UDC 81'322.4 DOI: 10.33910/2687-0215-2019-1-2-365-370

The impact of some linguistic features on the quality of neural machine translation

E. A. Shukshina✉[1]

[1] Saint Petersburg State University, 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Abstract. This paper investigates how different features influence the translation quality of a Russian-English neural machine translation system. All the trained translation models are based on the OpenNMT-py system and share the state-of-the-art Transformer architecture. The majority of the models use the Yandex English-Russian parallel corpus as training data. The BLEU score on the test data of the WMT18 news translation task is used as the main measure of performance. In total, five different features are tested: tokenization, lowercase, the use of BPE (byte-pair encoding), the source of BPE, and the training corpus. The study shows that the use of tokenization and BPE seems to give considerable advantage while lowercase impacts the result insignificantly. As to the BPE vocabulary source, the use of bigger monolingual corpora such as News Crawl as opposed to the training corpus may provide a greater advantage. The thematic correspondence of the training and test data proved to be crucial. Quite high scores of the models so far may be attributed to the fact that both the Yandex parallel corpus and the WMT18 test set consist largely of news texts. At the same time, the models trained on the Open Subtitles parallel corpus show a substantially lower score on the WMT18 test set, and one comparable to the other models on a subset of Open Subtitles corpus not used in training. The expert evaluation of the two highest-scoring models showed that neither excels current Google Translate. The paper also provides an error classification, the most common errors being the wrong translation of proper names and polysemantic words.

Keywords: machine translation, neural machine translation, neural networks, transformer, translation evaluation, translation quality, tokenization, training corpus, byte-pair encoding, Yandex parallel corpus, Yandex corpus, WMT18 test set, news texts, Yandex.Translate, BLEU score.

Introduction

In 2016 Google launched its machine translation system based on neural networks (Turovsky 2016) that significantly improved the quality of translation. It was a milestone in the development of the field as shortly afterwards most translation companies were seeking to introduce it in their systems too. Next year, in 2017, Yandex.Translate also implemented neural networks.

A neural network consists of simple processors that can receive data, perform simple operations, and convey the result to other neurons. They are usually organized in layers: the input layer, the output layer, and the hidden layers in between them. The data is transmitted from one layer to the next in feed-forward neural networks, while recurrent neural networks (RNN) have 'loops' that enable information to be transmitted backwards as well.

Until the Transformer architecture was introduced in (Vaswani et al. 2017), the dominant neural machine translation models were based on RNN used in the encoder-decoder architecture (Sutskever, Vinyals, Le 2014) with an attention mechanism (Bahdanau, Cho, Bengio 2015) that vaguely corresponds to alignment.

The Transformer is based solely on the attention mechanism and does not employ a recurrent network structure. Just as earlier models, it consists of the encoding and decoding components. The novelty is in the use of 'self-attention' layers that allow to find connections between the words

in a sentence, and the 'encoder-decoder attention' layers that are concerned with the correspondence between the input and the output sequence.

The paper explores the five features the quality of translation depends on: the use of lowercase, tokenization, and BPE (byte-pair encoding), the source of BPE, and the training corpus.

Setup of the experiment

The models under study are based on the OpenNMT-py open machine translation system that provides a variety of tools for preprocessing the data as well as for training and testing translation models. The experiment is run on the Yandex en-ru bilingual corpus that contains one million aligned sentences automatically extracted from the web.

To evaluate the models, we use BLEU score on 3,000 test sentences of WMT18 news translation task (Bojar et al. 2018).

We compare several models that differ in the number of preprocessing steps that were applied to the training data:

1. Tokenization — provided by the Moses tokenizer distributed as a part of OpenNMT-py;
2. Lowercase;
3. BPE (Sennrich, Haddow, Birch 2016) — an approach to segment a text into subword units based on their co-occurrence frequencies.

All the models share the same transformer architecture with 6 layers of decoding and encoding inspired by (Vaswani et al. 2017) except for the multi-GPU feature that was not used in our setup.

Results

Table 1. BLEU scores for all possible combinations of three preprocessing steps: tokenization, lowercase, BPE (learnt from the training data)

Model	Tokenization	Lowercase	BPE	BLEU score
1	0	0	0	14.86
2	1	0	0	19.71
3	0	1	0	15.50
4	1	1	0	20.74
5	0	0	1	21.57
6	1	0	1	23.32
7	0	1	1	21.81
8	1	1	1	24.82

As we can see, the least helpful step of the three in question is lowercase for it increases the BLEU score of the system only by 0.85 points on average while tokenization and BPE have a much greater impact on the score increasing it on average by 3.7 and 5.2 points respectively.

