CUNI Submission in WMT22 General Task

Josef Jon
Charles University
surname@mail.ufal.mff.cuni.cz

Abstract

We present CUNI-Bergamot submission for WMT22 General translation task. We compete in English → Czech direction. Our submission further explores block backtranslation techniques. In addition to the previous work, we measure performance in terms of COMET score and named entities translation accuracy. We evaluate performance of MBR decoding compared to traditional mixed backtranslation training and we show possible synergy when using both of the techniques simultaneously. The results show that both approaches are effective means of improving translation quality and they yield even better results when combined.

1 Introduction

This work focuses on exploring of two methods used in NMT in order to improve translation quality: backtranslation and Minimum Bayes Risk decoding using neural-based evaluation metric as a utility function. The methods used and related work are presented in the following section. In next section we describe our experimental setting and results.

2 Methods

We describe methods we used to build our system in this section.

2.1 Block backtranslation

The translation quality of NMT depends heavily on the amount of parallel training data. It has been shown that the authentic bilingual data can be partially supplemented by synthetically parallel, machine translated monolingual text (Bojar and Tamchyna, 2011; Sennrich et al., 2016; Xie et al., 2018; Edunov et al., 2018). Often the synthetic and authentic parallel data are mixed in the training dataset, but previous research shows that simply mixing the two types of text does not yield optimal translation quality. We are using block backtranslation (*block-BT*) in similar configuration to Popel et al. (2020). This method creates blocks of parallel and synthetic data and presents them to the neural network separately, switching between the two types during the training. Since in last year's WMT, the submission using block-BT by Gebauer et al. (2021) did not find any improvements, presumably due to improperly chosen block size, we decided to verify effectiveness of this method once again.

Averaging type Previous work on *block-BT* shows the importance of averaging the checkpoints to combine information from different blocks of training data in order to obtain good performance. We compare checkpoint averaging with another method of combining older sets of model's parameters with the current one – *exponential smoothing*. After each update u, the current parameters Θ_u are averaged (with smoothing factor α) with parameters after the previous update Θ_{u-1}:

$$\Theta_u = \alpha\Theta_u + (1-\alpha)\Theta_{u-1}$$

Previous work by Popel (2018) contains experiments with exponential averaging, but only on the level of already saved checkpoints, not online during the training after each update as for our work.

Minimum Bayes Risk Decoding NMT models predict conditional probability distribution over translation hypotheses given a source sentence. To select the most probable translation under the model (mode of the model's distribution), an approximation of MAP (*maximum-a-posteriori*) decoding is used, most commonly the beam search (Graves, 2012). However, beam search and MAP decoding in general has many shortcomings described in recent work (Stahlberg and Byrne, 2019; Meister et al., 2020) and other approaches have

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 280 - 289
December 7-8, 2022

been proposed to generate a high-quality hypothesis from the model.

One of them, MBR (Minimum Bayes Risk) decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004), has been proposed as an alternative to MAP. MBR does not produce a translation with the highest probability, rather a translation with the best value of utility function. This utility function is usually an automatic machine translation evaluation metric. However, to optimize towards best utility function value, it would necessary to know the ideal selection of hypothesis. In case of MT, that would mean a perfect, best possible translation, which of course is not known during the translation process. For this reason, an approximation of the ideal translation is used, based on the model's probability distribution (Bryan and Wilker, 2021). This can be implemented as generating a list of hypotheses (e.g. using sampling or beam search) and then computing utility function of each hypothesis using all the other hypotheses as the ideal translation approximation (i.e. as references). This approximation of MBR decoding can be seen as consensus decoding – the hypothesis that is the most similar to all the others is chosen.

Even though MBR is able to optimize towards many metrics and increase the scores, these gains did not translate into better human evaluation of the final translations, when using traditional metrics based on surface similarities like BLEU. Recent successes in development of novel metrics for machine translation has renewed interest in this method. (Amrhein and Sennrich, 2022a; Freitag et al., 2021; Müller and Sennrich, 2021).

3 Experiments

In this section we present our experimental setup and results.

3.1 Tools

We tokenize the text into subwords using FactoredSegmenter[1] and SentencePiece (Kudo and Richardson, 2018). We use MarianNMT (Junczys-Dowmunt et al., 2018) to train the models. BLEU scores are computed using SacreBLEU (Post, 2018), for COMET scores (Rei et al., 2020) we use the original implementation[2].

[1] https://github.com/microsoft/factored-segmenter

[2] https://github.com/Unbabel/COMET

3.2 Datasets

We train English-Czech NMT models for our experiments. We train our models on CzEng 2.0 (Kocmi et al., 2020). We use all 3 subsets of CzEng corpus: the originally parallel part, which we call *auth*, Czech monolingual data translated into English using MT (*csmono*) and English monolingual data translated into Czech using MT (*enmono*). We use `newstest2020` (Barrault et al., 2020) as our dev set and `newstest2021` (Akhbardeh et al., 2021) as our test set.

