Chasing COMET: Leveraging Minimum Bayes Risk Decoding for Self-Improving Machine Translation

Kamil Guttmann[*1], **Mikołaj Pokrywka**[*1], **Adrian Charkiewicz**[1], **Artur Nowakowski**[1,2]
[1] Laniqo, Poznań, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
{name}.{surname}@laniqo.com

Abstract

This paper explores Minimum Bayes Risk (MBR) decoding for self-improvement in machine translation (MT), particularly for domain adaptation and low-resource languages. We implement the self-improvement process by fine-tuning the model on its MBR-decoded forward translations. By employing COMET as the MBR utility metric, we aim to achieve the reranking of translations that better aligns with human preferences. The paper explores the iterative application of this approach and the potential need for language-specific MBR utility metrics. The results demonstrate significant enhancements in translation quality for all examined language pairs, including successful application to domain-adapted models and generalisation to low-resource settings. This highlights the potential of COMET-guided MBR for efficient MT self-improvement in various scenarios.

1 Introduction

Machine translation (MT) bridges the gap between languages, fostering global communication and information exchange. However, achieving high-quality translations across diverse languages and domains remains a significant challenge, especially for low-resource languages where limited training data hinders model performance. Even in well-resourced settings, continuous improvement and adaptation to specific domains are ongoing research efforts.

This paper explores the potential of Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) as a self-improvement strategy for MT models. MBR decoding leverages the model's predictions to select the best translation from a set of candidates, potentially improving overall translation quality.

We employ COMET (Rei et al., 2020) as the utility function in MBR decoding and rerank candidate translations generated by an MT model. This approach creates a synthetic parallel dataset from monolingual data in the source language, enabling further model self-improvement.

This study examines the effectiveness of MBR decoding for self-improvement in three language pairs: English–German (high-resource), Czech–Ukrainian (low-resource), and English–Hausa (low-resource). For English–German, the focus is on the biomedical domain, incorporating additional monolingual data, while for Czech–Ukrainian, self-improvement is explored using only the training data translated by the model and reranked through MBR decoding. We further investigate the potential of iterative self-improvement with MBR decoding in both English–German and Czech–Ukrainian language pairs. Finally, in the case of English–Hausa, we compare the use of COMET, a massively multilingual metric, with a metric specifically tailored to African languages i.e. AfriCOMET (Wang et al., 2023).

To determine the optimal configuration for MBR decoding, we investigate two decoding algorithms and various numbers of translation candidates.

[*]Equal contribution

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 80–99
June 24-27, 2024 ©2024 European Association for Machine Translation

2 Related Work

MBR and QE reranking with neural metrics MBR decoding, a technique commonly used in Statistical Machine Translation (SMT), has gained traction in Neural Machine Translation (NMT) in recent years. Freitag et al. (2022) proposed using reference-based metrics, such as BLEURT (Sellam et al., 2020a) and Quality Estimation (QE) models, such as COMET-QE (Rei et al., 2021) for reranking the set of hypotheses produced by the NMT model.

Similar work by Fernandes et al. (2022) proposed *quality-aware decoding*. They explored various reranking strategies, including the well-performing pre-ranking of the set of hypotheses with QE models before passing them into MBR decoding. They found that using MERT-tuned (Och, 2003) reranker, where multiple QE metrics and model log-likelihood scores are linearly combined with learned weights to maximize a reference-based metric on a validation set shows improvements over the baseline.

Amrhein and Sennrich (2022) used MBR decoding to identify biases and weaknesses in COMET, where they found that the early COMET models are not sufficiently sensitive to discrepancies in numbers and named entities.

MBR decoding performance is heavily dependent on the number of samples and the sampling strategy. Freitag et al. (2023) investigated various sampling strategies and found that epsilon sampling outperformed others. This sampling method discards tokens with a probability below a certain threshold (epsilon), guaranteeing that each token in the final sample has a fair chance of being included. The approach is particularly effective when generating a large set of samples, as it inherently yields greater sample diversity compared to beam search.

Vernikos and Popescu-Belis (2024) introduced QE-fusion, a method that combines spans from different candidates sampled from a model using QE metrics. They found that the method consistently improves translation quality in terms of neural evaluation metrics, especially if applied to LLM due to their ability to generate diverse outputs.

Due to its ease of implementation and use, MBR and QE reranking have been successfully applied in machine translation shared tasks, as demonstrated by the results in several studies (Nowakowski et al., 2022; Kudo et al., 2023; Jon et al., 2023). This highlights its potential to significantly improve translation quality.

Model self-improvement Recent research has shown a growing interest in leveraging model outputs for self-improvement. This approach holds significant promise in the case of machine translation, especially for low-resource and domain-specific translation scenarios, where there is access to the source-language data, but the corresponding target-language data is severely limited.

Gulcehre et al. (2023) describes reinforcement self-training (*ReST*) method for language modeling. The method is based on producing a dataset for fine-tuning by sampling from the model (LLM). The samples are then scored with a QE metric. Then, offline reinforcement learning algorithms are applied using a reward-weighted loss based on the QE scores. The method can be applied to all generative learning settings, but the authors focus on its application to machine translation, showing that the method increases translation quality.

Concurrent work by Finkelstein et al. (2023) describes self-tuning NMT models on a set of hypotheses reranked using either MBR, QE, or a combination of the two methods. They also experimented with using LLM as the teacher model, finding that it outperforms using a self-teacher and fine-tuning on references.

Our research expands on recent developments in the field by investigating the use of MBR-based fine-tuning in three key areas. Firstly, we examine its applicability in domain-specific translation tasks, specifically focusing on English–German translation in the biomedical domain. Secondly, we investigate its effectiveness for low-resource translation directions, exemplified by the Czech–Ukrainian language pair. This broadens the scope beyond English-centric language pairs, thus contributing to a more comprehensive analysis of MBR performance across less-represented languages in neural evaluation metrics. Finally, we explore the use of neural QE metrics tailored for specific languages, using AfriCOMET (Wang et al., 2023) as an example.

3 Experiment Overview

3.1 Model Self-Improvement

The self-improvement process leverages MBR decoding to guide the model to select high-quality translations according to the utility function. The process consists of 3 steps:

Step 1: Sample Generation Using beam search decoding with beam size equal to `N`, generate `N` translation candidates using the base model for each source sentence. While Freitag et al. (2023) suggested that epsilon sampling might yield better results with MBR decoding, it typically requires reranking a significantly larger number of translation candidates, which becomes computationally expensive for processing large datasets. Beam search, on the other hand, allows for generating a smaller set of high-quality candidates while providing sufficient data for effective MBR decoding.

Step 2: MBR Decoding Select a single translation for each source sentence from the list of candidates through MBR decoding utilizing COMET to guide the selection towards high-quality translations. For an efficient implementation of the MBR decoding algorithm, we use the code[1] from the Marian (Junczys-Dowmunt et al., 2018) framework.

Step 3: Model Fine-tuning Fine-tune the base model on the synthetically created dataset. Use COMET as an early stopping metric during training to ensure fitting to this metric.

3.2 English–German

The English–German experiment simulates a real-world domain adaptation scenario. In such settings, while a large general-purpose parallel corpus might be available, the specific domain often lacks extensive parallel data. To address this challenge, we leveraged both a smaller parallel dataset and a larger monolingual dataset in the source language containing biomedical terminology.

To leverage the monolingual data in the source language we propose a two-step approach:

1. Fine-Tuning: We fine-tune a general-purpose English–German model on a small parallel biomedical dataset.
2. Self-improvement: To enhance the model performance in the biomedical domain, we incorporate a larger monolingual biomedical dataset during the self-improvement process. This involves creating a synthetic parallel dataset via MBR decoding and subsequently fine-tuning the biomedical translation model on the generated data.

To assess the robustness of the self-improvement method, we conducted an additional experiment in which we applied this method to a model that was fine-tuned to the biomedical domain using general domain data for MBR decoding. This evaluated whether the model would retain its translation capabilities in the biomedical domain despite improvements based solely on out-of-domain data.

3.3 Czech–Ukrainian

The Czech–Ukrainian experiment addresses the challenge of machine translation between two low-resource languages. We aimed to evaluate whether self-improvement through MBR decoding leads to an increase in the overall translation quality when applied to language pairs that do not involve English, which typically dominate machine translation research.

In this setting, we used only the parallel data set without incorporating any additional monolingual data. To employ MBR decoding in this data-scarce environment, we directly translated the entire source side of the parallel dataset using the baseline translation model. This created a set of synthetic candidate translations, which were then reranked through MBR decoding.

In contrast to our English–German experiments where we incorporated external monolingual data, this setup explored self-improvement without relying on additional datasets. We achieved this by solely leveraging the information present within the data of the base model. This demonstrates the potential for self-improvement even in resource-constrained scenarios.

3.4 English–Hausa

The English–Hausa experiment delves into the critical question of how the choice of a quality evaluation metric influences the effectiveness of self-improvement with MBR decoding. We explored the impact of language coverage in the evaluation metric by comparing two approaches:

[1] https://github.com/marian-nmt/marian-dev/tree/master/scripts/mbr

- MBR decoding with WMT22 COMET: utilizing the *wmt22-comet-da* model, which has been trained on direct assessments between a diverse set of language pairs.

