CUNI Systems for the WMT 22 Czech-Ukrainian Translation Task

Martin Popel* **Jindřich Libovický*** **Jindřich Helcl***
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{popel,libovicky,helcl}@ufal.mff.cuni.cz

Abstract

We present Charles University submissions to the WMT 22 General Translation Shared Task on Czech-Ukrainian and Ukrainian-Czech machine translation. We present two constrained submissions based on block back-translation and tagged back-translation and experiment with rule-based romanization of Ukrainian. Our results show that the romanization only has a minor effect on the translation quality. Further, we describe Charles Translator, a system that was developed in March 2022 as a response to the migration from Ukraine to the Czech Republic. Compared to our constrained systems, it did not use the romanization and used some proprietary data sources.

1 Introduction

How fast can the machine translation (MT) community react to a sudden need of a high-quality MT system which was previously under low demand? This question motivated the new task at the WMT this year, which is Czech-Ukrainian translation.

Both languages belong to the Slavic language family (Czech is western Slavic, Ukrainian is eastern Slavic), and share some lexical and structural characteristics. Unlike Czech, which uses the Latin script, Ukrainian uses its variant of the Cyrillic alphabet.

We submit three systems to the WMT 22 General Translation Shared Task for this language pair in each translation direction. The first system, CUNI-JL-JH, implemented in Marian (Junczys-Dowmunt et al., 2018), uses tagged back-translation and is a result of our experiments with romanization of Ukrainian. Our second system, CUNI-TRANSFORMER, implemented in Tensor2Tensor (Vaswani et al., 2018), uses block back-translation. Finally, we submit an unconstrained system, CHARLES TRANSLATOR, implemented in Tensor2Tensor, which has been developed in spring 2022 as a response to the crisis caused by the Russian invasion of Ukraine and the following migration wave.

2 Constrained WMT Submissions

We submitted two systems in each translation direction that use the same parallel and monolingual data, but different techniques and different toolkits. This section first describes the shared data processing steps and then the specifics of each of the submissions in separate subsections.

2.1 Training Data

We use all parallel data allowed in the constrained task, along with 50 million Czech and 58 million Ukrainian sentences of monolingual data. In the following paragraphs we describe the data cleaning steps when preparing the training data. We further experiment with romanization of the Ukrainian Cyrillic alphabet and with artificial noising of the data.

Parallel data. The data for the constrained translation task consist of OPUS corpora (Tiedemann, 2012) that have a Czech-Ukrainian part, WikiMatrix (Schwenk et al., 2021) and the ELRC EU acts in Ukrainian.[1]

We clean the parallel data using rule-based filtering in the following way:

1. Filter out non-printable and malformed UTF-8 characters.
2. Detect language using FastText (Grave et al., 2018), only keep Czech and Ukrainian sentences on their respective source/target sides.
3. Only keep sentence pairs with character length ratio between 0.67 and 1.5 if longer than 10 characters.

*The author order was determined by a coin toss.

[1] https://elrc-share.eu/repository/search/?q=EU+acts+in+Ukrainian

Source	Original	Filtered
bible-uedin	8 k	8 k
CCMatrix	3,992 k	3,884 k
EUbookshop	2 k	1 k
GNOME	150	81
KDE4	134 k	64 k
MultiCCAligned	1,607 k	1,199 k
MultiParaCrawl	1,773 k	1,606 k
OpenSubtitles	731 k	273 k
QED	161 k	138 k
Tatoeba	3 k	2 k
TED2020	115 k	106 k
Ubuntu	0.2k	0.2k
wikimedia	2 k	2 k
XLEnt	695 k	695 k
WikiMatrix	105 k	99 k
ELRC EU Acts	130 k	108 k
Total	9,457 k	8,186 k

Table 1: Sizes of parallel data sources (number of sentence pairs).

4. Apply hand-crafted regular expressions to filter out the frequent errors, such that the system does not attempt to translate e-mail addresses, currencies, etc. In addition, regular expressions check translations of names of Czech[2] and Ukrainian[3] municipalities downloaded from Wikipedia.

We omit steps 2 and 3 for the XLEnt corpus, which seems to be very clean and consist of short phrases (likely to get misclassified for language).

The sizes of the used parallel data sources before and after cleaning are presented in Table 1.

