CUNI at WMT23 General Translation Task: MT and a Genetic Algorithm

Josef Jon and **Martin Popel** and **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{jon,popel,bojar}@ufal.mff.cuni.cz

Abstract

This paper presents the contributions of Charles University teams to the WMT23 General translation task (English to Czech and Czech to Ukrainian translation directions). Our main submission, CUNI-GA, is a result of applying a novel n-best list reranking and modification method on translation candidates produced by the two other submitted systems, CUNI-Transformer and CUNI-DocTransformer (document-level translation only used for the $en \rightarrow cs$ direction). Our method uses a genetic algorithm and MBR decoding to search for optimal translation under a given metric (in our case, a weighted combination of ChrF, BLEU, COMET22-DA, and COMET22-QE-DA). Our submissions are first in the constrained track and show competitive performance against top-tier unconstrained systems across various automatic metrics.

1 Introduction

Our submission for this year's WMT General translation task (Kocmi et al., 2023) is based on the previous submissions of our team (Popel et al., 2019, 2022) and MBR decoding in combination with genetic algorithm (GA). We describe the method in separate work (Jon and Bojar, 2023). The main goal of our submission is to find out whether our approach improves the translation quality perceived by humans. For this reason, we submitted both the base system translations and the mutated and reranked (i.e. GA-processed) translations for the human evaluation.

As all the parts of the approach are described in detail in the mentioned papers (as well as all the related work), we will restrict ourselves to providing a short overview of the main points in Section 2. In Section 3, we describe the datasets, tools and parameters used to obtain results presented in Section 4. Finally, we draw conclusions from the results.

2 Methods

Our submissions make use of two features that are not typical for current MT systems: document-level context and translation refinement through a genetic algorithm.

2.1 Document level translation

We use document-level NMT for the $en \rightarrow cs$ direction. The approach is described in Popel et al. (2019). Since all the training data for this direction have document boundaries, a document-level training set is created by extracting all sequences of consecutive sentences with at most 3000 characters. The final training set consists of pairs of such examples, where both sides have the same number of sentences. Sentences are separated by a special token. We also use Block backtranslation (Popel, 2018; Popel et al., 2020; Gebauer et al., 2021; Jon et al., 2022a).

2.2 Genetic algorithm

Our approach (Jon and Bojar, 2023) utilizes MBR decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004; Amrhein and Sennrich, 2022; Freitag et al., 2021; Müller and Sennrich, 2021; Jon et al., 2022b) in conjunction with the genetic algorithm (GA) (Fraser, 1957; Bremermann, 1958; Holland, 1975). By merging and mutating translations generated by an MT system, we aim to find the best translation under a specific metric. This is a new strategy for creating translation candidates in NMT. We illustrate one iteration of the whole process in Figure 1. The top, yellow part shows the steps that are the same as in simple reranking. We have an initial population of candidates, for example, n-best list produced by an MT model, that is scored by *fitness function*, in our case, a sum of MBR decoding scores using an MT evaluation metric and QE scores. At this point, for reranking, the process would stop after selecting the best-scoring translation candidate. In GA, we continue by splitting

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 119–127
December 6–7, 2023.