- MBR decoding with AfriCOMET: using AfriCOMET-STL, a novel COMET-like metric specifically designed for evaluating translations to and from multiple African languages, including Hausa.

The objective of this study was to investigate the effect of language contribution in the neural evaluation metric on the quality of translations decoded using MBR. The comparison of these two approaches specifically addresses whether self-improvement guided by the WMT22 COMET metric, which is trained on a diverse range of language pairs, can effectively generalize to low-resource language pairs. Furthermore, we explore the potential need to use language-specific metrics, such as AfriCOMET-STL for Hausa, to achieve better performance in such scenarios.

3.5 Iterative MBR Self-Improvement

Following the initial self-improvement through MBR decoding, we explored the possibility of applying it iteratively to further enhance the model's translation quality.

We started each iteration by selecting the best model checkpoint based on the WMT22 COMET metric on the validation set. Next, we performed MBR decoding on the entire training set using this checkpoint, generating a new iteration of the synthetic training set. Finally, we resumed the training of the model using the new training set, starting from the previously selected checkpoint.

The iterative process was repeated until a decrease was observed in the evaluation scores of metrics other than WMT22 COMET. In the case of English–German biomedical translation, the process was continued until the model's quality improved solely on an in-domain test set and decreased on a general domain test set, as this could indicate potential overfitting to the biomedical domain.

4 Experimental Setup

4.1 Data Filtering

We filtered the general training data using the following heuristic filters:

- average length of words in each sentence (character-wise) ≤ 15;
- number of characters in each sentence ≤ 500;
- digits in a sentence (character-wise) $\leq 15\%$;
- number of characters in the longest word ≤ 28;
- number of words in sentence ≤ 100;
- Levenshtein distance between source and target sentences ≥ 2;
- number of characters in each sentence ≥ 5;
- probability that each sentence is in the correct language $\geq 10\%$.

To ensure that each sentence is in the correct language, we have used the fastText LID-201 language identification model (Burchell et al., 2023).

The Bicleaner-AI model (Zaragoza-Bernabeu et al., 2022) is also used to filter the English–German dataset. This tool estimates the likelihood that a sentence pair constitutes a mutual translation. A threshold of 50% is established for the Bicleaner score within this language pair. Bicleaner-AI is not utilized for other language pairs due to the unavailability of open-source models for those languages.

4.2 Vocabulary

We employed SentencePiece (Kudo and Richardson, 2018), a subword tokenization library, to train unigram tokenizers for each language pair in our experiments.

For the English–German and English–Hausa setups, we created a joint vocabulary containing 32,000 subword tokens and tied all embeddings during the training of the MT model. In contrast, for Czech–Ukrainian, due to different scripts (Latin and Cyrillic), we created separate vocabularies of 32,000 subword tokens and tied only the target and output layer embeddings.

4.3 Baseline Model Hyperparameters

For all experiments, we trained Transformer (big) (Vaswani et al., 2017) models using the Marian framework. These models were trained on four NVIDIA A100 GPUs, each equipped with 80GB of VRAM.

Hyperparameter Settings:

- learning rate: 2e-4;

- learning rate warmup: 8000 updates;
- learning rate decay: inverse square root;
- mini-batch size determined automatically to fit GPU memory;
- early stopping after 10 consecutive validations with no improvement in mean word cross-entropy score.

4.4 Evaluation metrics

We use sacreBLEU (Post, 2018) to calculate BLEU[2] (Papineni et al., 2002) and chrF[3] (Popović, 2015).

We acknowledge the potential for overfitting to the WMT22 COMET[4] metric used for MBR decoding. Therefore, we extended the evaluation to also include CometKiwi[5] (Rei et al., 2022), UniTE[6] (Wan et al., 2022), UniTE-DA[7] (Rei et al., 2023) and BLEURT-20[8] (Sellam et al., 2020b).

For the English–Hausa experiments, we additionally calculated scores using AfriCOMET-STL (Wang et al., 2023), which was specifically trained to evaluate translations involving certain African languages.

4.5 English to German

To train the baseline model, we used all corpora from the MTData toolkit (version 0.4.0) (Gowda et al., 2021), excluding the validation sets and the test sets from the available datasets. Our filters described in Section 4.1 reduced the dataset from approximately 800 million sentences to 400 million.

In the context of domain adaptation, we employed the following list of domain data:

1. 40 thousand sentences from biomedical-translation-corpora (Neves et al., 2016);
2. 3 million sentences from Ufal medical corpus shared in WMT23 (Kocmi et al., 2023);
3. 2 million sentences from EMEA corpus downloaded from OPUS (Tiedemann and Nygaard, 2004).

After deduplication, we were left with 3 million sentences which we split into two datasets. We considered a scenario with 1 million bilingual parallel sentences and approximately 2 million monolingual sentences in the source language. Khresmoi-dev (Dušek et al., 2017) concatenated with FLORES-200 (NLLB Team et al., 2022) was utilized as the validation set during training. We did not apply any filtering to the domain data.

We used the above data to train the following models:

- Baseline (**Baseline**) – model trained only on data from the MTdata toolkit.
- Baseline + mix-tuning (**Mix-tune**) – fine-tuned **Baseline** model on 1 million in-domain bilingual data concatenated with 1 million general-domain data randomly sampled from the **Baseline** training set.
- Baseline + domain MBR (**Base-domain-mbr**) – fine-tuned **Baseline** model on 2 million domain-specific sentences from MBR-decoded forward translations.
- Mix-tuned + domain MBR (**Mix-tune-domain-mbr**) – fine-tuned **Mix-tune** model on 2 million domain-specific sentences from MBR-decoded forward translations.
- Mix-tuned + MBR-iteration2 (**Mix-tune-domain-mbr-iter2**) – fine-tuned **Mix-tune-domain-mbr** on the 2 million domain-specific sentences from MBR-decoded forward translations.
- Mix tuned + general-MBR (**Mix-tune-general-mbr**) – fine-tuned **Mix-tune** model on 2 million sentences sampled from the general-domain corpora from the **Baseline** training set as MBR-decoded forward translations.

When fine-tuning the **Mix-tune** model, we tailor the learning rate setup to meet specific requirements: learn-rate: 1e-7, lr-decay-inv-sqrt: 16000, lr-warmup: 16000. All remaining fine-tuning procedures employ an adjusted learning rate set to 5e-6.

[2] BLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1
[3] chrF signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1
[4] https://huggingface.co/Unbabel/wmt22-comet-da
[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da
[6] https://huggingface.co/Unbabel/unite-mup
[7] https://huggingface.co/Unbabel/wmt22-unite-da
[8] https://storage.googleapis.com/bleurt-oss-21/BLEURT-20.zip

4.6 Czech to Ukrainian

We leveraged all of the Czech–Ukrainian parallel data from the WMT23 MTData recipe, resulting in approximately 8 million sentence pairs after filtering as described in Section 4.1. We did not include any additional monolingual data in this experiment.

We utilized the FLORES-200 dataset for validation during training, while the WMT22 test set served as an additional benchmark.

We trained the baseline model only on the parallel data, using hyperparameters as described in Section 4.3. Next, we translated the source side of the parallel corpus used in training with our baseline model, saving a list of translation candidates. We performed MBR decoding, selecting the best translation of each set of candidate translations, resulting in a synthetic training dataset.

We investigated the following approaches to leverage the MBR-decoded data for model improvement:

- Standard fine-tuning (**MBR-finetuned**) – we fine-tuned the baseline model on the MBR-decoded data, using a learning rate of 5e-6.
- Fine-tuning with a high learning rate (**MBR-ft-high-lr**) – we fine-tune the baseline model on MBR-decoded data, using a learning rate of 2e-4.
- Resuming training with MBR-decoded data (**MBR-resumed**) – we switched the training set to the MBR-decoded version and resumed training, restoring the optimizer state and effectively continuing its training with the improved data.

4.7 English to Hausa

To train the models in the English–Hausa direction, we used data from the WMT shared tasks from previous years. Specifically, we used:

1. 7 million sentences from OPUS;
2. 2.4 million data from the WMT23 African MT Shared Task (Kocmi et al., 2023);
3. 150 thousand sentences from ParaCrawl v8.0 (Bañón et al., 2020).

The deduplication process reduced the data size to approximately 9 million sentences. Following the filtering criteria detailed in Section 4.1, a total of 3.1 million sentences were retained. We used FLORES-200 for validation during training. After training, we evaluated the model on the FLORES-200 and NTREX test sets.

We took similar steps as in the Czech–Ukrainian experiment, training a baseline model with hyperparameters set as described in Section 4.3. We conducted experiments employing MBR decoding, comparing its performance using two distinct metrics as the utility function:

- WMT22 COMET – based on XLM-RoBERTa (Conneau et al., 2020), covering a diverse set of 100 languages,
- AfriCOMET-STL – based on AfroXLM-RoBERTa (Alabi et al., 2022), covering 17 African languages and 3 high-resource languages.

We investigated the impact of the chosen metric for MBR decoding by training two models using the refined translations:

- **MBR-COMET** – training resumed with the training set switched to the WMT22 COMET MBR-decoded version.
- **MBR-AfriCOMET** – training resumed with the training set switched to the AfriCOMET-STL MBR-decoded version.