For experiments concerning translation of named entities, we used a test set originally designed for Czech NLG in restaurant industry domain[3](Dušek and Jurčíček, 2019). It contains sentences which include names of restaurants and addresses in Czech and their translations in English. We will call this test set the `restaurant` test set.

3.3 Models

We train Transformer-base (which we denote *base*) and Transformer-big (*big 6-6*) models with standard parameters (Vaswani et al., 2017) as preconfigured in MarianNMT. For the largest model (*big 12-6*), we use Transformer-big with 12 encoder layers and depth scaled initialization (Junczys-Dowmunt, 2019; Zhang et al., 2019)[4]. We also used learning rate of $1e{-}4$ for the 12 layer model instead of $3e{-}4$, which was used for other models. We trained all models for at least 1.4M updates. After that, we computed validation BLEU scores every 5k updates and we stopped if the score did not improve for 30 consecutive validations. We trained the models on heterogenous grid server, which includes combinations of Quadro RTX 5000, GeForce GTX 1080 Ti, RTX A4000 and GeForce RTX 3090 cards. Typical training time on 4 108Ti of the base models for 1.4M updates was 7 days.

3.4 Block-BT settings

For all our experiments, we create a checkpoint each 5k updates and we vary only the size of the blocks during which the training data have the same type (20k, 40k, 80k and 160k updates). The size is the same for all block types. We circle through the block types in the following order: *auth*→*csmono*→*auth*→*enmono*.

[3] https://github.com/UFAL-DSG/cs_restaurant_dataset

[4] Training scripts available at: https://github.com/cepin19/wmt22_general

For checkpoint averaging, we average 8 checkpoints. For exponential smoothing, we use default Marian configuration ($\alpha = 0.001$, but there are some slight modifications based on number of updates since start of the training and batch size).

We also look at the effects of using only backtranslation, or both back- and forward-translation.

3.5 Block-BT results

Training regime and averaging method First, we compare different training regimes: *mixed-BT*, where all the training datasets are concatenated and shuffled together and *block-BT* with 40k updates long blocks and two possible averaging types – exponential smoothing (*exp*) or checkpoint averaging (*avg8*).

Figure 1 shows behavior of BLEU and COMET scores on `newstest2020` during the training for these configurations. We opt to present the interval between 480k and 1280k updates. We chose the lower bound because the behavior is more stabilized than in the beginning of the training and the upper bound because all the models were trained for at least 1400k updates and 1280k is the nearest lower multiplicative for the largest block size. *40k block* curve represents a model without any averaging, *40k block avg8* is a model trained without exponential smoothing, but each checkpoint was averaged with 7 previous checkpoints for the evaluation, *40k block exp* model was trained with continuous exponential smoothing. Finally, we also experimented with combination of both - trained with exponential smoothing and averaged after the training. The combination does not improve over the separate averaging techniques and we omitted the curve from the figure to make it more readable.

In both metrics, *block-BT* with either form of averaging outperforms *mixed-BT* training. Without any averaging, the advantage of *block-BT* over *mixed-BT* is smaller. Type of averaging does not seem to play a large role – checkpoint averaging, exponential smoothing and their combination yield very similar best scores. The best scores on `newstest2020` for each combination of parameters are presented in Table 1.

The curves for checkpoint averaging and exponential smoothing behave similarly, with exponential averaging reacting faster to change of the block. Additionally, the *avg8* models have higher peaks in *enmono* (red) blocks, especially for BLEU scores. The shape of the curves could be tuned by changing frequency of saving checkpoints and number of checkpoints to be averaged for checkpoint averaging method, or by changing the α factor for exponential smoothing.

There are differences in behaviour between BLEU and COMET score curves. Most notably, COMET is less sensitive to transition from *auth* (green) to *csmono* (blue) blocks. We hypothesize this is caused by lower sensitivity of COMET score to wrong translation of named entities and rare words (Amrhein and Sennrich, 2022a). We present further experiments in this direction later.

Block size We asses influence of block size for both of the two averaging methods. We compare block sizes of 20k, 40k, 80k and 160k updates. Behaviour of COMET and BLEU scores is presented in Figures 2 and 3 for exponential smoothing and checkpoint averaging, respectively. The best scores are again shown in Table 1.

We see that 20k block size yields noticeably worse results when using checkpoint averaging that the other sizes. The negative effect of the small block size is less pronounced when using exponential smoothing, yet still present. Other block sizes perform similarly in both metrics. This results is expected, since for 8-checkpoint averaging with 5k updates checkpointing interval, it is necessary to have a block size of at least 40k updates to fit all the 8 checkpoints and thus explore all possible ratios of *auth* and *mono* data.