Monolingual data. The overview of the monolingual data sources is in Table 2. For Czech, we use the Czech monolingual portion of the CzEng 2.0 corpus (Kocmi et al., 2020). For Ukrainian, we used all resources, available for WMT, i.e., the NewsCrawl, the Leipzig Corpora (Biemann et al., 2007), UberText corpus (Khaburska and Tytyk, 2019) and Legal Ukrainian Crawling by ELRC. The Uber corpus and the Ukrainian Legal corpus are distributed tokenized with removed punctuation. We automatically restored the punctuation and detokenized the models using a lightweight Transformer model (Vaswani et al., 2017; Base model with 3 layers, 8k vocabulary) trained on the NewsCrawl corpus.

For Ukrainian, we only keep sentences shorter than 300 characters. For Czech, we keep all sentence lengths from the CzEng corpus (up to 1400 characters). For both languages, we remove non-printable and malformed UTF-8 characters.

[2] https://uk.wikipedia.org/wiki/Міста_Чехії

[3] https://cs.wikipedia.org/wiki/Seznam_měst_na_Ukrajině

Source		Original	Filtered
Czech	CzEng 2.0		50.6 M
Ukrainian	NewsCrawl	2.3 M	2.0 M
	Leipzig Corpora	9.0 M	7.6 M
	UberText Corpus	47.9 M	41.2 M
	ELRC Legal	7.6 M	7.2 M
	Total	66.8 M	58.1 M

Table 2: Monolingual data sizes in number of sentences before and after filtering.

Romanization. We develop a reversible romanization than transcribes between the Ukrainian and Czech alphabets. For example, Зараз у нас є 4-місячні миші is transcribed to *Zaraz u nas je 4-misjačni myši*. This way the model can better exploit the lexical similarities between the two languages (e.g. миші should be translated to Czech as *myši*), while keeping all the necessary information to reconstruct the original Cyrillic text. Note that the transcription of Cyrillic changes when changing the target language, reflecting the phonology of that language (e.g. ш transcribes to *sh* in English, but *š* in Czech). We introduce special tags for words and characters that are written in Latin script found in Cyrillic text. The romanization is specifically designed for Ukrainian (e.g. и transcribes to *y*, not *i* as would be the case in Russian), so its reversibility occasionally fails for Russian names.

Artificial noise. We apply synthetic noise on the source side that should simulate the most frequent deviations from the standard orthography (missing capitalization, lower- or upper-casing parts of the sentences, missing or additional punctuation).

All scripts for training data processing are available at `https://github.com/ufal/uk-cs-data-scripts`. We use Flores 101 (Goyal et al., 2022) development set for validation.

2.2 Tagged-back-translation-based System (CUNI-JL-JH)

The CUNI-JL-JH submission is a constrained system and uses the data described in the paragraphs above. We train the system in 3 iterations of tagged back-translation (Caswell et al., 2019) with greedy decoding. Each iteration, we filter the back-translated data using Dual Cross-Entropy filtering (Junczys-Dowmunt, 2018) when keeping

40,930,735 synthetic sentences, i.e., 5× the size of clean authentic parallel data.

The first two back-translation iterations were done with the Cyrillic script on the Ukrainian side. In the final back-translation iteration, we performed romanization and noising of the source side. We train three models with random initialization and submit the ensemble.

For all iterations, we used a Transformer Big model with tied embeddings and a shared SentencePiece vocabulary size of 32k (fitted on 5M randomly sampled sentences; with sampling at the training time, α=0.1; Kudo and Richardson, 2018). We set the learning rate to 0.0003 and use 8,000 warm-up steps. We initialize the models randomly in each back-translation iteration.

For validation, we use greedy decoding. At test time, we decode with beam search with beam size of 4 and length normalization of 1.0 (estimated on validation data).

The system is implemented using Marian (Junczys-Dowmunt et al., 2018).

Negative results. We experimented with Dual-Cross-Entropy filtering (Junczys-Dowmunt, 2018) for parallel data selection and came to inconclusive results. Therefore, we used all parallel data after rule-based filtering.[4]

Additionally, we experimented with MASS-style (Song et al., 2019) pre-training using monolingual data only and continue with training on parallel data. We were not able to find a hyper-parameter setting where the pre-trained model would outperform the models trained from random initialization. Therefore, we only use model trained from random initialization.