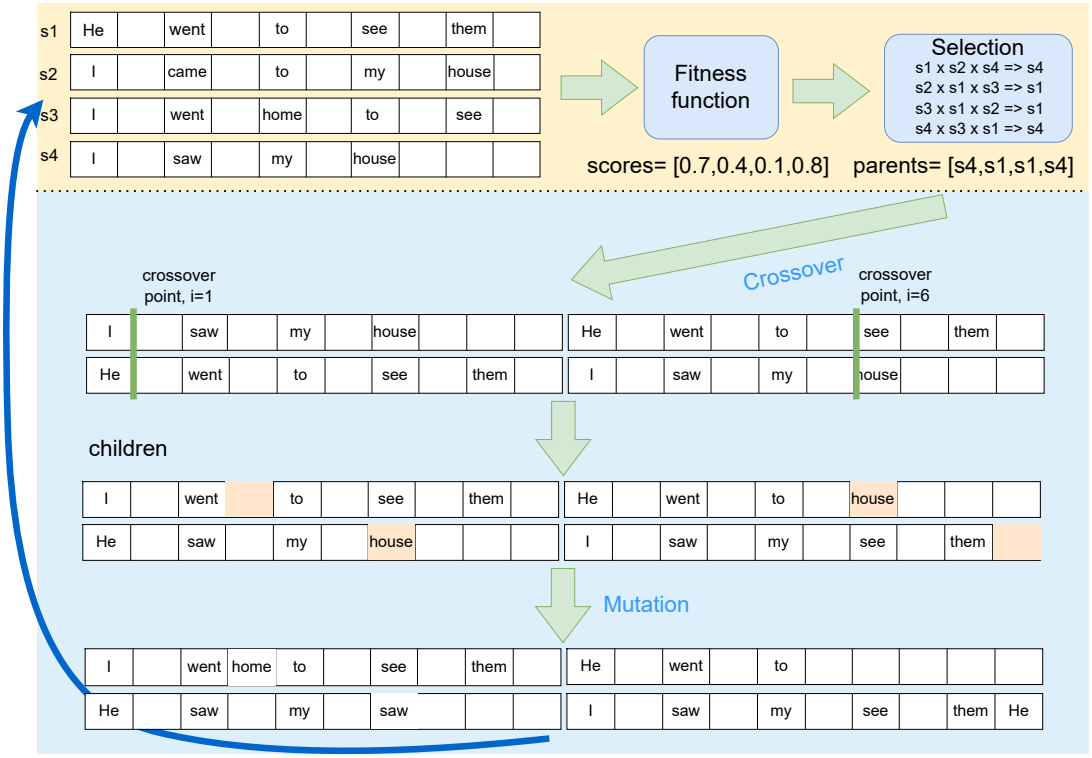

Figure 1: One iteration of the GA algorithm for a population of 4 individuals. The steps with a yellow background are equivalent to simple reranking, the steps with blue background introduce the operations of the genetic algorithm. Figure taken from Jon and Bojar (2023)

a well-scoring subset of the candidate sentences at random points and reattaching them in a different order, by a process called cross-over. These combined candidates are mutated at random places, meaning some of the tokens are either deleted or replaced by different tokens from a set of suitable candidate tokens. Also, new tokens can be added this way. These modifications result in a new population of translation candidates and the whole process is repeated from the start. A more detailed description of our approach is available in Jon and Bojar (2023).

MBR decoding NMT models generate a probability distribution over potential translations for a specified input sentence. The widely used method to derive the ultimate translation from this distribution is "maximum-a-posteriori" (MAP) decoding. However, the computational demands of precise MAP decoding lead to the adoption of approximations like beam search, referenced by Koehn et al. (2003). Recent literature, such as Stahlberg and Byrne (2019) and Meister et al. (2020), has shed light on several constraints of MAP and proposed alternatives.

MBR decoding is one such alternative. It uses a utility function to select the translation, aiming to minimize expected loss or *risk*. Typically, MT metrics are employed as these utility functions. In practice, candidate translations produced by the MT model are used as an approximation of the set of all possible translations. In such case, if we only use purely reference-based metrics (like BLEU), MBR decoding becomes a consensus decoding, where the chosen candidate is the one closest to all the others. However, novel MT metrics also take source sentence into account, so the process is more complex than a simple search for the most average translation. The MBR decoding has seen renewed interest with the introduction of the new generation of metrics (Amrhein and Sennrich, 2022; Freitag et al., 2021; Müller and Sennrich, 2021; Jon et al., 2022b).

3 System description

Our models are based on submissions of our team from previous years (Popel et al., 2022, 2019). We resubmit those (*CUNI-Transformer* and *CUNI-DocTransformer* submissions) and we also submit an additional translation: the outputs of these models combined, mutated and rescored by the GA

described in Section 2.2 (*CUNI-GA* submission).

3.1 Tools and data

All our submissions are constrained, using only the training data provided by the task organizers, specifically the CzEng 2.0 (Kocmi et al., 2020) corpus. We used English to Czech `newstest-18` and `newstest-22` as validation sets for the genetic algorithm approach. Due to the computational requirements of our method, we only evaluate the first 150 sentences of each test set. We didn't run any validation experiments for GA in the $cs \rightarrow uk$ language pair, we used the same parameters as for $en \rightarrow cs$. We have only translated the general translation test set using GA, the test suits translations for *CUNI-GA* are copied from the *CUNI-DocTransformer* submission.

3.2 Models

We use Transformer models. For the dev set experiments, we use same models as Jon and Bojar (2023) (i.e. `transformer-big` using MarianNMT (Junczys-Dowmunt et al., 2018) with default hyperparameters). For the final submissions, the models are the same as in last year's submissions: Popel et al. (2022) for $cs \rightarrow uk$ and Popel et al. (2019) for $en \rightarrow cs$.