Reverse direction For the reverse direction, Czech to English, we performed less extensive evaluation. We only compare *mixed*, *block-BT* with 40k blocks and either exponential smoothing or checkpoint averaging. Behavior of the metrics is shown in Figure 4 and final best scores on `newstest2020` are presented in Table 2. *Block-BT* still outperforms *mixed* training, but by a smaller margin than in the other direction.

Backtranslation direction We also evaluate influence of using only backtranslations as additional synthetic data (monolingual data in target language to automatically translated to source language) or adding also forward translations (from source language to target target) and we present the results in Table 3. Interestingly the results show large gains in both BLEU and COMET when using forward translation. We hypothesize this is caused by the good quality of the model used to perform the forward translation. In such case, the translation

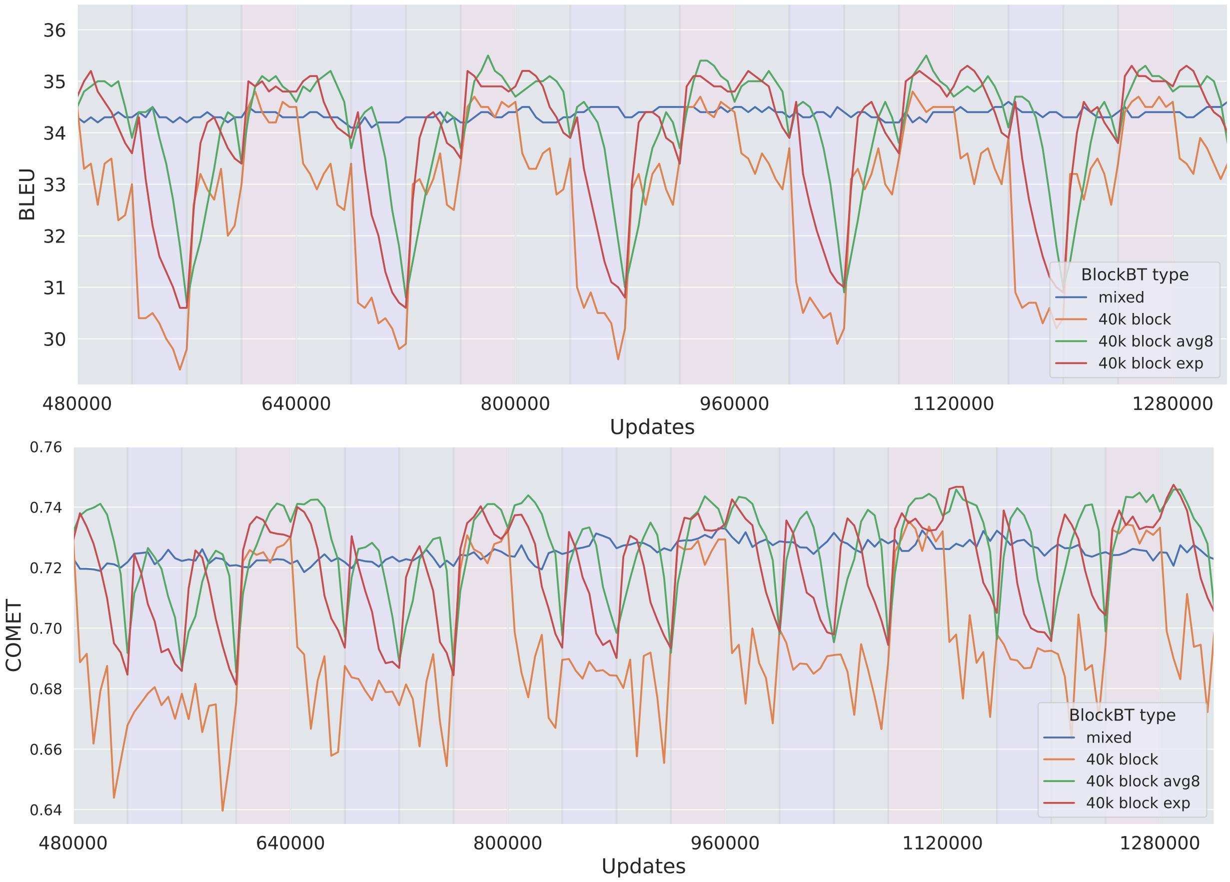

Figure 1: Comparison of different training regimes for EN-CS translation on newstest20 in terms of BLEU (top) and COMET (bottom). Background colors for block-BT regime show which part of training data was used for given part of the training. Green means authentic parallel data, blue is CS->EN backtranslation and red is EN->CS forward translation.

model assumes the role of the teacher in teacher->student training and might lead to a good quality results.