5 Results

The statistical significance of the evaluation results is assessed using a paired bootstrap resampling test (Koehn, 2004), involving 1000 resampling trials to confirm the statistical significance of the model improvements ($p < 0.05$).

5.1 Number of translation samples and search algorithm

To determine the optimal setup for MBR decoding, we conducted experiments involving the translation and evaluation of chosen test sets with various MBR decoding sample sizes and two decoding algorithms. This approach offers the advantages of being both representative and computationally efficient compared to training MT models on the entire MBR-decoded training set.

We evaluated two decoding algorithms – beam search and top-k. For the top-k setup, we experimented with temperature values of 0.1 and 1, keeping the k parameter equal to 10. These choices

were based on the work done by Freitag et al. (2023). To determine the best number of samples for MBR decoding we conducted experiments with the following numbers of samples: 10, 25, 50, 100, 200, 300, 400, 500.

Firstly we noted that beam search is the preferred option, given its high scores and greater stability across different metric results, as observed in Figure 1 and 2. We provide more specific results in the Appendix Figures 4, 5.

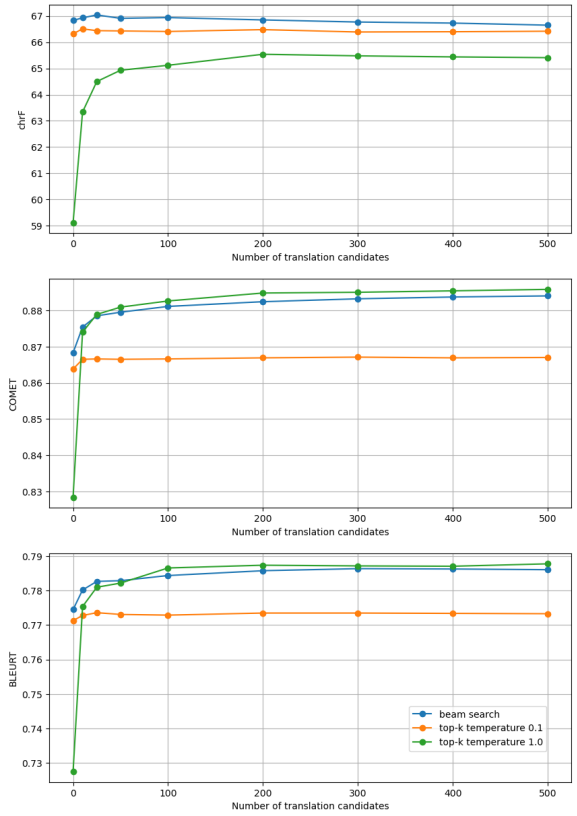

Figure 1: Comparison of beam search and top-k algorithms of the **Mix-tune** English–German model for the khresmoi test set. Top-k algorithm with temperature 1.0 showed superior performance on neural metrics over top-k with temperature 0.1 and slightly better performance than beam search. However, beam search achieved the highest score on the chrF metric, while the top-k algorithm with temperature 1.0 had the lowest score (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

Secondly, we decided to train our models on MBR-decoded data from 50 candidates selected by the beam search decoding algorithm. We considered the balance between improvement in evaluation metrics based on neural language models, stability across lexical metrics, and the execution time of MBR decoding, as shown in Figure 3.

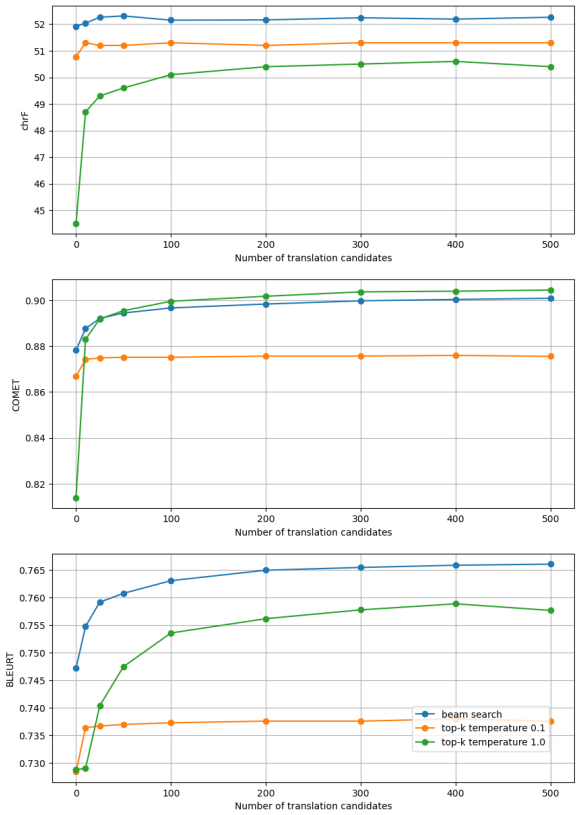

Figure 2: Comparison of beam search and top-k algorithms of the **baseline** Czech–Ukrainian model for the FLORES-200 test set. Beam search seems to be the superior option with the best performance on chrF and BLEURT metrics and slightly worse results on COMET over top-k with temperature 1.0 (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

We provide more detailed results in the Appendix Figures 6, 7, 8, 9, 10, 11, 12.

5.2 English to German

Table 1 shows the evaluation results on the in-domain test set khresmoi. All models self-improved with MBR decoding have shown enhanced performance. However, model **Mix-tune-domain-mbr-iter2** did not exhibit improvement over its first iteration **Mix-tune-domain-mbr**, even on COMET, which was the utility metric of MBR decoding. **Mix-tune-general-mbr** model shows a slightly better performance on BLEURT metric compared to models fine-tuned on in-domain MBR-decoded forward translations.

Table 2 presents the evaluation results on the FLORES-200 test set. Although chrF did not increase, the neural evaluation metrics showed improvement. Similar to the khresmoi test set, the **Mix-tune-domain-mbr-iter2** model showed a decrease in quality during the second iteration of self-

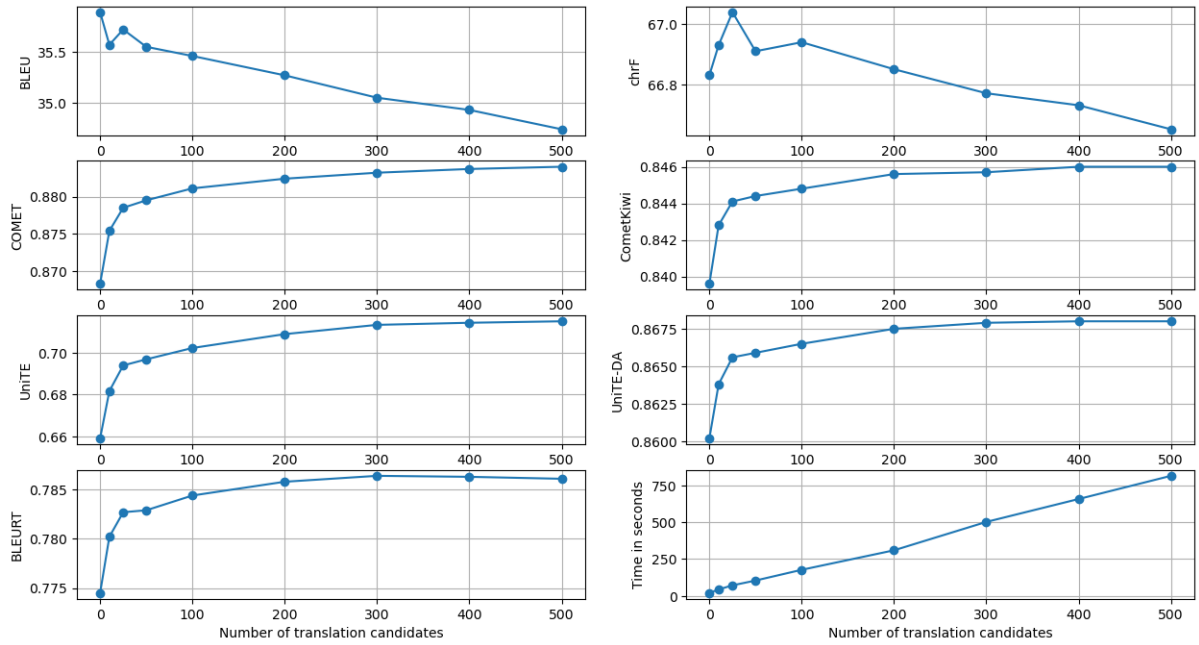

Figure 3: Comparison of beam search performance with a different number of samples of the **Mix-tune** English–German model for the khresmoi test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains, and performance on the n-gram metrics deteriorated (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

Model	chrF	COMET	BLEURT
Baseline	66.6	0.8653	0.7693
Mix-tune	66.8	0.8682	0.7749
Base-domain-mbr	**66.9**	0.8711*	0.7755
Mix-tune-domain-mbr	**66.9**	**0.8728***	0.7792*
Mix-tune-domain-mbr-iter2	**66.9**	0.8727*	0.7791*
Mix-tune-general-mbr	**66.9**	0.8720*	**0.7799***

Table 1: English–German khresmoi set results for the MBR self-improvement approaches. All models fine-tuned with MBR self-improvement technique have shown better performance over **Baseline** and **Mix-tune** models, including the **Mix-tune-general-mbr** model, which was finetuned on general-domain MBR-decoded data. The results marked with an asterisk (*) are statistically significant compared to the **Mix-tune** model.

