MUNI-NLP Submission for Czech-Ukrainian Translation Task at WMT23

Pavel Rychlý and **Yuliia Teslia**
NLP Centre, Faculty of Informatics, Masaryk University
pary@fi.muni.cz, 531354@mail.muni.cz

Abstract

The system is trained on officially provided data only. We have heavily filtered all the data to remove machine translated text, Russian text and other noise. We use the DeepNorm modification of the transformer architecture in the TorchScale library with 18 encoder layers and 6 decoder layers. The initial systems for back-translation uses HFT tokenizer, the final system uses custom tokenizer derived from HFT.

1 Introduction

The annual Conference in Machine Translation (WMT) provides an invaluable platform for researchers to showcase their advancements in this domain. This paper serves as the submission from the Natural Language Processing Centre of Masaryk University (MUNI-NLP) team for the Czech-Ukrainian Translation Task at WMT23 (Kocmi et al., 2023).

Central to our approach is a commitment to data quality and we have also done some experiments with different subword tokenizers. Furthermore, we recognize the paramount importance of data filtering, a process that distinguishes signal from noise. To this end, we employ a rigorous data filtering procedure to eliminate machine-translated text, Russian content, and other sources of potential distraction, enabling our model to focus on the genuine linguistic intricacies of Czech and Ukrainian.

In terms of model architecture, we employ the DeepNorm modification (Wang et al., 2022) of the transformer architecture. This modified architecture, integrated within the TorchScale library, boasts 18 encoder layers and 6 decoder layers.

Our approach further delves into the critical aspects of back-translation and tokenization. Initially, we leverage the HFT (High Frequency Tokens) tokenizer for back-translation, harnessing the power of synthetic data to enhance model robustness. In the final iteration of our system, we introduce a custom tokenizer derived from HFT.

2 Data selection and preprocessing

Our system participates in the constrained track, we use only data allowed for this year. We do not use any pretrained models, only selected parallel and monolingual texts.

Many of the provided text are very noisy. We excluded some of them from training completely.

To mitigate the adverse effects of noisy data on our translation system, we conduct a comprehensive analysis of the data problems that emerge from these issues. In subsequent sections of this paper, we delve into the specific strategies and techniques we employ to filter out machine-translated text, Russian content, and other sources of noise. Our data filtering pipeline is designed to rigorously curate the training data set, ensuring that our model is exposed to high-quality, human-generated translations that align closely with the nuances of the Czech and Ukrainian languages. Through these meticulous filtering strategies, we aim to enhance the overall performance and translation quality of our system.

2.1 Parallel data

We use only official Task 1 data downloaded by the `mtdata` command (Gowda et al., 2021). The majority of segments comes from the OPUS corpus (Tiedemann, 2012), the biggest single source is Facebook-wikimatrix (Schwenk et al., 2019).

OPUS-opensubtitles Sometimes contains wrong or missing diacritics in Czech part. We removed segments containing meta data like authors of the subtitles. There are many parts on the Ukrainian side with Russian language instead of Ukrainian. We have removed such segments.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 162–165
December 6–7, 2023.

Source	Used segments	Original segments	Used words
ELRC-acts-ukrainian	130003	130003	2.5M
OPUS-ccmatrix	3916740	3991954	44M
OPUS-opensubtitles	515216	730804	2.7M
OPUS-multiparacrawl	941349	2200276	12M
OPUS-qed	155346	161020	2M
OPUS-tatoeba	2932	2933	11k
OPUS-ted2020	112689	114229	1.6M
Facebook-wikimatrix	824606	848961	9.9M
Total	6602828		

Table 1: The sizes of all sources used for the final system. *Used words* means number of words used in one language, these are almost same for both languages.

In total, almost 30% of segments were removed from this source.

Facebook-wikimatrix Many segments are not aligned, they contains similar texts but the sentences are not translations. We can see such situations in sentences about different sport teams, towns and history persons.

We have removed segments with special formatting options, lines containing *Dostupné online (available online)* and similar strings.

OPUS-wikimedia Removed HTML formatting, notes in tested parenthesis which are not translated anyway.

Removed segments containing URL, references to online sources.

Removed segments with Czech texts in Ukrainian part and vice versa.

OPUS-multiccaligned Excluded from processing.

The Czech part contains almost exclusively a very bad machine translations in domains like: game playing, health recommendations, porn, bitcoins, garden.

Only a few good Czech sentences are copied from Czech Wikipedia.

OPUS-bible Excluded from processing.

It contains very old language with unusual vocabulary and grammar.

OPUS-elrc-5179 Excluded from processing.

The same text as ELRC-acts with some errors (missing characters).

OPUS-eubookshop Excluded from processing.

Contains concatenated words on both sides.

Several sources contains duplicated segment, we keep only the first instance of such duplicates.

The sizes of all parallel sources used for the final system are listed in the Table 1.

2.2 Monolingual data

Statmt-news-crawl-2021-ces Removed time indications at the beginning of lines.

LangUk-* There is no punctuation in text.

We have used simple rules to add probable punctuation marks.

Removed markdown formatting.

We use additional filtering of back-translated segments. We use `filter-acktranslation.py` from (Popel et al., 2022). Unfortunately, there were still some Russian texts at the early stages of development and some Czech sentences were translated into Russian instead of Ukrainian. We filtered such segments out for the final system.

The sizes of monolingual data are listed in Table 2.