Named entities test sets From anecdotal evidence, we have seen that checkpoints with large influence of backtranslated data perform worse on named entities translation and COMET and BLEU scores might not reflect this drop of accuracy. We evaluate the models in terms of accuraccy of named entitiy translation on the `restaurant` test set. We selected Czech to English direction, since the evaluation is easier given lower morphological richness of target language. Figure 5 shows comparison of behavior of named entities translation accuracy on the restaurant test set and COMET and BLEU scores on `newstest2020` for exponential smoothing and checkpoint averaging. NE accuracy peaks towards the end of *auth* regions (green). Both COMET and BLEU scores peak also during the *auth* part of the training, but, especially for COMET, the peak occurs in earlier stages after the switch to *auth*. Overall, BLEU curve correlates better with the NE accuracy curve. We hypothesize this might be related to the fact that COMET was found to be insensitive to named entities errors by Amrhein and Sennrich (2022b).

However, it seems that the shift between the accuracy and the other two metrics is not too large in our settings and choosing the best performing model in terms of either COMET or BLEU should not hurt NE translation by a large amount. We further investigate that in Table 4 – we chose the checkpoint with best COMET (first row) and best BLEU (second row) on the `newstest2020` and the checkpoint with best NE translation accuracy on the restaurant test set (third row). We compute all three metrics for these three models. The best COMET checkpoint obtains accuracy of 60.7% on the restaurant test set, the best BLEU checkpoint reaches accuracy of 62.9%, while the best accuracy reached by any checkpoint is 63.6%.

Model size	Block size	Avg type	update (k)	BLEU	update (k)	COMET
base	mixed	exp	1340	34.7	1760	0.7337
	mixed	exp+avg8	1365	34.7	965	0.7326
	20k	-	1360	34.6	640	0.7324
		exp	410	34.9	725	0.7406
		avg8	660	34.8	1385	0.7349
		exp+avg8	420	34.9	735	0.7399
	40k	-	610	34.8	1415	0.7363
		exp	1130	35.3	1290	**0.7474**
		avg8	780	35.5	1420	0.7462
		exp+avg8	1150	35.5	1075	0.7466
	80k	-	1250	34.9	960	0.7393
		exp	1210	35.2	1450	0.7447
		avg8	985	35.5	665	**0.7474**
		exp+avg8	585	35.3	1150	0.7455
	160k	-	1130	34.9	1210	0.7387
		exp	1125	35.3	1285	0.7453
		avg8	1135	35.5	1305	0.7467
		exp+avg8	1145	35.3	1310	**0.7473**
big 6-6	40k	exp	445	35.4	1125	0.7546
		exp+avg8	300	35.4	1310	0.7567
big 12-6	40k	exp	130	36.1	1210	0.7848

Table 1: Best COMET and BLEU scores on EN-CS newstest2020 for all the combinations of models size, training regime and block size. We report the best score and an number of updates after which was this score reached.

