Neural machine translation system for the Kazakh language

Ualsher Tukeyev
Al-Farabi Kazakh National University
71 Al-Farabi ave., Almaty, 050040,
Kazakhstan
ualsher.tukeyev@gmail.com

Zhandos Zhumanov
Al-Farabi Kazakh National University
71 Al-Farabi ave., Almaty, 050040,
Kazakhstan
z.zhake@gmail.com

[1] Abstract

"Development and research of the Kazakh language neural machine translation system" is a 3-year-long project funded by the Science Committee of Ministry of Education and Science of the Republic of Kazakhstan. The project is being executed by the team of the Laboratory of Intelligent Information Systems at the Research Institute of Mathematics and Mechanics of al-Farabi Kazakh National University. Duration of the project: January 2018 – December 2020.

1 The purpose of the project

The purpose of the project is to create neural machine translation technology for the Kazakh language aiming at a high quality of machine translation, specifically adapted to the features of the Kazakh language.

Since 2013 the direction of machine translation based on recurrent neural networks, that is, neural machine translation, is intensively developing, and is actively explored for popular world languages. Practical applications of neural machine translation, in particular, in Google Translate, show impressive results. At the same time, neural machine translation research for the Kazakh language as a low resource language is a topical task. This project aims to fill that gap.

2 Project objectives

One problem that affects the quality of neural machine translation for Kazakh is the fact that parallel corpora with large volumes of data are not developed for the Kazakh language. Currently we have a small parallel corpus of approximately 140 000 Kazakh–English sentences, and a Kazakh–Russian corpus of a similar size. To solve this problem, in the project we will leverage on existing corpora of related languages.

The objectives of the project are investigation of:

- the basic version of the neural machine translation of the Kazakh language, using standard technology of NMT to Kazakh;
- the morphological segmentation for the Kazakh language NMT;
- the development of syntactic corpora for NMT of Kazakh;
- the development of models and algorithms for solving the problem of unknown words for the neural machine translation of the Kazakh language;
- the evaluation of the quality of the neural machine translation of the Kazakh language.

Our team had worked on a project titled "Development of free/open-source machine translation system for Kazakh–English and Kazakh–Russian (and vice versa) on the base of Apertium platform." in 2015–2017. Language resources created in the previous project are used in the current one [1, 2, 3].

Investigation of the basic version of the neural machine translation of the Kazakh language is based on the use of recurrent neural networks using the "encoder–semantic representation–decoder" model [4]. Along with that we will explore transformer architecture as well.

3 Expected results of the project

Project results will include technology (models, algorithms and software) of neural machine translation, adapted to the features of the Kazakh language. The system of neural machine translation of the Kazakh language will be developed as a free/open-source system.

4 Current results and future plans

The first part of the project is focused on Kazakh–English language pair. The second part will be focused on Kazakh–Russian language pair. At the end of the first year, the following results were achieved: the technology of hybrid automaton-neural machine translation of the Kazakh language based on the complete system of Kazakh language endings [5]; the preprocessor of morphological segmentation in Kazakh–English neural machine translation system; the postprocessor of morphological desegmentation in Kazakh–English neural machine translation system. Because of agglutinative nature words in Kazakh are formed by adding affixes. Different forms of the same word could be used in text and treated as different words if segmentation is not applied. The fact that the rules for adding affixes are very strict allows for creation of a complete system of Kazakh language endings, which simplifies segmentation and helps reducing vocabulary.

Current work is directed at gathering more parallel corpora by crawling multilingual web-sites with various tools, experimenting with different neural machine translation architectures, translation of unknown (out of vocabulary) words and integrating all of techniques that prove to be effective into one neural translation system. Our team consists from 7 members, described project have 45 million tenge (100 000 euro) on three year (2018–2020) funded by Kazakh government.

References

[1] Tukeyev U.A., Rakhimova D.R., Zhumanov Zh.M., Kartbayev A.Zh. Single state transducer model for Kazakh and Russian morphology // KazNU BULLETIN, Mathematics, Mechanics, Computer Science Series. – Алматы, «Қазақ университеті». – 2016. – №2 (89). – P. 110-117.

[2] Rakhimova D. R., Tukeyev U.A., Zhumanov Zh.M. Methodology of the automated enrichment of machine translation system dictionaries for Kazakh–Russian and Kazakh–English language pair // Proceedings of 4th International conference on Turkic languages ("TurkLang–2016"). – Bishkek, Kyrgyzstan, 2016. – C. 81-85.

[3] Zhumanov Zh., Madiyeva A., Rakhimova D. New Kazakh Parallel Text Corpora with On-line Access. Lecture Notes in Computer Science. – 2017. – 10449. – pp. 501-508.

[4] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. The Association for Computational Linguistics. In HLTNAACL, p. 746–751(2013).

[5] Tukeyev U., Sundetova A., Abduali B., Akhmadiyeva Zh., Zhanbussunov N. Inferring of the morphological chunk transfer rules on the base of complete set of Kazakh endings // LNAI 9876, Computational Collective Intelligence, Part 2, Springer, 2016, pp. 563-574