Model	chrF	COMET	BLEURT
Baseline	**67.5**	0.8751	0.7735
Mix-tune	**67.5**	0.8756	0.7744
Base-domain-mbr	67.2	0.8772	0.7743
Mix-tune-domain-mbr	67.3	0.8787*	0.7766
Mix-tune-domain-mbr-iter2	67.1	0.8766	0.7748
Mix-tune-general-mbr	**67.5**	**0.8813***	**0.7784***

Table 2: English–German FLORES-200 test set results for the MBR self-improvement approaches. **Mix-tune-general-mbr** model has shown superior performance, however, models with domain-specific forward translation maintain performance. The results marked with an asterisk (*) are statistically significant compared to the **Mix-tune** model.

improvement. **Mix-tune-general-mbr** showed superior performance over other models.

In summary, our findings demonstrate that applying MBR decoding significantly improves the performance of the high-resource English–German model for low-resource biomedical domain translation, particularly on neural network metrics. While lexical metrics show lower stability, they also hold potential for improvement.

Experiments demonstrated the robustness of self-improving models with the MBR decoding technique. Model fine-tuned on general forward translation had great performance on the in-domain test set and the model fine-tuned on domain-specific forward translation maintained performance on the general domain test set. We provide a broader evaluation in the Appendix Tables 9, 10, 11, 12.

5.3 Czech to Ukrainian

The results of the three MBR self-improvement approaches described in Section 4.6 are presented in Tables 3 and 4 for the FLORES-200 and WMT22 test sets, respectively.

We find that standard fine-tuning of the baseline model with MBR-decoded data yields the smallest improvements across all metrics, suggesting its limited effectiveness in this context. We note that both fine-tuning with a higher learning rate and

Model	chrF	COMET	BLEURT
Baseline	52.0	0.8779	0.7466
MBR-finetuned	52.4	0.8839	0.7522
MBR-ft-high-lr	**52.7**	**0.8869**	0.7553
MBR-resumed	**52.7**	0.8864	**0.7557**

Table 3: Czech–Ukrainian FLORES-200 test set results for the three MBR self-improvement approaches. All self-improved models exhibit improvements on all metrics compared to the baseline model, regardless of the fine-tuning approach used. Notably, both **MBR-ft-high-lr** and **MBR-resumed** models achieve the highest gains, demonstrating comparable performance. All self-improved models show statistical significance compared to the **Baseline** model.

Model	chrF	COMET	BLEURT
Baseline	58.4	0.8721	0.7498
MBR-finetuned	60.0	0.8803	0.7574
MBR-ft-high-lr	**60.2**	0.8844	0.7619
MBR-resumed	60.0	**0.8852**	**0.7639**

Table 4: Czech–Ukrainian WMT22 test set results for the three MBR self-improvement approaches. Similar to the FLORES-200 results, all self-improved models exhibit improvements on all metrics compared to the baseline model. However, on the WMT22 test set, the neural metrics favour the **MBR-resumed** model over the **MBR-ft-high-lr** model. All self-improved models show statistical significance compared to the **Baseline** model.

Model	chrF	COMET	BLEURT
Baseline	52.0	0.8779	0.7466
MBR-resumed	52.7*	0.8864*	0.7557*
MBR-resumed-iter2	**52.8**	0.8888*	**0.7567**
MBR-resumed-iter3	52.6	**0.8901**	0.7557

Table 5: Czech–Ukrainian iterative self-improvement results on the FLORES-200 test set. While the COMET score consistently improves across all three iterations, the chrF and BLEURT scores show a decrease in the third iteration. This suggests that the model overfits to COMET, harming the quality of the translation. Results with an asterisk (*) are statistically significant in comparison with the model in the row directly above it.

resuming the training exhibit comparable performance, with resumed training achieving slightly better results on the WMT22 test set. This may indicate that resuming training helps mitigate overfitting to the FLORES-200 validation set used during training.

Tables 5 and 6 showcase the impact of iterative training with MBR decoding on the FLORES-200 and WMT22 test sets, respectively. The second iteration consistently improves scores across all metrics, demonstrating the effectiveness of the

Model	chrF	COMET	BLEURT
Baseline	58.4	0.8721	0.7498
MBR-resumed	60.0*	0.8852*	0.7639*
MBR-resumed-iter2	**60.3***	0.8885*	**0.7641**
MBR-resumed-iter3	60.1	**0.8896**	0.7578

Table 6: Czech–Ukrainian iterative self-improvement results on the WMT22 test set. Consistent with the FLORES-200 results, the COMET score improves across all iterations, while other metrics show a decrease in the last iteration. Notably, the BLEURT score not only decreases but falls below the score achieved by the first self-improved model. Results with an asterisk (*) are statistically significant in comparison with the model in the row directly above it.

iterative self-improvement process in refining the model's translation capabilities. However, the third iteration leads to a decrease in both chrF and BLEURT scores. This suggests potential overfitting to the MBR decoding utility metric, where the model prioritizes aspects that score well according to COMET but may not translate to overall translation quality.

We provide extended evaluations in the Appendix in Tables 13, 14, 15, 16.

5.4 English to Hausa

Model	chrF	COMET	BLEURT	AfriCOMET
Baseline	49.9	0.7569	0.7931	0.6984
MBR-COMET	50.9	**0.7720**	**0.8083**	0.7207
MBR-AfriCOMET	**51.2**	0.7692	0.8061	**0.7239**

Table 7: English–Hausa FLORES-200 test set results for MBR self-improvement with different metrics. Both self-improved models achieve gains compared to the baseline model on all evaluation metrics. While the AfriCOMET-based model achieves a higher AfriCOMET score, reflecting its alignment with the specific evaluation metric, the COMET-based model surpasses it in both BLEURT and COMET scores, while showing a comparable gain on the AfriCOMET score. All self-improved models show statistical significance compared to the **Baseline** model.

Model	chrF	COMET	BLEURT	AfriCOMET
Baseline	51.6	0.7596	0.7791	0.6800
MBR-COMET	**53.1**	**0.7752**	**0.7986**	0.7046
MBR-AfriCOMET	53.0	0.7721	0.7956	**0.7062**

Table 8: English–Hausa NTREX test set results for MBR self-improvement with different metrics. Similar to the FLORES-200 results, both self-improved models using MBR decoding demonstrate improvements over the baseline model on all evaluation metrics. All self-improved models show statistical significance compared to the **Baseline** model.

This section compares the performance of

two MBR decoding self-improvement approaches for English–Hausa translation: one utilizing the WMT22 COMET model and another using the AfriCOMET model. The results are presented in Tables 7 and 8 for the FLORES-200 and NTREX test sets, respectively.

We observe that the AfriCOMET MBR-tuned model achieves gains over the WMT22 COMET MBR-tuned model on chrF for the FLORES-200 test set, but this advantage is not replicated on the NTREX test set. Additionally, the gains from AfriCOMET MBR-tuning are mainly limited to the AfriCOMET metric.

Our analysis reveals that the **MBR-AfriCOMET** model exhibits improvements over the **MBR-COMET** model primarily on lexical metrics in the case of the FLORES-200 test set, but not in the case of NTREX. The gains of the **MBR-AfriCOMET** model are mainly limited to AfriCOMET metrics, while other neural-based metrics consistently favour the **MBR-COMET** model.

While WMT22 COMET might exhibit a lower correlation with human judgment for the English–Hausa language pair than AfriCOMET, as reported by Wang et al. (2023), both self-improved models achieved significant and comparable gains on AfriCOMET. This suggests that WMT22 COMET, can still correctly rerank translation candidates and effectively guide the self-improvement process, leading to improvements on AfriCOMET, a metric specifically designed for African languages. This finding suggests that self-improvement guided by WMT22 COMET, with its diverse language coverage, might be effective even in low-resource settings, potentially reducing the need for additional adaptation of neural evaluation models to individual languages.

Additional evaluations are provided in the Appendix in Tables 17, 18.

6 Conclusion

This study demonstrated the effectiveness of model self-improvement through MBR decoding in improving translation quality. This approach proves beneficial for both high and low-resource languages, offering versatility in its application across diverse scenarios. Examples include domain-specific translation and the enhancement of general translation models.

We conducted experiments with various sample sizes for MBR decoding, using two decoding algorithms: beam search and top-k. The aim was to find a balance between automatic metric gains and time efficiency. Our experiments have shown that the beam search algorithm with a beam size set to 50 is the optimal choice.

In the field of high-resource English-to-German biomedical translation, we investigated the impact of domain adaptation using various self-improvement approaches on MBR-decoded forward-translated data. Experiments showed that all MBR-based fine-tuning, regardless of the domain of the test set, improved performance compared to the baseline model. This finding highlights the robustness of the self-improvement technique.