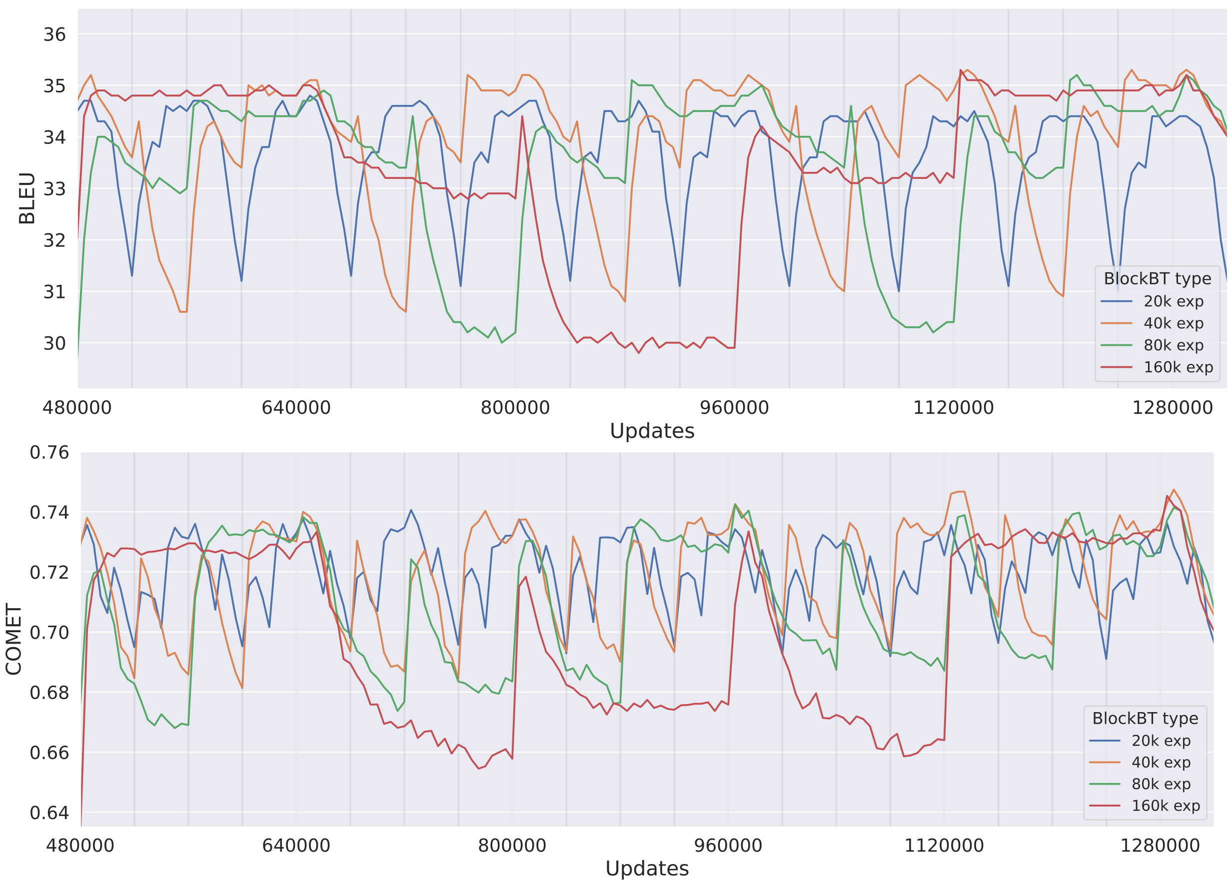

Figure 2: Comparison of how the block size affects behavior of BLEU (top) and COMET (bottom) scores during the training for block-BT with exponential smoothing of the parameters, without checkpoint averaging, on EN-CS `newstest2020`.

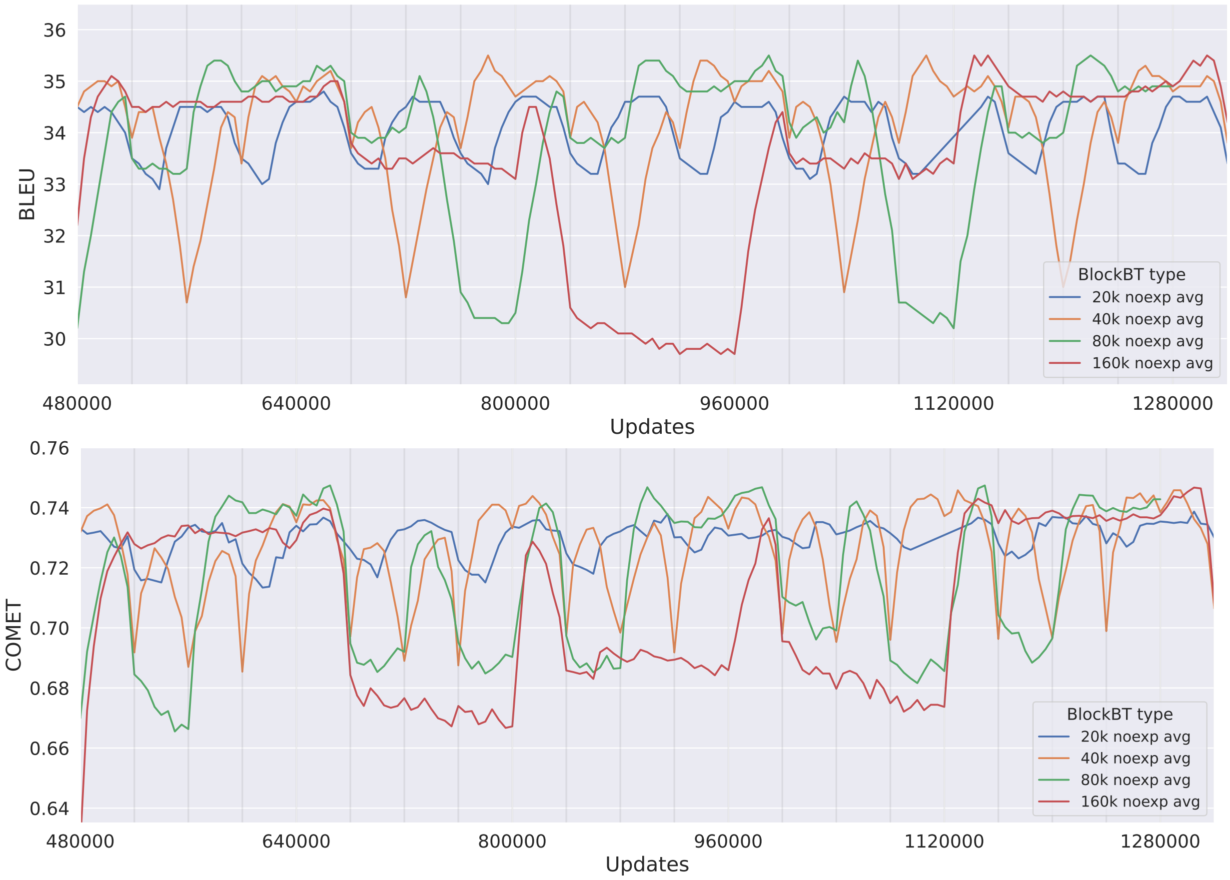

Figure 3: Comparison of how the block size affects behavior of BLEU (top) and COMET (bottom) scores during the training or block-BT with checkpoint averaging and no exponential smoothing of the parameters, on EN-CS `newstest2020`.