2.3 Block back-translation System (CUNI-TRANSFORMER)

The CUNI-Transformer submission is also constrained, trained on the same data as CUNI-JL-JH. The system was trained in the same way as the sentence-level English-Czech CUNI-Transformer systems submitted to previous years of WMT shared tasks (Popel, 2018, 2020; Gebauer et al., 2021). It uses Block back-translation (BlockBT) (Popel et al., 2020), where blocks of authentic (human-translated parallel) and synthetic (back-translated) training data are not shuffled together, but checkpoint averaging is used to find the optimal ratio of checkpoints from the authentic and synthetic blocks (usually 5:3). The uk→cs system was trained with a non-iterated BlockBT (i.e. cs-mono data was translated with an authentic-only trained baseline). The cs→uk was trained with two iterations of BlockBT (i.e. the uk-mono data was translated with the above mentioned uk→cs non-iterated BlockBT system). We had not enough time to train more iterations and apply noised training and romanization. The system was implemented using Tensor2Tensor (Vaswani et al., 2018).

Inline casing. We experimented with Inline casing (InCa) pre-processing in the cs→uk direction. The main idea is to lowercase all training data and insert special tags `<titlecase>` and `<all-uppercase>` before words in the respective case, so that the original casing can be reconstructed (with the exception of words like *McDonald* or *iPhone*, which use different casing patterns than all-lowercase, all-uppercase and titlecase). We improved this approach by remembering the most frequent casing variant of each (lowercased) word in the training data. The most frequent variant does not need to be prefixed with any tag, which makes the length of training sequences shorter. We also introduced a third tag `<all-lowercase>` for encoding all-lowercased words whose most frequent variant is different. For example, if the InCa vocabulary includes only two items: *iPhone* and *GB*, sentence *My iPhone 64GB and iPod 64 GB or 32 gb* will be encoded as `<titlecase>` *my iphone* `<all-uppercase>` *64gb and iPod 64 gb or 32* *<all-lowercase> gb*. Note that *iPod* was kept in the original case because it was not included in the InCa vocabulary and it does not match any of the three "regular" casing patterns. We applied InCa on both the source and target side and experimented with training the InCa vocabulary on the authentic data only or on authentic plus synthetic (monolingual backtranslated).

Inline casing showed promising results in preliminary experiments (without backtranslation), especially when combined with romanization and artificial noise in training. Unfortunately, we had not enough time to train the backtranslated model long enough, so we submitted it only as a contrastive run and plan to explore it more in future.

[4]Note that we use Dual-Cross-Entropy for filtering the monolingual data, as described in the first paragraph of this section, but we have not done any experiments with keeping all the monolingual data.

Model	cs→uk	uk→cs
Authentic only	20.91	22.95
BT iteration 1	21.69	23.70
BT iteration 2	21.87	23.98
BT iteration 3 (seed 1)	21.53	23.76

Table 3: Validation BLEU scores for the first two iterations of BT for the tagged BT systems.

3 Charles Translator for Ukraine

Charles Translator for Ukraine is a free Czech-Ukrainian online translation service available for public at `https://translator.cuni.cz` and as an Android app. It was developed at Charles University in March 2022 to help refugees from Ukraine by narrowing the communication gap between them and other people in Czechia. Similarly to CUNI-TRANSFORMER, it is based on Transformer and iterated Block back-translation (Popel et al., 2020). The training used source-side artificial noising, but no romanization and no inline casing. It was trained on most (but not all) of the training data provided by WMT plus about one million uk-cs sentences from the InterCorp v14 corpus (Čermák and Rosen, 2012; Kotsyba, 2022), so this submission is unconstrained.

4 Results

In this section, we report BLEU scores on the Flores 101 development set that we used to make our decisions about the system development and the final automatic scores. Note that the validation set is very different from the test set. The validation set consists of clean and rather complicated sentences from Wikipedia articles, whereas the WMT 22 test set is noisy user-generated text from the logs of the production deployment of Charles Translator.[5]

Tagged BT systems. Table 3 shows validation BLEU scores from the first three iterations of back-translation. The second and third iteration did not bring substantial improvements, so we decided not to further iterate.