Experiments on the Czech–Ukrainian language pair revealed that fine-tuning the MT model on MBR-decoded translations of the training data set significantly improves translation performance. Applying this process iteratively improves quality, but further iterations yield diminishing gains and at some point, the quality may even degrade due to overfitting to the MBR decoding utility metric.

In the English–Hausa experiments, we employed two models for MBR decoding: WMT22 COMET and AfriCOMET. Both models yielded comparable and significant improvements in automatic metrics, indicating their effectiveness in guiding the self-improvement process. While AfriCOMET, specifically trained on African languages, might intuitively seem favourable for this language pair, the performance of the **MBR-COMET** model highlights the potential of utilizing more widely applicable metrics like WMT22 COMET even for low-resource settings.

References

Alabi, Jesujoba O., David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Amrhein, Chantal and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In He, Yulan, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the*

12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1125–1141, Online only, November. Association for Computational Linguistics.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Burchell, Laurie, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada, July. Association for Computational Linguistics.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Dušek, Ondřej, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Fernandes, Patrick, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In Carpuat, Marine, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States, July. Association for Computational Linguistics.

Finkelstein, Mara, Subhajit Naskar, Mehdi Mirzazadeh, Apurva Shah, and Markus Freitag. 2023. Mbr and qe finetuning: Training-time distillation of the best and most expensive decoding methods.

Freitag, Markus, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Freitag, Markus, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for\\minimum bayes risk decoding for machine translation. *arXiv preprint arXiv:2305.09860*.

Gowda, Thamme, 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, August. Association for Computational Linguistics.

Gulcehre, Caglar, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Jon, Josef, Martin Popel, and Ondřej Bojar. 2023. CUNI at WMT23 general translation task: MT and a genetic algorithm. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 119–127, Singapore, December. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, 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, July. Association for Computational Linguistics.

Kocmi, Tom, 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, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

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

Kudo, Keito, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 128–136, Singapore, December. Association for Computational Linguistics.

Kumar, Shankar 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, May 2 - May 7. Association for Computational Linguistics.

Neves, Mariana, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia, May. European Language Resources Association (ELRA).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Nowakowski, Artur, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam Mickiewicz University at WMT 2022: NER-assisted and quality-aware neural machine translation. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326–334, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

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

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

Rei, Ricardo, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online, November. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rei, Ricardo, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada, July. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur P Parikh. 2020b. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Tiedemann, Jörg and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In Lino, Maria Teresa, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vernikos, Giorgos and Andrei Popescu-Belis. 2024. Don't rank, combine! combining machine translation hypotheses using quality estimation.

Wan, Yu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland, May. Association for Computational Linguistics.

Wang, Jiayi, David Ifeoluwa Adelani, Sweta Agrawal, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Marek Masiak, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Hamam Mokayede, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Tosin Adewumi, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, and Pontus Stenetorp. 2023. Afrimte and africomet: Empowering comet to embrace under-resourced african languages.

Zaragoza-Bernabeu, Jaume, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France, June. European Language Resources Association.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	66.6	35.6	0.8653	0.8373	0.6441	0.8574	0.7693
Mix-tune	66.8	**35.9**	0.8682	0.8397	0.6594	0.8602	0.7749
Base-domain-mbr	**66.9**	35.7	0.8711	0.8416	0.6694	0.8621	0.7755
Mix-tune-domain-mbr	**66.9**	35.8	**0.8728**	**0.8423**	0.6766	0.8631	0.7792
Mix-tune-domain-mbr-iter2	**66.9**	35.6	0.8727	**0.8423**	0.6757	**0.8633**	0.7791
Mix-tune-general-mbr	**66.9**	35.5	0.8720	0.8422	**0.6775**	0.8631	**0.7799**

Table 9: English–German khresmoi set results for the MBR self-improvement approaches. All models fine-tuned with MBR self-improvement technique have shown better performance over **Baseline** and **Mix-tune** models, even **Mix-tune-general-mbr** model with general forward translations.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	63.1	35.0	0.8505	0.8336	0.5368	0.8470	0.7500
Mix-tune	63.5	35.6	0.8525	0.8360	0.5418	0.8495	0.7541
Base-domain-mbr	63.5	35.8	**0.8549**	0.8374	0.5549	0.8501	0.7522
Mix-tune-domain-mbr	63.6	35.7	0.8540	0.8379	0.5552	0.8508	0.7530
Mix-tune-domain-mbr-iter2	**63.7**	**35.9**	0.8543	**0.8383**	0.5575	**0.8510**	0.7535
Mix-tune-general-mbr	63.4	35.4	0.8547	0.8378	**0.5613**	0.8501	**0.7542**

Table 10: English–German WMT22-medline set results for the MBR self-improvement approaches. All models fine-tuned with MBR self-improvement technique have shown better performance over **Mix-tune** model except on metric BLEURT. On this specific test set, **Mix-tune-domain-mbr-iter2** outperformed the **Mix-tune-domain-mbr** model, unlike the results observed on other test sets.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	**67.5**	42.0	0.8751	0.8454	0.6630	0.8614	0.7735
Mix-tune	**67.5**	**42.2**	0.8756	0.8457	0.6657	0.8617	0.7744
Base-domain-mbr	67.2	41.7	0.8772	0.8469	0.6677	0.8632	0.7743
Mix-tune-domain-mbr	67.3	41.7	0.8787	0.8477	0.6719	0.8641	0.7766
Mix-tune-domain-mbr-iter2	67.1	41.5	0.8766	0.8466	0.6653	0.8629	0.7748
Mix-tune-general-mbr	**67.5**	41.8	**0.8813**	**0.8484**	**0.6824**	**0.8654**	**0.7784**

Table 11: English–German FLORES-200 test set results for the MBR self-improvement approaches. **Mix-tune-general-mbr** model has shown superior performance, however, models with domain-specific forward translation maintain performance.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	63.8	36.6	0.8428	0.8328	0.5308	0.8420	0.7106
Mix-tune	63.7	36.5	0.8427	0.8322	0.5283	0.8414	0.7107
Base-domain-mbr	63.3	35.8	0.8463	0.8359	0.5376	0.8454	0.7138
Mix-tune-domain-mbr	63.2	35.9	0.8468	0.8358	0.5404	0.8464	0.7132
Mix-tune-domain-mbr-iter2	63.0	35.5	0.8460	0.8345	0.5348	0.8455	0.7119
Mix-tune-general-mbr	**64.1**	**36.7**	**0.8629**	**0.8399**	**0.5622**	**0.8492**	**0.7202**

Table 12: English–German Statmt test set results for the MBR self-improvement approaches. **Mix-tune-general-mbr** model has shown significantly improved performance on every metric, however models with domain-specific forward translation maintain performance.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	52.0	22.2	0.8779	0.8449	0.4441	0.9017	0.7466
MBR-finetuned	52.4	22.3	0.8839	0.8513	0.4715	0.9063	0.7522
MBR-ft-high-lr	**52.7**	22.6	**0.8869**	**0.8543**	**0.4829**	0.9085	0.7553
MBR-resumed	**52.7**	**22.8**	0.8864	0.8540	0.4824	**0.9086**	**0.7557**

Table 13: Extended Czech–Ukrainian FLORES-200 test set results for the three MBR self-improvement approaches. All approaches lead to an increase in evaluation scores. Both **MBR-ft-high-lr** and **MBR-resumed** models achieve the highest gains, demonstrating comparable performance.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	58.4	31.1	0.8721	0.8046	0.3744	0.8795	0.7498
MBR-finetuned	60.0	32.3	0.8803	0.8121	0.4112	0.8846	0.7574
MBR-ft-high-lr	**60.2**	**33.2**	0.8844	0.8152	**0.4246**	0.8880	0.7619
MBR-resumed	60.0	33.0	**0.8852**	**0.8162**	0.4236	**0.8890**	**0.7639**

Table 14: Extended Czech–Ukrainian WMT22 test set results for the three MBR self-improvement approaches. As in the case of evaluation results on the FLORES-200 test set, all approaches improve upon the baseline model, although **MBR-resumed** stands out across all neural metrics apart from UniTE.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	52.0	22.2	0.8779	0.8449	0.4441	0.9017	0.7466
MBR-resumed	52.7	**22.8**	0.8864	0.8540	0.4824	0.9086	0.7557
MBR-resumed-iter2	**52.8**	22.6	0.8888	0.8557	**0.4882**	**0.9099**	**0.7567**
MBR-resumed-iter3	52.6	22.3	**0.8901**	**0.8562**	0.4873	0.9097	0.7557

Table 15: Extended Czech–Ukrainian iterative self-improvement results on the FLORES-200 test set. Models increase in quality across all neural metrics until the third iteration, when the quality measured by metrics other than COMET and CometKiwi decreases. It's worth noticing that the BLEU score increases only in the first iteration and slowly degrades in consecutive iterations.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT
Baseline	58.4	31.1	0.8721	0.8046	0.3744	0.8795	0.7498
MBR-resumed	60.0	**33.0**	0.8852	0.8162	0.4236	0.8890	0.7639
MBR-resumed-iter2	**60.3**	32.6	0.8885	**0.8183**	**0.4349**	**0.8900**	**0.7641**
MBR-resumed-iter3	60.1	31.9	**0.8896**	0.8174	0.4312	0.8887	0.7578

Table 16: Extended Czech–Ukrainian iterative self-improvement results on the WMT22 test set. Evaluations across all metrics show similar tendencies as in the case of FLORES-200, except for CometKiwi which also decreases in the third iteration.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT	AfriCOMET
Baseline	49.9	22.3	0.7569	0.5597	-0.2297	0.6082	0.7931	0.6984
MBR-COMET	50.9	23.2	**0.7720**	**0.5707**	**-0.1777**	**0.6233**	**0.8083**	0.7207
MBR-AfriCOMET	**51.2**	**23.4**	0.7692	0.5638	-0.1878	0.6183	0.8061	**0.7239**

Table 17: Extended English–Hausa results on the FLORES-200 test set. According to lexical metrics and AfriCOMET, the **MBR-AfriCOMET** model shows the greatest improvement. However, other neural metrics suggest that the **MBR-COMET** model is superior.