Model	Block	Avg type	update (k)	best BLEU	update (k)	best COMET
base	mixed	exp	1405	25.2	1220	0.4149
		exp+avg8	1430	25.1	1220	0.4114
	40k	-	580	25.3	1040	0.4086
		exp	755	25.3	570	0.4183
		avg8	765	25.4	1060	0.4175
		exp+avg8	1080	25.2	1230	0.4186

Table 2: COMET and BLEU scores for Czech to English directions. The best checkpoints were chosen based on their performance on `newstest2020`.

Update (k)	COMET	BLEU	Acc
570	**0.4183**	24.9	0.607
755	0.4038	**25.3**	0.629
590	0.4099	24.9	**0.636**

Table 4: Best checkpoints of Czech to English model trained with 40k blocks and exponential smoothing in terms of COMET (first row), BLEU (second row) on newstest2020 and NE translation accuracy on restaurant test set (third row).

dir	regime	datasets	D BLU	T BLU	D CMT	T CMT
encs	mixed	all	34.7	20.9	0.7337	0.6206
		auth+cs	31.5	19.5	0.6904	0.5779
		auth+en	34.8	20.6	0.7258	0.6097
	block	all	35.3	**21.1**	0.7474	**0.6245**
		auth+cs	33.9	19.9	0.7232	0.5908
		auth+en	**35.4**	20.7	**0.7497**	0.6147
csen	mixed	all	25.2	-	0.4149	-
	block	all	25.3	-	0.4183	-
		auth+en	24.3	-	0.3682	-

Table 3: Results on newstest2020 and newstest2021 for various dataset combinations. *D/T* mean dev (*newstest2020*) and test (*newstest2021*) sets respectively, *CMT* stands for wmt20-comet-da scores.

3.6 MBR decoding

We used MBR decoding to rerank concatenation of n-best lists produced by various checkpoints. In total, we used 6-best lists from 12 checkpoints. We divided the checkpoints based on which block of the training data they were saved in and sorted them by COMET score on `newstest2020`. Using different strategies we selected the best performing checkpoints to provide the n-best lists. We present the results in Table 5. The first row shows results for mixed-BT regime, i.e. we concatenated n-best lists produced by the 12 best performing mixed-BT