Using a different corpus for learning BPE

The models with BPE presented above train BPE on the training data. However, its size may be insufficient to provide relevant vocabulary for the test data. The obvious step is to extract BPE vocabulary from larger monolingual corpora. This is supposed to provide a more general BPE vocabulary for each language that would not be specific to the training data.

For this purpose, we chose News Crawl with 8,233,935 sentences for Russian and 26,861,180 sentences for English. For better results, we deleted all the sentences that have no Cyrillic characters from the Russian News Crawl corpus, which reduced its size to 7,879,149 sentences.

Table 2 shows how 30,000 new BPE vocabularies impacted the tokenized and lowercased data.

Table 2. Results of the experiments with the BPE source and the training corpus. BLEU* stands for BLEU score measured on a part of OpenSubtitles corpus not used in training

Model	Tokenization	Lowercase	BPE	BLEU WMT 18	BLEU*
9	yes	yes	News Crawl	25.18	
10	yes	yes	News Crawl	17.85	26.50
11	yes	yes	Open Subtitles	16.02	25.72

Training on a bigger corpus

The size of the Yandex corpus is both an advantage, as it increases the speed of our experiments, and a disadvantage. To test performance on a bigger corpus, we used the Open Subtitles corpus that contains 25,910,105 Russian-English sentence pairs. We trained two models that differ in the source of the BPE vocabulary applied to the training data — News Crawl corpus for model 10 and the training data itself for model 11.

The results presented in Table 2 indicate that the use of a separate corpus for the extraction of BPE vocabulary proves to be more advantageous. The lower results of the models on the WMT 18 test data may be due to the difference in subject and register of the training and test data. The Open Subtitles corpus contains sentences of a more colloquial style, while WMT 18 test data is in line with the task through its focus on news text translation. To illustrate this difference, we also provide the BLEU score obtained on the test portion of the Open Subtitles corpus that was not used in training (BLEU* column of Table 2), that happens to be comparable to that of previous models.

Human evaluation

We decided to have a closer look at the translations provided by the two highest ranked models (models 9 and 10). For our evaluation we took 100 sentences randomly sampled from the test data. The performance of our models was compared to that of Google Translate. The raters were asked to rank the translations provided by the two models and Google Translate. They were given the input text and the reference translation from the test data. The rater could give the same rank to the sentences if they were equally good or bad. The results are presented in Table 3. It is clearly seen that neither of our models could excel Google translate.

Table 3. Results of human evaluation of models 9 and 10. Average comparison scores with Google Translate pairwise

	v9 vs Google	v10 vs Google	v9 vs v10
better	21	20	45
worse	49	56	30
equal	30	24	25

Error analysis

We also decided to examine the errors that occur in the translations of model 9 more thoroughly and sort them into the types that loosely correspond to the classification provided in (Vilar et al. 2006).

In the output of our model we found the following error types:

1. Missing words

a. Missing part of sentence: 12 sentences (8% of the total error count).

Полчища владельцев прогулочных корабликов и артистов, изображающих Статую Свободы, линчевали бы его, если бы он попробовал это сделать.	*A horde of boat-trip owners and Liberty impersonators would have lynched him if he did.*	*he would have done it if he tried to do it.*

b. Missing content words: 23 sentences (15% of the total error count).

*На мой взгляд, Коулмен **сегодня** один из самых выдающихся барабанщиков мира.*	*In my opinion, Coleman is one of the most accomplished drummers in the world **today**.*	*in my opinion, coleman is one of the world's most prominent drummers.*

c. Missing filler words: 13 sentences (8% of the total error count).

*Все в порядке, — шепчет одна **из** женщин.*	*It's okay, one **of** the women whispers.*	*everything is fine, one woman whispers.*

2. Incorrect words

a. Mistranslated proper names: 30 sentences (20% of the total error count). The majority were transliterated quite closely (*Ameliya Chesca* instead of *Amelia Chasse*, *Ranan Raffferti* instead of *Ronan Rafferty*). Some names were translated as if they were common nouns (*the wolves* instead of *Volkov*, *orphan* instead of *Sirotin*), and very few were far from correct (*Gennady Chelsi* instead of *Rod Chapel*).

b. Wrong sense of the word: 49 sentences (32% of the total error count).

Мы также серьезно относимся к своему уставу и к власти, который он нам дает».	*We also take the statute and the authority it gives us seriously."*	*we are also seriously **concerned with** our charter and the power it gives us."*

c. Wrong form of the word: 13 sentences (8% of the total error count).