3.3 GA parameters

We refrained from searching for the optimal values of GA parameters due to the significant computational demands of our method.

For the results on the validation set, we used exactly the settings described by Jon and Bojar (2023), i.e. Transformer model trained on the CzEng 2.0 (Kocmi et al., 2020) corpus in $cs \rightarrow en$ direction (i.e. the opposite direction to the task). We used beam search with size 20 to produce a 20-best list and sampled an additional 20 translations from the model to create an initial population of 40 candidates, which we copied 50 times to obtain a population size of 2000.

We used different NMT models (see Section 3.2) and a different number of initial sentences for the shared task submissions. For the $cs \rightarrow uk$ direction, the starting population consists of the top 35 hypotheses produced by beam search from the two models described in Popel et al. (2022) (top-10 from the *CUNI-Transformer-inca-roman* and top-25 from the *CUNI-Transformer* model).[1] This set is replicated 50 times, leading to a total population of 1750 candidates. For $en \rightarrow cs$ we use a concatenation of n-best lists with beam sizes 4 and 10 from both *CUNI-DocTransformer* and *CUNI-Transformer* (28 candidates in total), also copied 50 times over, resulting in population size of 1400. To combine document-level and sentence-level translations, we re-split the translated documents back into sentences.

To choose parents for the succeeding generation, we use tournament selection with $n = 3$. These parents are then merged at a crossover rate of $c = 0.1$. The mutation rate, for altering non-empty genes (i.e. tokens) to other non-empty genes m, is $1/l$, where l denotes the chromosome's (chromosome is a sequence of tokens, representation of one translation candidate) length[2] For transitions from an empty to a non-empty gene (i.e. addition of a word) and vice versa (i.e. deletion), the rate is $\frac{m}{10}$. The GA runs for 250 and 130 generations for $cs \rightarrow uk$ and $en \rightarrow cs$, respectively.

3.4 Metrics

The translations are evaluated by the following metrics: ChrF (Popović, 2015), BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), multiple versions of COMET (Rei et al., 2020, 2021, 2022b,a,c) and UniTE (Wan et al., 2022). We abbreviate some of the longer metrics' names further in the text in order to save space.[3]

For both BLEU and ChrF, we utilize SacreBLEU (Post, 2018). In all experiments, ChrF uses a $\beta = 2$ setting (ChrF2). We rely on the original implementations for COMET,[4] BLEURT,[5] and UniTE[6] scores.

4 Results

This section presents automatic metric scores on validation sets and the official test set.

4.1 English to Czech

The first translation direction is English to Czech, where we submitted the outputs of our older sentence-level (*CUNI-Transformer*) and document-level (*CUNI-DocTransformer*) systems, as well as

[1] The *CUNI-Transformer-inca-roman* uses preprocessing using romanization and inline casing (Popel et al., 2022).

[2] See Jon and Bojar (2023) for a more detailed description.

[3] CMT20 (wmt20-comet-da), CMT21 (wmt21-comet-mqm), CMTH22 (eamt22-cometinho-da), QE20 (wmt20-comet-qe-da-v2), QE22 (wmt22-cometkiwi-da), BLEURT (BLEURT-20), UniTE (UniTE-MUP)

[4] `https://github.com/Unbabel/COMET`

[5] `https://github.com/google-research/bleurt`

[6] `https://github.com/NLP2CT/UniTE`

Method	Fitness	ChrF	BLEU	CMT20	CMT21	CMTH22	QE20	BLEURT	UniTE	% new
baseline	-	56.7	30.1	0.5007	0.0399	0.5017	0.2477	0.7078	0.3018	0
Reranking	CMT20	57.4	31.2	0.5853	0.0409	0.5390	0.2930	0.7193	0.3413	0
Reranking	CMT20+QE20+BLEU	**57.5**	**31.2**	0.5983	0.0417	**0.5596**	0.3620	**0.7255**	0.3686	0
GA	CMT20	56.2	28.4	0.6247	0.0410	0.5382	0.2893	0.7177	0.3366	52
GA	CMT20+QE20+BLEU	**57.5**	29.5	**0.6266**	**0.0429**	0.5403	**0.4198**	0.7174	**0.3946**	70

Table 1: Comparison of the scores of baseline MT output, reranked output, and GA-modified output. The last column shows the percentage of finally selected best translations that were not present in the initial population (i.e. they were newly created by the GA operations). Table from Jon and Bojar (2023).