3 Tokenization

We use HFT tokenizer (Signoroni and Rychlý, 2022) for all stages. The tokenizer uses special characters to annotate word boundaries and character capitalization, they are listed in Table 3.

An example of tokenized text from the Czech part of the training data is in Figure 1. All uppercase letters are transformed to lower-case and the special characters preserve the original format. White spaces around punctuation marks are annotated explicitly in the same way as in Sentencepiece

```
A on: "Ne, tady jsme už asi rok."
↑¦a¦ ¦on¦ :▁" ↑¦ne¦ ,▁ ¦tady¦ ¦jsme¦ ¦už¦ ¦asi¦ ¦rok¦ ."

Trvalo 17 dní, než US senát zakázal
↑ ¦trvalo¦ ¦17¦ ¦dní¦ ,▁ ¦než¦ Δ¦us¦∇ ¦sen át¦ ¦zak ázal¦

Ona napsala: " MŮJ BANKÉŘ BUDE POTŘEBOVAT PŘEVOD 3 000$.
↑¦ona¦ ¦napsala¦ :▁ "▁ Δ ¦můj¦ ¦bank é ř¦ ¦bude¦ ¦potřebovat¦ ¦převo d¦∇ ¦3¦ ¦000¦ $ .
↑ ¦ona¦ ¦napsala¦ :▁ "▁ Δ ¦můj¦ ¦bank é ř¦ ¦bude¦ ¦potřebovat¦ ¦převo d¦ ∇ ¦3¦ ¦0 0 0¦ $ .
```

Figure 1: Example of the tokenization. The first line of each group is the plain text, the second line is the respective tokenization. The very last line is the modified tokenization used in the final system.

Source	Used segments
Leipzig-news	1M
Leipzig-newscrawl	1M
Leipzig-wikipedia	1M
Statmt-news-crawl.ces	11M
LangUk-ubercorpus	22M
LangUk-news	15M
Total UK	41M
Total CS	15M

Table 2: The sizes of all sources used for back-translation.

¦	<token-delimiter>
↑	<single-uppercase>
▁	<explicit-whitespace>
∇	<all-uppercase>
Δ	<end-of-uppercase>

Table 3: Special characters in the HFT tokenizations.

(Kudo and Richardson, 2018), but spaces between words are assumed as default.

For the final system we have made the following extra changes in tokenization:

- Separate special capitalization symbols from tokens, they are always separate tokens.
- Split numbers into digits.

An example of this changes is displayed on the very last line in the Figure 1, the first token is split into two tokens, the number “3 000” at the end of the line is tokenized into two tokens in the original tokenization and into four tokens in the final one.

For the initial systems for translation we use vocabulary size of 32,000 items. The final translations system use only 12,000 items in the vocabulary on each side.

These modifications in the final system were motivated by an experiment on smaller data where BLEU score increased from 22.4 to 24.9. Separating individual digits is also an option (disabled by default) in the Sentencepiece (Kudo and Richardson, 2018) tokenizer. We will do a detailed evaluation of this modifications in the future.

4 Model

We use the DeepNorm (Wang et al., 2022) modification of the transformer (Vaswani et al., 2017) architecture in the TorchScale library (Ma et al., 2022). Our early experiments with the number of encoder and decoder layers shows with the agreement of Wei et al. (2022) that asymmetric configuration with more encoder layer performs better. We use 18 encoder layers and 6 decoder layers in all our models.

The first stage of the system in CS-UK direction is trained only on parallel data (6.6M segments) for 30 epochs, second stage in UK-CS direction uses also 15M Czech segments and is trained for 17 epochs. The final system uses parallel data and back-translated Ukrainian monolingual data (41M segments). It is trained for only 4 epochs. Checkpoints are created every 2000 updates and the final submission is the average of 8 checkpoints (1 following and 6 preceding) around the top-scoring checkpoint on development data.

The performance of the individual models are detailed in Table 4.

Stage	direction	segments	BLEU
ST1	CS-UK	6.6M	31.17
ST2	UK-CS	19M	34.57
final	CS-UK	48M	35.87

Table 4: Progress of scores

5 Results

Our first submission evaluated by OCELoT system on the test data received very low scores (BLEU 15.6) which don't correlate to the scores on our development set. We noticed Russian sentences instead of Ukrainian in our translations. For the final submission, we have done more filtering of both parallel and monolingual (back-translated) data as described in Section 2. The same system on cleaned data received much better scores (BLEU 28.3)

The official scores of our final system on the test data (Kocmi et al., 2023) are listed in the Table 5.

	final	**first**
COMET	87.0	
chrF	57.0	41.0
BLEU	28.3	15.6

Table 5: Automatic Scores of the final system and the first submission.

6 Conclusion

This paper presents the MUNI-NLP submission to the WMT 2023 General Machine Translation Task. Our results show that it is very important to clean the training data, especially foreign languages.

The paper also introduces a novel tokenization into subwords, a detailed evaluation of it is part of our future work.

Acknowledgments

The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. TorchScale: Transformers at scale. *CoRR*, abs/2211.13184.

Martin Popel, Jindrich Libovicky, and Jindrich Helcl. 2022. Cuni systems for the wmt 22 czech-ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 352–357, Abu Dhabi. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Edoardo Signoroni and Pavel Rychlý. 2022. HFT: High frequency tokens for low-resource NMT. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 56–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. DeepNet: Scaling Transformers to 1,000 layers. *CoRR*, abs/2203.00555.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang, and Ying Qin. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410, Abu Dhabi. Association for Computational Linguistics.