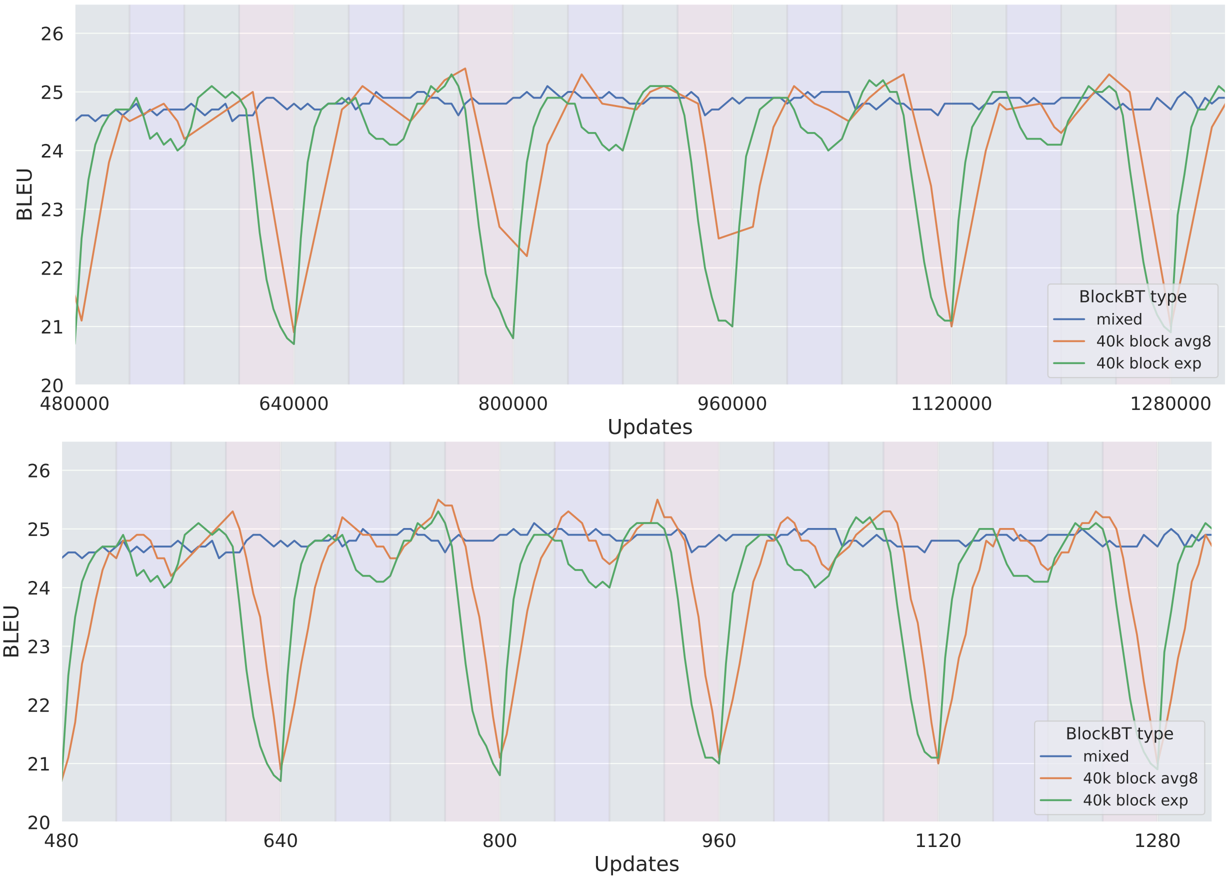

Figure 4: Comparison of different training regimes for CS-EN translation on `newstest2020` in terms of BLEU (top) and COMET (bottom). Background colors for block-BT regime show which part of training data was used for given part of the training. Green means authentic parallel data, blue is CS->EN forward translation and red is EN->CS backtranslation.

i	auth	cs	en	AVG comet20	MBR comet20	comet21
1	-	-	-	0.7322	0.7888	0.0885
2	9	2	1	**0.743**	0.8082	0.0946
3	4	4	4	0.7408	0.8182	0.0972
4	12	0	0	0.7425	0.801	0.0929
5	0	12	0	0.7303	0.8104	0.0949
6	0	0	12	0.7372	0.796	0.0918
7	1	7	4	0.737	**0.8232**	**0.0981**
8	0	7	5	0.7361	**0.8232**	**0.098**
9	2	7	3	0.7377	**0.8231**	**0.0981**

Table 5: Results of MBR decoding on `newstest2020` for different selection of the hypotheses n-best lists produced by checkpoints from different training blocks. In total, 12 n-best lists produced by transformer-base models are concatenated and the first three columns show how many n-best lists are used from each block (the checkpoints for each block are sorted by COMET (wmt20-da model), so these are produced by the best performing checkpoints). The *AVG COMET20* shows the average wmt20-da COMET scores for the first hypotheses of each n-best list that was used, *MBR COMET20* shows wmt20-da score of the final sentences after MBR decoding, COMET21 shows results of the same sentences from wmt21-da model.

checkpoints. In the second row, the block-BT training checkpoints were used to create n-best lists, selected only based on their COMET scores, without any regard on the block type they were saved in. In third row, we combine n-best lists from 4 best performing checkpoints from each type of block. In rows 4-6, we use best performing checkpoints from each type of block separately. In the final row, we show the optimal selection which yielded the highest score. The results suggest that larger diversity in terms of block type of the checkpoints improves MBR results: the combination of n-best lists produced by checkpoints from diverse block types provides a better pool of hypotheses for MBR, even though the average COMET score of these checkpoints is lower than for the less diverse selection. This can be observed in rows 2 and 3.

3.7 Submission

Our primary submission is based on the *big 12-6* model and MBR decoding. We explored all the possible combinations of 18 checkpoints from dif-