Model	w_{CMT}	w_{QE}	w_{BLEU}	w_{chrF}	chrF	BLEU	CMT20	CMT21	CMTH22	QE20	CMT22	BLEURT	UniTE	New
Baseline	-	-	-	-	56.6	30.1	0.500	0.040	0.504	0.244		0.707	0.301	0.00
	0.15	0.15	0.35	0.35	57.2	29.8	0.619	**0.043**	0.542	0.401	0.856	0.715	0.384	0.64
	0.1	0.1	0.4	0.4	**57.4**	30.0	0.616	**0.043**	0.541	0.403	0.856	0.714	0.385	0.63
	0.25	0.25	0.25	0.25	**57.4**	29.8	0.616	**0.043**	0.541	0.410	0.857	0.713	0.388	0.64
	0.2	0.2	0.3	0.3	57.3	29.6	0.619	**0.043**	0.540	0.406	0.857	0.715	0.388	0.64
CMT20	0.4	0.2	0.2	0.2	57.2	**30.7**	0.629	**0.043**	**0.549**	0.388	0.856	**0.720**	0.384	0.51
	0.4	0.3	0.1	0.2	**57.4**	30.2	0.630	**0.043**	0.548	0.405	0.857	0.718	0.384	0.65
	0.4	0.4	0.1	0.1	57.1	29.0	0.631	**0.043**	0.542	0.427	**0.859**	0.716	**0.389**	0.68
	0.5	0.5	0	0	55.2	25.1	**0.633**	**0.043**	0.514	**0.470**	0.856	0.705	0.372	0.86
	1	0	0	0	56.8	29.7	0.614	0.041	0.533	0.289	0.844	0.712	0.336	0.51
	0.15	0.15	0.35	0.35	57.5	32.0	0.601	0.042	0.560	0.332	0.858	0.729	0.388	0.27
	0.1	0.1	0.4	0.4	**57.7**	**32.2**	0.602	0.042	**0.562**	0.331	0.858	0.730	0.392	0.28
	0.25	0.25	0.25	0.25	57.5	32.0	0.601	0.042	0.560	0.330	0.857	0.729	0.388	0.29
CMT22	0.2	0.2	0.3	0.3	57.5	32.0	0.601	0.042	0.561	0.331	0.858	0.730	0.394	0.32
	0.4	0.2	0.2	0.2	57.2	31.5	0.593	0.042	0.550	0.326	0.857	0.727	0.370	0.25
	0.4	0.3	0.1	0.2	**57.7**	32.1	0.597	0.042	0.555	0.332	0.857	0.728	0.386	0.27
	0.4	0.4	0.1	0.1	57.6	32.0	0.606	0.042	0.560	0.334	0.859	0.730	0.393	0.29
	0.5	0.5	0	0	**57.7**	31.7	**0.620**	**0.043**	**0.562**	**0.359**	**0.866**	**0.731**	**0.406**	0.57
	1	0	0	0	56.8	29.8	0.570	0.042	0.528	0.328	0.863	0.714	0.344	0.49

Table 2: Scores of translations on the first 150 sentences of `newstest-18` created by GA. The fitness metric is a weighted sum of COMET, COMET-QE, BLEU and chrF, with weight shown in columns 2 to 5. The first column shows which version of COMET and COMET-QE was used. Higher is better for all the metrics. The best results for each metric are bold.