Model	chrF	BLEU	COMET	CometKiwi	UniTE	UniTE-DA	BLEURT	AfriCOMET
Baseline	51.6	23.9	0.7596	0.5704	-0.1763	0.6294	0.7791	0.6800
MBR-COMET	**53.1**	**25.3**	**0.7752**	**0.5865**	**-0.1051**	**0.6484**	**0.7986**	0.7046
MBR-AfriCOMET	53.0	24.9	0.7721	0.5803	-0.1273	0.6409	0.7956	**0.7062**

Table 18: Extended English–Hausa results on the NTREX test set. In contrast to evaluations on the FLORES-200 test set, in this case only the AfriCOMET metric favours the **MBR-AfriCOMET** model.

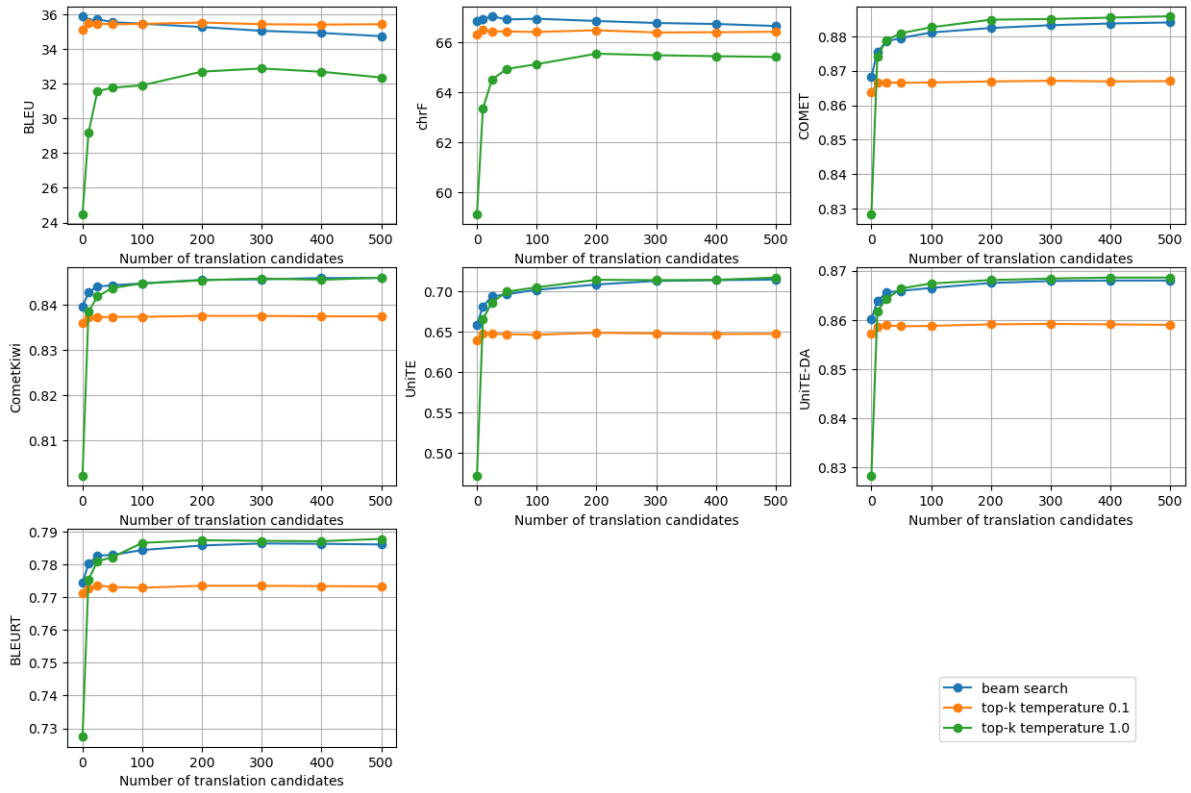

Figure 4: Comparison of beam search and top-k algorithms of the **Mix-tune** English–German model for the khresmoi test set. Top-k algorithm with temperature 1.0 showed superior performance on neural metrics over top-k with temperature 0.1 and slightly better performance than beam search. However, beam search achieved the highest score on the chrF metric, while the top-k algorithm with temperature 1.0 had the lowest score for lexical metrics (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

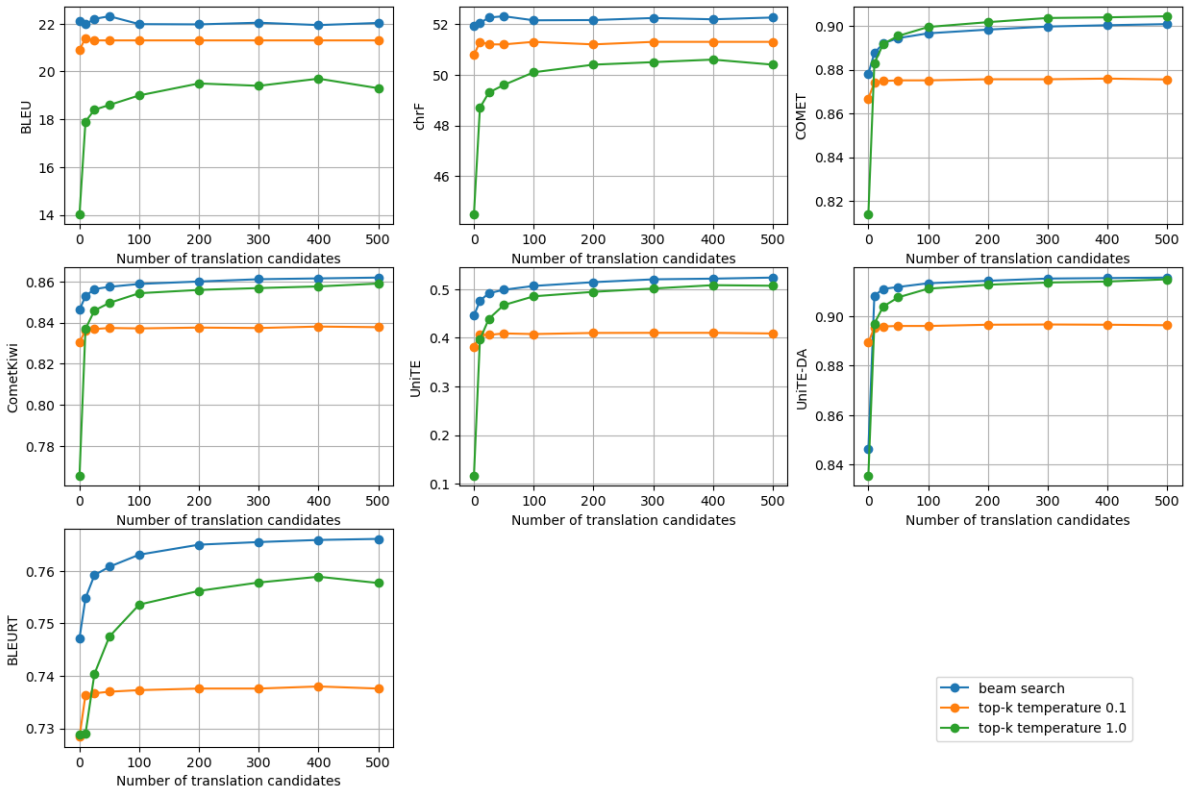

Figure 5: Comparison of beam search and top-k algorithms of the **baseline** Czech–Ukrainian model for the FLORES-200 test set. Beam search seems to be the superior option with the best performance on every metric except COMET (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

