Development and Evaluation of an English-to-Russian Neural Machine Translation System

Dominique S. Loyer
PhD student and AI researcher
UQAM, April 2025
loyer.dominique@uqam.ca

Abstract

This report presents the development and evaluation of an English-to-Russian (EN→RU) Neural Machine Translation (NMT) system. Using the Hugging Face Transformers library (Wolf et al., 2020) and a pre-trained model, `Helsinki-NLP/opus-mt-en-ru` (Junczys-Dowmunt et al., 2018), based on the Transformer architecture (Vaswani et al., 2017), we fine-tuned the model on a subset of 10,000 sentence pairs from the `opus_books` parallel corpus (Tiedemann, 2012). The system was trained for 3 epochs in a Kaggle environment (T4 GPU). Evaluation on a held-out test set from the same corpus yields a SacreBLEU score of 22.6 and a chrF score of 48.5.

Keywords: Neural Machine Translation, EN→RU, Transformer, Fine-tuning, SacreBLEU, chrF.

1 Introduction

Machine Translation (MT) aims to translate text from a source language to a target language without human intervention. Historically dominated by statistical approaches (SMT), MT was revolutionized by Neural Machine Translation (NMT) in the mid-2010s (Bahdanau et al., 2014). This project:

- implements a complete pipeline (data, fine-tuning, evaluation),
- utilizes the `Helsinki-NLP/opus-mt-en-ru` model (Junczys-Dowmunt et al., 2018),
- leverages 10,000 sentence pairs from the `opus_books` corpus (Tiedemann, 2012).

We measure translation quality using SacreBLEU and chrF and compare our results to the state of the art.

2 Related Work

Early NMT architectures combined Recurrent Neural Networks (RNNs) with attention mechanisms (Bahdanau et al., 2014). The Transformer architecture (Vaswani et al., 2017) subsequently established itself as the dominant paradigm due to its efficiency and performance. The MarianMT models (Junczys-Dowmunt et al., 2018), pre-trained on the extensive OPUS corpus, provide a solid foundation for fine-tuning. State-of-the-art results for EN→RU, as reported in WMT campaigns and by models like NLLB-200 (NLLB Team et al., 2022), typically achieve BLEU scores in the 30–45 range.

3 Methodology

3.1 General Pipeline

The process follows these steps:

1. **Data Loading:** The `opus_books` EN–RU parallel corpus (Tiedemann, 2012) is loaded via the `datasets` library (Lhoest et al., 2021).
2. **Preprocessing:** The data undergoes cleaning and tokenization with a maximum sequence length of 128 tokens.
3. **Model Loading:** The pre-trained `Helsinki-NLP/opus-mt-en-ru` model (Junczys-Dowmunt et al., 2018) is loaded.

4. **Fine-tuning:** The model is fine-tuned for 3 epochs with a batch size of 8, a learning rate of 2×10^{-5}, and mixed-precision (fp16) training using the `Seq2SeqTrainer`.

5. **Evaluation:** Performance is measured using SacreBLEU and chrF with the `evaluate` library (Lhoest et al., 2021).

3.2 Environment

The development process was significantly streamlined by leveraging the Hugging Face ecosystem, particularly the Transformers library. The experiments were conducted on a Kaggle Notebook equipped with a Tesla T4 GPU, which provided adequate computational power for fine-tuning the model within a reasonable timeframe.

4 Data

The `opus_books` corpus (Tiedemann, 2012) was chosen for its accessibility and parallel nature. The dataset was split into training, validation, and test sets to evaluate the model's generalization capabilities and prevent overfitting. The training set size of 10,000 pairs is relatively small by modern standards for large-scale MT tasks.

5 Results

The system achieved a SacreBLEU score of 22.6 and a chrF score of 48.5 on the held-out test set. These metrics provide a quantitative measure of the system's performance.

5.1 Discussion

The SacreBLEU score of 22.6 indicates a moderate translation quality. While this score is not competitive with state-of-the-art results for EN→RU (BLEU 30–45), it is a respectable outcome given the limited training data (10,000 pairs). The high-performing models mentioned in the related work section were trained on vastly larger corpora.

The chrF score of 48.5, which measures character-level similarity, corroborates this finding. A direct comparison with state-of-the-art systems, which primarily report BLEU, is difficult without corresponding chrF values.

6 Future Work

While the results are encouraging given the data constraints, several avenues for improvement exist:

- **Data Augmentation:** Increasing the training data size using larger corpora like WMT (Bojar et al., 2016) or ParaCrawl (Barrault et al., 2019) is a clear next step. Techniques like back-translation (Sennrich et al., 2016) could also be employed to generate synthetic training data.

- **Advanced Models:** Experimenting with larger, more powerful models such as NLLB-200 (NLLB Team et al., 2022), mBART (Liu et al., 2020), or T5 (Raffel et al., 2020) could yield significant performance gains.

- **Hyperparameter Optimization:** A more extensive search for optimal fine-tuning hyperparameters (e.g., learning rate, batch size, number of epochs) could further improve performance.

- **Qualitative Analysis:** A detailed qualitative analysis of translation errors would provide deeper insights into the model's weaknesses and guide future improvements.

Limitations

The primary limitation of this work is the small scale of the training data. Using only 10,000 sentence pairs from the `opus_books` corpus restricts the model's ability to learn the full complexity and diversity of the English and Russian languages. Consequently, the model may struggle with out-of-domain sentences, complex syntactic structures, and rare vocabulary. The maximum sequence length of 128 tokens may also lead to truncation and poor translation of longer sentences. The evaluation is conducted on a test set from the same domain as the training data, which may not reflect performance on other text genres. Finally, this study is limited to a quantitative analysis; a qualitative error analysis would be necessary to fully understand the system's behavior.

Ethics Statement

This research was conducted using publicly available datasets and open-source models. The

opus_books corpus consists of translated literary works, and its use for research purposes is standard practice in the field. The pre-trained model, Helsinki-NLP/opus-mt-en-ru, was released by the Helsinki-NLP group for public use. We acknowledge that machine translation systems can perpetuate or amplify existing biases present in their training data. While a bias analysis was outside the scope of this project, we recognize it as an important consideration for the deployment of any MT system.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Thomas Etchegoyhen, Pascale Bougares, Benjamin Cartoni, Laurent Maillard, Julien Carré, Samuel Bordes, Carolina Scarton, Xavier Tannier, Laurent Guérin, and Marek Šimko. 2019. ParaCrawl: Large Scale Extracted Parallel Corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5892, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Petya Nœv, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2016. Findings of the 2016 conference on machine translation (WMT16). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 1–30, Berlin, Germany. Association for Computational Linguistics.

Gemma Team and Google. 2024. Gemma: Open models based on gemini research and technology. https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf. Accessed: April 2025.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.

Quentin Lhoest, Albert Villanova, Patrick von Platen, Mariama Drame, Yacine Jernite, Julien Plu, Clara Ma, Lewis Tunstall, Nicolas Patry, Clement Delangue, Julien Chaumond, Victor Sanh, Thomas Wolf, and Lysandre Debut. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xuan Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 178–195.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth He, Kevin Hoffman, Marcin Junczys-Dowmunt, Philipp Koehn, Varvara Lakshminarasimhan, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04671*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Com-*

putational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.