Neural Machine translation from Russian into English

Vyacheslav Shults, Aleksei Zhuchkov, Susanna Gimaeva

I. Framing the problem

Neural machine translation (NMT) is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model. The advent of NMT certainly marks one of the major milestones in the history of machine translation [1]. This report describes our work on neural machine translation from Russian into English.

II. Data exploration and preparation

As a final solution we took a pretrained model created by Facebook for WMT19 News Translation Task Submission [2] without further fine-tuning, thus we did not use any training data.

However, it is worth mentioning what data was used for pretraining of the model we chose as described by [2].

The authors used all available bitext data, and for their monolingual data they used English and Russian Newscrawl as well as a filtered portion of Russian Commoncrawl for augmentation. For preprocessing they normalized punctuation and tokenized all datawith the Moses tokenizer (Koehn et al.,2007), then they learnt separate BPE encodings with 24K split operations for each (Russian, English) language. To filter the data, Nathan N. et.al. applied language identification filtering(langid; Lui et al., 2012), keeping only sentence pairs with correct languages on both sides and removed sentences longer than 250 tokens as well as sentence pairs with a source/target length ratio exceeding 1.5.

III. Model description

The baseline system of the model [2] is large BPE-based transformer model [3] trained with the FAIRSEQ [4] sequence modeling toolkit which rely on sampled back-translations.

The authors of [2] tried to increase the base system network capacity by increasing embed dimension, FFN size, number of heads, and number of layers. They reported that using a larger FFN size (8192) gives a reasonable improvement in performance while maintaining a manageable network size. Thus, all subsequent models, including ensembles and the model we utilized, used this larger FFN Transformer architecture.

Moreover, this approach utilizes beam search. It is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set and uses breadth-first search to build its search tree. Once reaching the configured maximum search depth (i.e. translation length), the algorithm will evaluate the solutions found during the search at various depths and return the best one (the one with the highest probability). Fine-tuning of its parameters gives a score gain of 0.2.

The authors stated that for their experiments with langid filtering, larger FFN, and ensembling performance of a bitext-only system improved by 1.6 BLEU. Back-translation added 3 BLEU, most likely due to the lower quality bitext data available, they assume. Finetuning, ensembling, and noisy channel model approach reranking added almost 4 BLEU, with reranking contributing 1.2 BLEU.

IV. Experiments

The first set of our experiments included trying existing models for NMT pretrained for Russian-English pair. We mainly explored Transformer-based architectures for this task, as these architectures provide high performance results. Among a variety of such pretrained models, Russian language models are quite rare. Finally, we took WMT19 News Translation Task Submission [2]. We additionally experimented with the default parameters given to the model. Specifically, we tried various beam size parameters, but failed to find a better value than the one reported by [2], which is 50.

The second set of our experiments was devoted to fine-tuning of the chosen model on UN Parallel Corpus and ParaCrawl Corpus, mainly following the official guideline of huggingface transformers repository. However, we were not able to finish this set of experiments as the documentation of the model is not exhaustive and the experiments required a lot of troubleshooting.

V. Results and evaluation

The examples of translation from Russian into English are comparable to human translation: (original punctuation and grammar is left)

Source language: Russian	Target language: English
С учетом изложенного выше Япония хотела бы предложить следующие элементы заключительных документов обзорной конференции	In view of the foregoing, Japan would like to propose the following elements of the final documents of the 2015 Review Conference for further consideration by the States parties.

Source language: Russian	Target language: English
2015 года для дальнейшего обсуждения государствами-участниками	
Главными движущими силами благого управления должны быть уважение закона, соблюдение принципов диалога и сотрудничества, защита демократии и поощрение прав человека.	Respect for the rule of law, adherence to the principles of dialogue and cooperation, the protection of democracy and the promotion of human rights should be the main drivers of good governance.

As for numeric evaluation of the model performance, we used BLEU score [5], which is considered to be highly correlated to human judgements of translation quality [6]. BLEU scores are calculated by comparing generated translations with reference translations and averaging the results. This value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts and score equal to 1 for identical ones.

Our model achieved BLEU score of 52.02 % on the first provided test set, and 50.95 % on the second one accordingly. The test data and results can be found in the GitHub repository alongside the code.

VI. Conclusion

In conclusion, we achieved results highly similar to human experts. If further increase of the accuracy is required, fine-tuning on a large and diverse datasets (or domain-specific datasets, depending on the downstream task) deserves attention. Moreover, despite the BLEU score being one of the most popular automated and inexpensive metrics with high correlation to human percieved quality of text, other metrics can give additional information about the quality of generated texts.

References

[1] F. Stahlberg, "Neural machine translation: A review," Journal of Artificial Intelligence Research, vol. 69, pp. 343–418, 2020.

[2] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook fair's wmt19 news translation task submission," arXiv preprint arXiv:1907.06616, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.

[4] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," arXiv preprint arXiv:1904.01038, 2019.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[6] D. Coughlin, "Correlating automated and human assessments of machine translation quality," in Proceedings of MT summit IX. Citeseer, 2003, pp. 63–70.