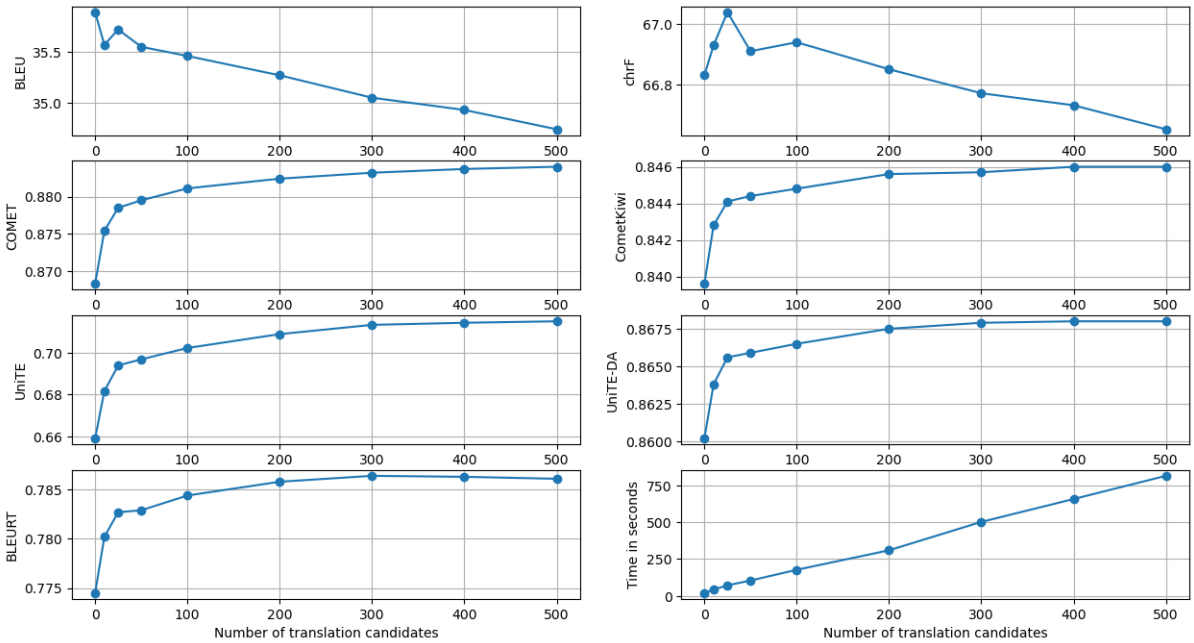

Figure 6: Comparison of beam search performance with a different number of samples of the **Mix-tune** English–German model for the khresmoi test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains and performance on the n-gram metrics deteriorated (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

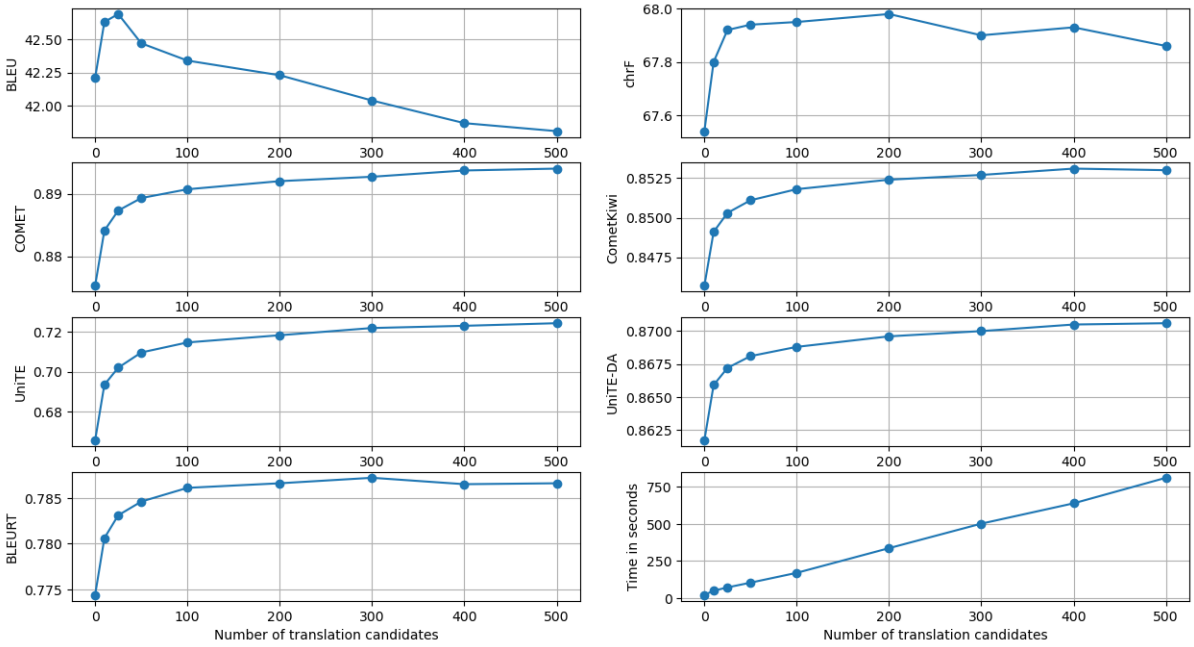

Figure 7: Comparison of beam search performance with a different number of samples of the **Mix-tune** English–German model for the FLORES-200 test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains and performance on the n-gram metrics deteriorated (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

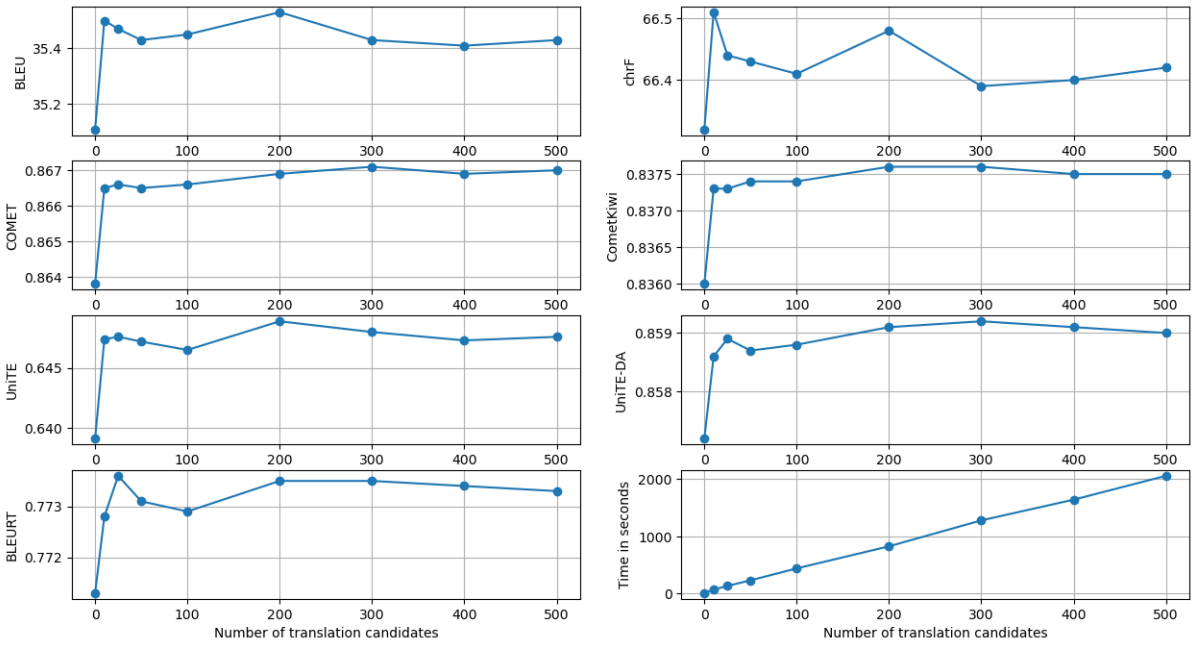

Figure 8: Comparison of top-k performance (temperature 0.1, k=10) with different number of samples of the **Mix-tune** English–German model for the khresmoi test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

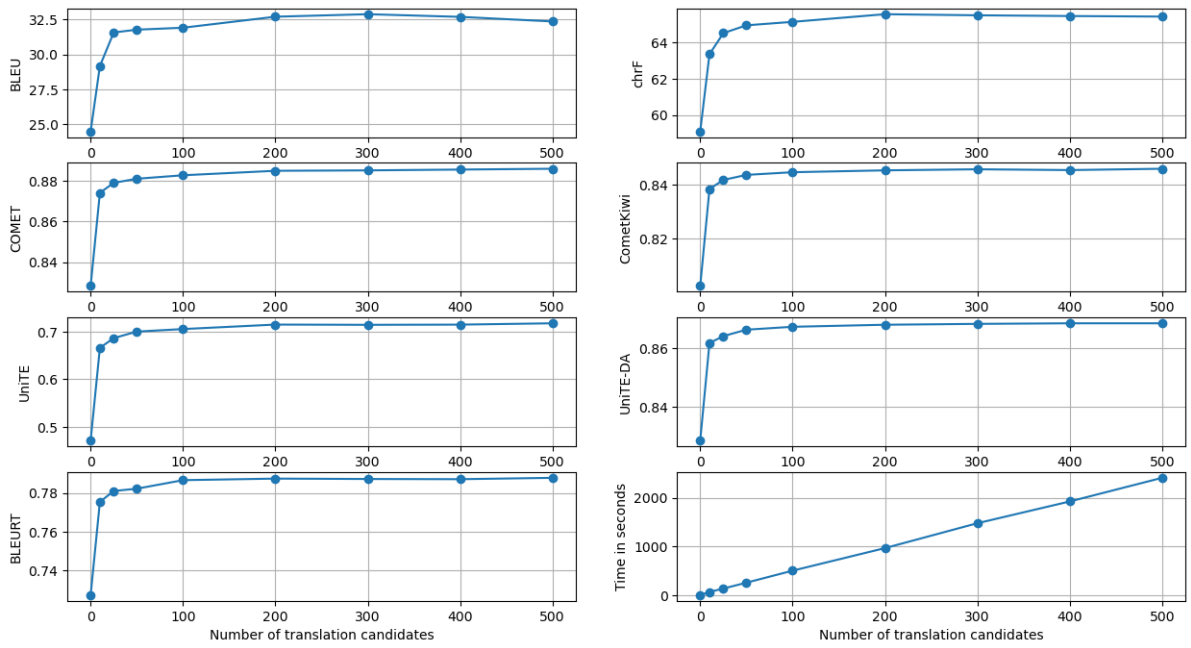

Figure 9: Comparison of top-k performance (temperature 1.0, k=10) with different number of samples of the **Mix-tune** English–German model for the khresmoi test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