...но он получил повестку, обязывающую его явиться в отделение полиции...	*...but he received a summons obliging him to appear at the police station...*	*...but he was given **a** agenda that would oblige him to appear in the police department...*

3. Word order errors — including word and phrase level reordering — were found in 13 sentences and correspond to 8% of the total error count.

По словам Пола, и Люк, и Марк были «недовольны финансовыми условиями своего отделения».	*Both Luke and Mark had become, Paul says, "bitter about the terms of their financial separation."*	***paul and luke said that mark** were "dissatisfied with the financial conditions of his division."*

Conclusion and future work

We investigated the impact of five different features on the quality of neural machine translation. The application of tokenization and BPE leads to a drastic growth in BLEU score. It is more effective to use larger monolingual corpora for BPE training. The use of lowercase does not seem to provide the advantage significant enough to compensate for missing capitalization

in proper names, abbreviations and beginnings of sentences. The study shows that thematic and register correspondence between the training corpus and the intended use of the system is quite important. This implies that a general-purpose translation system must be trained on a large representative parallel corpus with texts in different styles and registers as well as a wide range of topics.

It is worth mentioning that these conclusions are drawn from a single study based on Russian-English translation. All the statements remain to be verified for other language pairs, which is something we will focus on in our future work.

Sources

Anglo-russkij parallel'nyj korpus (versiya 1.3). [Online]. Available at: https://translate.yandex.ru/corpus (accessed 15.08.2019). (In Russian)

Index of /news-crawl. [Online]. Available at: http://data.statmt.org/news-crawl/ (accessed 11.09.2019). (In English)

OpenSubtitles.org. [Online]. Available at: http://www.opensubtitles.org (accessed 13.08.2019). (In Russian)

References

Bahdanau, D., Cho, K., Bengio, Y. (2015) Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473v7*. [Online]. Available at: https://arxiv.org/abs/1409.0473 (accessed 15.08.2019). (In English)

Barrault, L., Bojar, O., Costa-jussà, M. R. et al. (2019) Findings of the 2019 Conference on Machine Translation (WMT19). In: *Proceedings of the Fourth Conference on Machine Translation (WMT). Vol. 2: Shared Task Papers (Day 1). Florence, Italy, August 1–2, 2019.* Stroudsburg, PA: Association for Computational Linguistics, pp. 1–61. (In English)

Bojar, O., Federmann, Ch., Fishel, M. et al. (2018) Findings of the 2018 Conference on Machine Translation (WMT18). In: *Proceedings of the Third Conference on Machine Translation (WMT). Vol. 2: Shared Task Papers. Brussels, Belgium, October 31 — Novermber 1, 2018.* Stroudsburg, PA: Association for Computational Linguistics, pp. 272–307. (In English)

Lison, P., Tiedemann, J. (2016) OpenSubtitles 2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, May 23–28, 2016.* Pp. 923–929. [Online]. Available at: http://www.lrec-conf.org/proceedings/lrec2016/summaries/947.html (accessed 13.08.2019). (In English)

One model is better than two. Yandex.Translate launches a hybrid machine translation system. (2017) *Yandex Blog*. 14 September. [Online]. Available at: https://yandex.com/company/blog/one-model-is-better-than-two-yu-yandex-translate-launches-a-hybrid-machine-translation-system (accessed 15.08.2019) (In English)

Sennrich, R., Haddow, B., Birch, A. (2016) Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, August 7–12, 2016.* Vol. 1. Stroudsburg, PA: Association for Computational Linguistics, pp. 1715–1725. (In English)

Sutskever, I., Vinyals, O., Le, Q. V. (2014) Sequence to Sequence Learning with Neural Networks. In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Red Hook, NY: Curran Associates, pp. 3104–3112. (In English)

Turovsky, B. (2016) Found in translation: More accurate, fluent sentences in Google Translate. *Translate. News about Google Translate.* 15 November. [Online]. Available at: https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/ (accessed 15.08.2019). (In English)

Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, California, USA, 4–9 December 2017.* Red Hook, NY: Curran Associates, pp. 5998–6008. (In English)

Vilar, D., Xu, J., D'Haro, L. F., Ney, H. (2006) Error analysis of statistical machine translation output. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy, May 22–28, 2006.* Pp. 697–702. [Online]. Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf (accessed 10.08.2019). (In English)

Author:
Elena A. Shukshina, ORCID: 0000-0002-6014-9136, e-mail: elena.shukshina@gmail.com

For citation: Shukshina, E. A. (2019) The impact of some linguistic features on the quality of neural machine translation. *Journal of Applied Linguistics and Lexicography*, 1 (2): 365–370. DOI: 10.33910/2687-0215-2019-1-2-365-370

Received 29 August 2019; reviewed 11 September 2019; accepted 12 September 2019.