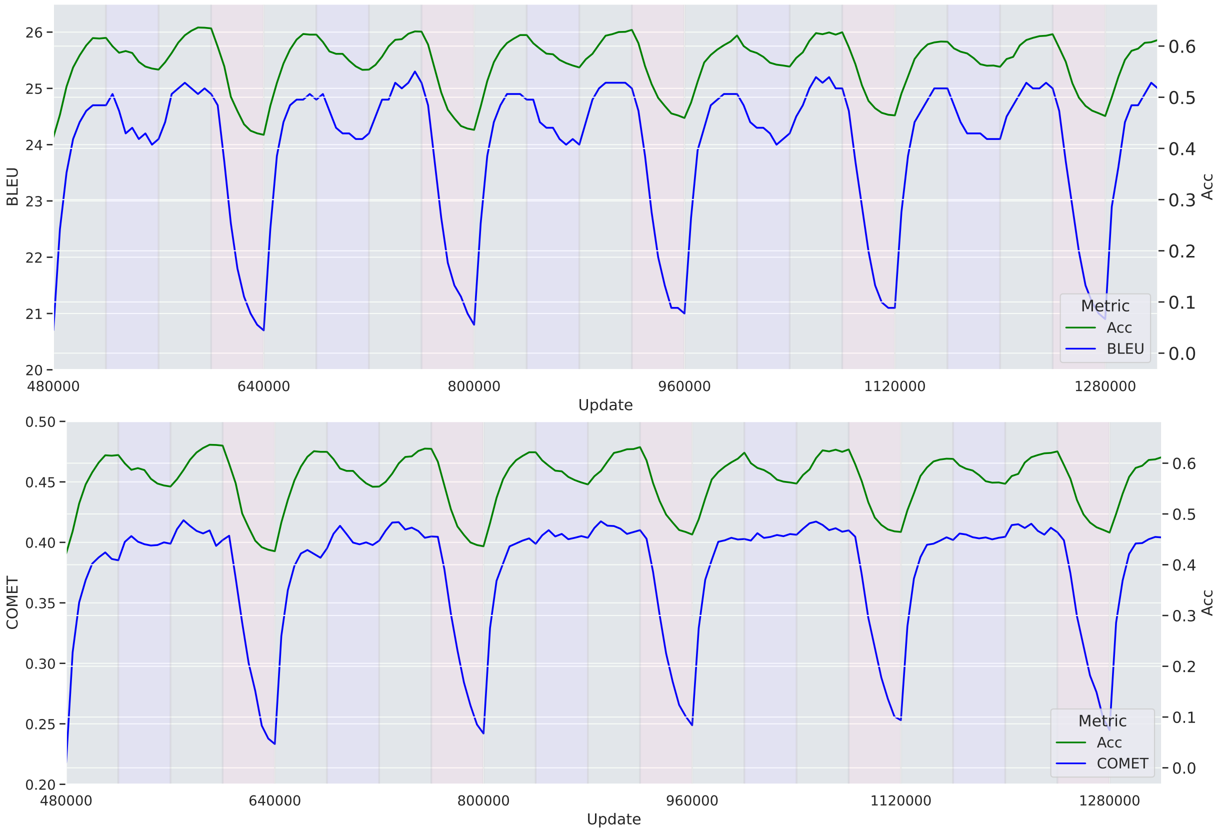

Figure 5: Behaviour of BLEU (top), COMET (bottom) on `newstest2020` and NE translation accuracy on `restaurant` test set for Czech to English translation with block-BT using exponential smoothing.

auth	cs	en	AVG comet20	MBR comet20	comet21
9	2	8	0.7802	0.8566	0.1114

Table 6: Our final submission for the EN-CS general translation task, based on outputs of the transformer-big 12-6 model. Meaning of the columns is identical to Table 5.

System	COMET-B	COMET-C	ChrF-all
Online-W	97.8	79.3	70.4
Online-B	97.5	76.6	71.3
CUNI-Bergamot *	**96.0**	**79.0**	65.1
JDExploreAcademy *	95.3	77.8	**67.2**
Lan-Bridge	94.7	73.8	70.4
Online-A	92.2	71.1	67.5
CUNI-DocTransformer *	91.7	72.2	66.0
CUNI-Transformer *	86.6	68.6	64.2
Online-Y	83.7	62.3	64.5
Online-G	82.3	61.5	64.6

Table 7: Results of automatic metrics on wmt22 general task test set. Constrained submissions are marked by an asterisk, the best scores among constrained submissions are bold. COMET-B and COMET-C are COMET scores for the two different references, ChrF is computed using both references together.

ferent blocks as described in the previous section. The results of the best combination are shown in Table 6. We present the results of the official evaluation in our task in Table 7. In total, there were 5 submitted systems (4 constrained) and 5 online services. Our submission ranked first in COMET score among the constrained systems and third in ChrF score.

4 Acknowledgements

This work was supported by GAČR EXPRO grant NEUREM3 (19-26934X, RIV: GX19-26934X) and we used services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101). We would also like to thank Martin Popel for his feedback on the paper and Ondřej Bojar for overall guidance in the field.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022a. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet.

Chantal Amrhein and Rico Sennrich. 2022b. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet. *arXiv preprint arXiv:2202.05148*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Eikema Bryan and Aziz Wilker. 2021. Sampling-based minimum bayes risk decoding for neural machine translation.

Ondřej Dušek and Filip Jurčíček. 2019. Neural generation for Czech: Data and baselines. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 563–574, Tokyo, Japan. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021. Minimum bayes risk decoding with neural metrics of translation quality.

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.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech Language*, 14(2):115–135.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. 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.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

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.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint

Conference on Natural Language Processing (Volume 1: Long Papers), pages 259–272, Online. Association for Computational Linguistics.

Martin Popel. 2018. Machine translation using syntactic analysis.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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 Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.