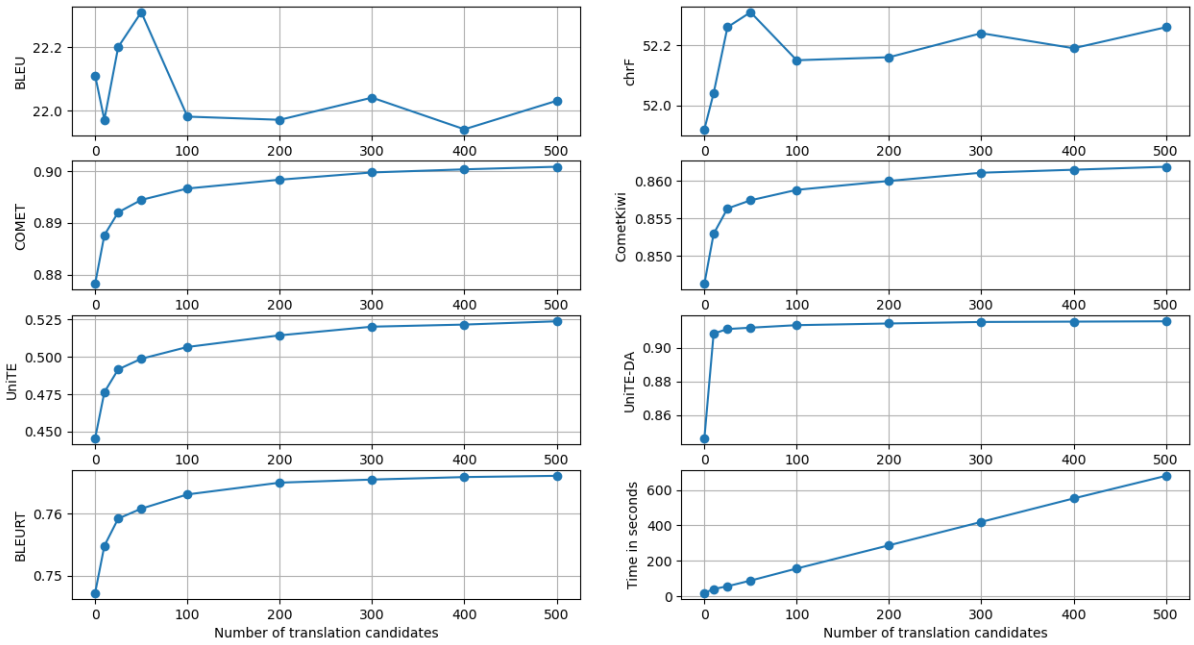

Figure 10: Comparison of beam search performance with different number of samples of the **Baseline** Czech–Ukrainian model for the FLORES-200 test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains and performance on the n-gram metrics deteriorated (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

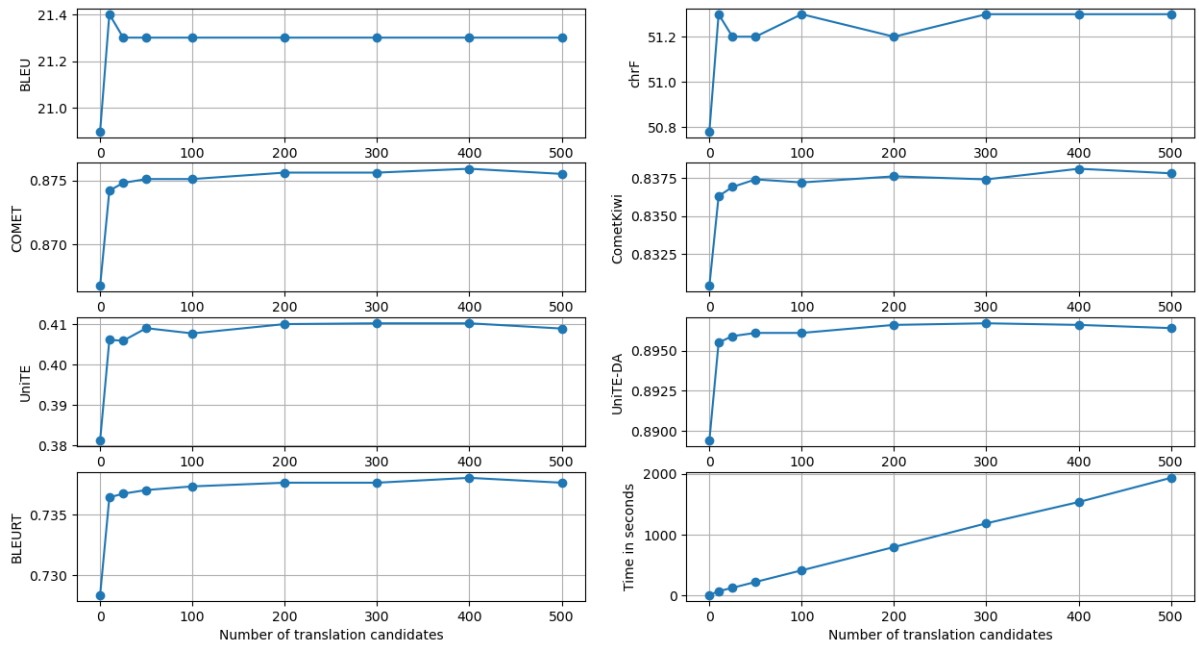

Figure 11: Comparison of top-k performance (temperature 0.1, k=10) with different number of samples of the **Baseline** Czech–Ukrainian model for the FLORES-200 test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).

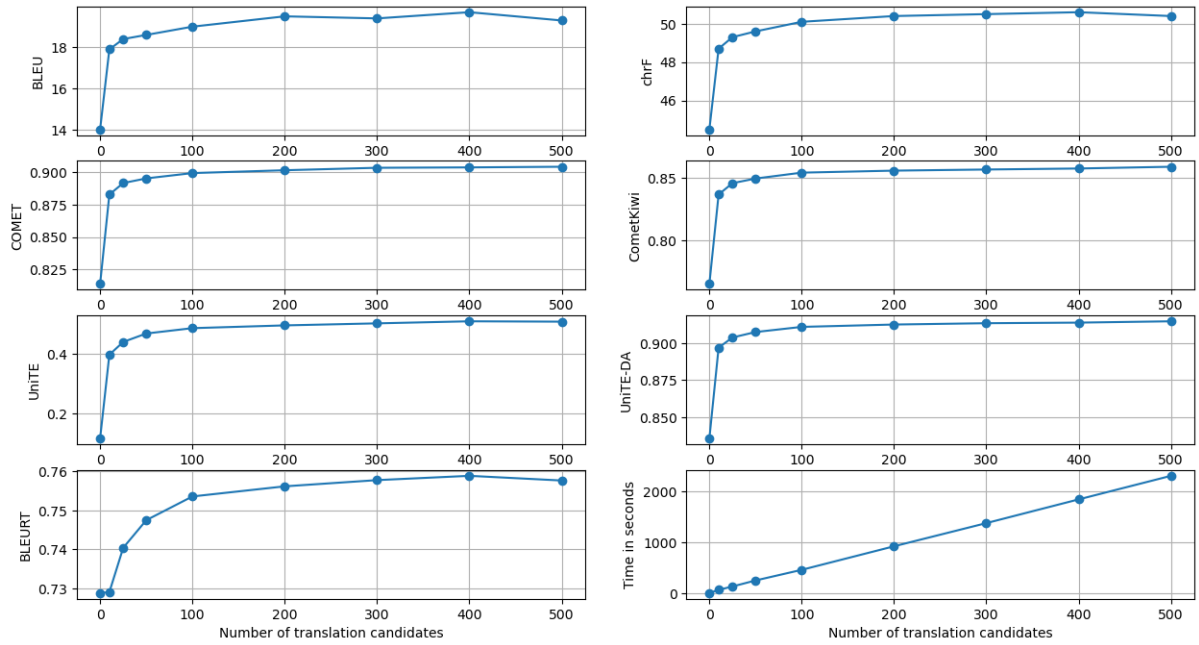

Figure 12: Comparison of top-k performance (temperature 1.0, k=10) with different number of samples of the **Baseline** Czech–Ukrainian model for the FLORES-200 test set. Initial increases in the number of samples for MBR decoding showed very rapid gains, but further increases no longer resulted in such large gains (translation without MBR decoding is represented on the chart as the number of translation candidates equal to 0).