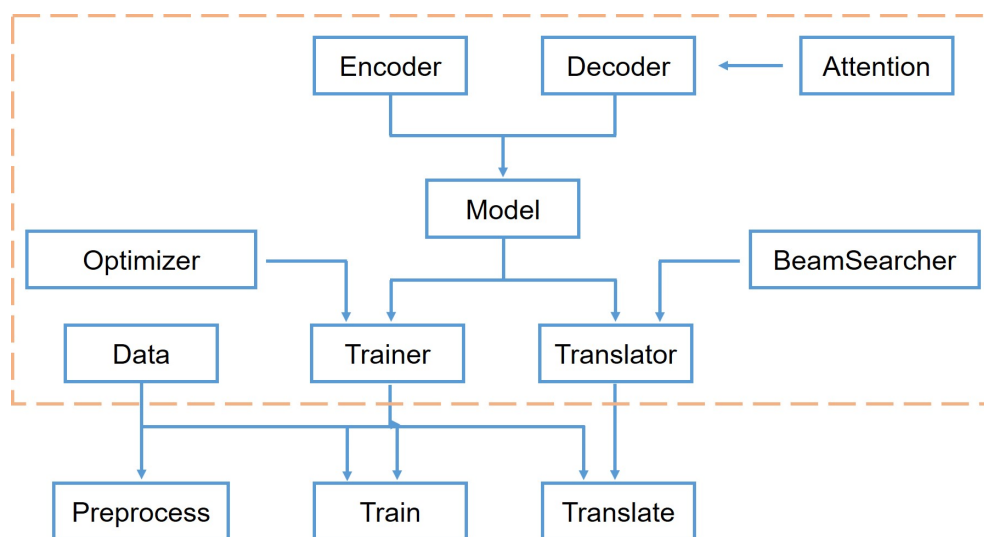


Figure 1: Building an Opennmt Model

4 Results

A Kazakh-English language pair was used for training. 109,772 sentences were used in the corpus. These proposals were taken from the website [11]: Akorda, Primeminister, mfa.gov.kz, economy.gov.kz, strategy2050.kz. Of these, it was taken for train 80000, test 20000, validation 9772. It took 36 hours to train at Opennmt.

Table 3. Obtained result in Opennmt for the English-Kazakh language pair

Language pair	Speed tok/sec	BLEU
Kazakh-English	4185	20.56
English-Kazakh	4185	20.05

As you can see in the table BLEU is less compared to other languages as for English-German, English-French pairs. Because the structure of the language of the Turkic languages is different from these languages. More parallel data is required to improve this metric.

5 Conclusion

This article covered learning parallel corpuses in Opennmt. In the experiment, it was used for the English-Kazakh language pair. To improve the translation, you need to add sentences

to the corpus for the English-Kazakh language pair. In the future, corpora will be developed for English-Turkic language pairs and training according to this system.

References

- [1] Moore R.C., "A discriminative framework for bilingual word alignment", *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT. Vancouver (2005): 81-88.
- [2] Bekbulatov, E. and Kartbayev A., "A study of certain morphological structures of Kazakh and their impact on the machine translation quality", *Proceedings of the IEEE 8th International*
- [3] Conference on Application of Information and Communication Technologies. Astana (2014): 495-501.
- [4] Nirenburg S., "Knowledge-Based Machine Translation", *Machine Translation*, Springer 1 (4) (1989): 5-24.
- [5] Nagao M., "A framework of a mechanical translation between Japanese and English by analogy principle", *Proceedings of the international NATO symposium on Artificial and human intelligence* (1984): 173-180.
- [6] Ziemski M., Junczys-Dowmunt M. and Pouliquen B., "The United Nations Parallel Corpus", *Proceedings of Language Resources and Evaluation LREC*. Slovenia (2016): 3530-3534.
- [7] Koehn P., "Europarl: A Parallel Corpus for Statistical Machine Translation", *Proceedings of the 10th Machine Translation Summit Phuket* (2005): 79-86.
- [8] Boitet C., "Bernard Vauquois' contribution to the theory and practice of building MT systems", *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering Beijing* (2010): 331-334.
- [9] Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig, "Linguistic Regularities in Continuous Space Word Representations", *The Association for Computational Linguistics. In HLTNAACL* (2013): 746-751.
- [10] Nal Kalchbrenner, Phil Blunsom, "Recurrent Continuous Translation Models", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA* (2013): 1700-1709.
- [11] Mikel L. Forcada and Ramon P. Neco, "Recursive Hetero-Associative Memories for Translation", *International Work-Conference on Artificial and Natural Neural Networks, IWANN'97 Lanzarote, Canary Islands, Spain* (1997): 453-462.