Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2018 vol.11107 LNAI, pages 163-170

Building the tatar-Russian NMT system based on re-translation of multilingual data

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

This paper assesses the possibility of combining the rule-based and the neural network approaches to the construction of the machine translation system for the Tatar-Russian language pair. We propose a rule-based system that allows using parallel data of a group of 6 Turkic languages (Tatar, Kazakh, Kyrgyz, Crimean-Tatar, Uzbek, Turkish) and the Russian language to overcome the problem of limited Tatar-Russian data. We incorporated modern approaches for data augmentation, neural networks training and linguistically motivated rule-based methods. The main results of the work are the creation of the first neural Tatar-Russian translation system and the improvement of the translation quality in this language pair in terms of BLEU scores from 12 to 39 and from 17 to 45 for both translation directions (comparing to the existing translation system). Also the translation between any of the Tatar, Kazakh, Kyrgyz, Crimean Tatar, Uzbek, Turkish languages becomes possible, which allows to translate from all of these Turkic languages into Russian using Tatar as an intermediate language.

http://dx.doi.org/10.1007/978-3-030-00794-2_17

Keywords

Data augmentation, Low-resourced language, Neural machine translation, Rule-based machine translation, Turkic languages

References

[1] ABBYY Aligner 2.0 (2017). https://www.abbyy.com/ru-ru/aligner/

[2] ABBYY SmartCAT tool for professional translators (2017). https://smartcat.ai/workspace

[3] Baisa, V.: Problems of machine translation evaluation. In: Sojka, P., s Horák, A. (eds.) Proceeding of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, Brno, pp. 17–22 (2009). https://nlp.fi.muni.cz/raslan/2009/papers/2.pdf

[4] Bojar, O., et al.: Findings of the 2017 conference on machine translation (WMT17). In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pp. 169–214. Association for Computational Linguistics, Copenhagen, September 2017. http://www.aclweb.org/anthology/W17-4717

[5] Bojar, O., et al.: Findings of the 2016 conference on machine translation. In: Proceedings of the First Conference on Machine Translation, pp. 131–198. Association for Computational Linguistics, Berlin, August 2016. http://www.aclweb.org/anthology/W/W16/W16-2301

[6] Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2017), pp. 567–573, January 2017

[7] Moses, the machine translation system (2017). https://github.com/moses-smt/mosesdecoder/

[8] Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318 (2002)

[9] Schiebinger, L., Klinge, I.: Gendered Innovations: How Gender Analysis Contributes to Research. Publications Office of the European Union, Luxembourg (2013)

[10] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. ArXiv e-prints, August 2015

[11] Sennrich, R., Haddow, B., Burch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, pp. 86–96 (2016)

[12] Sennrich, R., et al.: The University of Edinburgh’s neural MT systems for WMT17. In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. Stroudsburg, PA, USA (2017)

[13] Subword Neural Machine Translation (2017). https://github.com/rsennrich/subword-nmt/

[14] Suleimanov, D., Gatiatullin, A., Almenova, A., Bashirov, A.: Multifunctional model of the Turkic morpheme: certain aspects. In: Proceedings of the International Conference on Computer and Cognitive Linguistics TEL-2016, Kazan, pp. 168–171 (2016)

[15] Open-Source Neural Machine Translation in Theano (2017). https://github.com/rsennrich/nematus

[16] Wu, Y., et al.: Google’s neural machine translation system: bridging the gap between human and machine translation. ArXiv e-prints, September 2016

[17] Yandex translate (2017). https://translate.yandex.com/

[18] One model is better than two. Yandex. Translate launches a hybrid machine translation system (2017). https://goo.gl/PddtYn

[19] Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. ArXiv e-prints, April 2016