Table 4 shows validation BLEU scores from the last (third) BT iteration – three independently trained systems and their ensembles, and the Cyrillic and romanized versions of the data. In general, ensembling only brings a small improvement. Romanization does not bring a significant difference

[5]The test set only contains sentences from users who provided their consent for this usage and the sentences were pseudonymized.

Model		cs→uk	uk→cs
Cyrillic	Seed 1	21.53	23.76
	Seed 2	22.28	**25.10**
	Seed 3	21.96	24.39
	Ensemble	22.45	24.86
Romanized	Seed 1	21.42	23.99
	Seed 2	21.76	23.91
	Seed 3	22.37	24.18
	Ensemble	**22.62**	24.22

Table 4: Validation BLEU scores for the last (i.e., the third) iteration of BT comparing romanized and original script.

compared to using the Cyrillic script. In the Czech-to-Ukrainian direction, the best system was the ensemble of the romanized systems. However, in the Ukrainian-to-Czech direction, the best system was one of the Cyrillic systems that used accidentally 3 times higher batch size than the remaining ones. This result suggests that the batch size has a much stronger effect than most of the techniques that we experimented with and that we might have reached better results if we opted for higher batch size.

Results on WMT test. Automatic evaluation on the WMT22 test set is presented in Table 5. Both the constrained systems and Charles Translator show comparable results. The tagged BT system reaches a slightly higher COMET score than the Block BT system, however, Czech-Ukrainian was not in the training data of the COMET score, which make the score unreliable for this particular language pair. For Czech-to-Ukrainian, Charles Translator reaches a slightly higher COMET score and slightly lower BLEU and chrF scores than both the constrained systems, but we do not consider such small differences of automatic metrics relevant.

5 Conclusions

We presented Charles University submissions to the WMT 22 General Translation Shared Task on Czech-Ukrainian and Ukrainian-Czech machine translation. We present two constrained submissions based on block back-translation and tagged back-translation and experiment with rule-based romanization of Ukrainian. Further, we describe Charles Translator, a system that was developed in March 2022 as a response to the migration from Ukraine to the Czech Republic. Compared to our constrained systems, it did not use the romanization

System	cs→uk			uk→cs		
	BLEU	chrF	COMET	BLEU	chrF	COMET
Best constrained (HuaweiTSC/AMU)	36.0	62.6	0.994	37.0	60.7	1.048
CUNI-Transformer	35.0	61.6	0.873	35.8	59.0	0.885
CUNI-JL-JL	34.8	61.6	0.900	35.1	58.7	0.890
Best unconstrained (Lan-Bridge/Online-B)	38.1	64.0	0.942	36.5	60.4	0.965
Charles Translator	34.3	61.5	0.908	35.9	59.0	0.901

Table 5: Final automatic results on the WTM22 test data compared to the best overall score achieved in each metric.

and used some proprietary data sources.

Our results show that the romanization only has a minor effect on the translation quality, compared to machine-learning aspects that affect translation quality. Block back-translation seems to deliver slightly better results that tagged back-translation, however the differences are only small.

Acknowledgements

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ, by the Czech Science Foundation (GACR) grant 20-16819X (LUSyD), and by the European Commission via its Horizon 2020 research and innovation programme no. 870930 (WELCOME), Horizon Europe Innovation programme no. 101070350 (HPLT).

References

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, 2007.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Petr Gebauer, Ondřej Bojar, Vojtěch Švandelík, and Martin Popel. 2021. CUNI systems in WMT21: Revisiting backtranslation techniques for English-Czech NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 123–129, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Anastasiia Khaburska and Igor Tytyk. 2019. Toward language modeling for the ukrainian. *Advances in Data Mining, Machine Learning, and Computer Vision. Proceedings*, pages 71–80.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing CzEng 2.0 parallel corpus with over 2 gigawords. *CoRR*, abs/2007.03006.

Natalia Kotsyba. 2022. Ukrainian-Czech part of InterCorp v14. `https://intercorp.korpus.cz`.

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. Association for Computational Linguistics.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel. 2020. CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

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. European Language Resources Association (ELRA).

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 6000–6010, Long Beach, CA, USA. Curran Associates, Inc.