Model	w_{CMT}	w_{QE}	w_{BLEU}	w_{chrF}	chrF	BLEU	CMT20	CMT21	CMTH22	QE20	CMT22	BLEURT	UniTE	New
Baseline	-	-	-		68.3	44.9	0.738	0.045	0.751	0.357	0.876	0.785	0.540	0.00
	0.15	0.15	0.35	0.35	68.4	43.0	0.777	0.047	0.777	0.464	0.890	0.787	0.607	0.52
	0.1	0.1	0.4	0.4	68.6	43.5	0.779	0.047	0.779	0.464	0.891	0.787	0.609	0.51
	0.25	0.25	0.25	0.25	68.3	43.0	0.785	0.047	**0.783**	0.469	0.892	0.789	**0.617**	0.52
	0.2	0.2	0.3	0.3	68.5	43.3	0.780	0.047	0.777	0.465	0.891	0.787	0.610	0.52
CMT20	0.4	0.2	0.2	0.2	**68.6**	**44.2**	0.778	0.047	0.773	0.441	0.887	**0.791**	0.586	0.33
	0.4	0.3	0.1	0.2	68.2	43.0	0.785	0.047	0.777	0.470	0.891	0.789	0.612	0.49
	0.4	0.4	0.1	0.1	67.7	42.1	**0.787**	**0.047**	0.777	**0.485**	**0.892**	0.788	0.614	0.55
	0.5	0.5	0	0	65.1	36.4	0.782	0.047	0.747	0.514	0.887	0.771	0.574	0.77
	1	0	0	0	67.9	42.1	0.772	0.046	0.760	0.386	0.880	0.785	0.552	0.36
	0.15	0.15	0.35	0.35	68.8	45.1	0.771	0.047	0.794	0.417	0.890	0.799	0.604	0.25
	0.1	0.1	0.4	0.4	68.8	45.1	0.772	0.047	**0.795**	0.417	0.890	0.798	0.605	0.25
	0.25	0.25	0.25	0.25	68.8	44.8	0.774	0.047	0.792	0.418	0.890	0.799	0.603	0.27
CMT22	0.2	0.2	0.3	0.3	68.9	45.3	0.772	0.047	0.794	0.417	0.890	0.799	0.604	0.25
	0.4	0.2	0.2	0.2	68.9	45.1	0.771	0.047	0.794	0.408	0.889	0.795	0.602	0.22
	0.4	0.3	0.1	0.2	**69.1**	**45.7**	0.772	0.047	0.794	0.410	0.889	0.798	0.607	0.25
	0.4	0.4	0.1	0.1	68.8	45.2	0.771	0.047	0.792	0.420	0.890	0.798	0.603	0.25
	0.5	0.5	0	0	68.6	43.6	**0.782**	**0.047**	0.793	**0.431**	**0.893**	**0.800**	**0.612**	0.46
	1	0	0	0	68.2	43.5	0.762	0.046	0.778	0.401	0.890	0.788	0.576	0.39

Table 3: Scores of translations on the first 150 sentences of `newstest-22` created by GA. The fitness metric is a weighted sum of COMET, COMET-QE, BLEU and chrF, with weight shown in columns 2 to 5. The first column shows which version of COMET and COMET-QE was used. Higher is better for all the metrics. The best results for each COMET version are bold.

a combination and modification of both using our GA approach.

GA vs. reranking Jon and Bojar (2023) provide a comparison of the genetic algorithm approach to a simple reranking using the same objective metrics. In that work, a sum of CMT20, QE20 and BLEU is used as the fitness metric. The results are copied in Table 1. The baseline translations are obtained via beam search. The same work also shows that for UniTE, CMT22, CMT21-MQM held-out metrics[7], GA significantly outperforms simple reranking with the same objective metric. However, BLEURT, CMTH22 and chrF seem to favor reranking only.

For our current work, we ran additional experiments. We use a weighted sum of COMET, COMET-QE, chrF and BLEU as the objective (fitness) metric. We compare older and newer versions of both COMET and COMET-QE, represented by CMT20/QE20 and CMT22/QE22, respectively. Since the objective metrics lose their relevance for evaluation once we optimize for them, a set of held-out metrics is selected to better estimate the translation quality. The results for the first 150 sentences of `newstest18` are presented in Table 2, and the scores for the first 150 sentences of `newstest22` are presented in Table 3.

We vary the weights of the different fitness metrics to see the effect on the held-out metrics (columns $\mathbf{w_{CMT}}$, $\mathbf{w_{QE}}$, $\mathbf{w_{BLEU}}$ and $\mathbf{w_{chrF}}$). The last column shows a portion of cases where the final selected candidate was not part of the initial population, the other columns show values of the respective scores.

We see an interesting difference between CMT22/QE22 and CMT20/QE20. While optimizing only for CMT20 or CMT20+QE20 hurts other scores greatly (for example UniTe and BLEURT), optimizing solely for CMT22+QE22 does not have such an adverse effect on other metrics. We hypothesize multiple factors play a role in this. One of them might be the better robustness of the newer versions, which are designed to deal better with hallucinations and unexpected target tokens that could be introduced by the GA. CMT20 and especially QE20 were previously shown to be partially insensitive to this kind of errors (Guerreiro et al., 2023), but they could be detected by the other metrics, hence the lower scores.

Final submission Overall, the results suggest the best choice is to simply average CMT22 and QE22 scores ($w_{CMT} = 0.5$, $w_{QE} = 0.5$). We did not have the complete evaluation at hand by the time of the submission, so we used weights $w_{CMT} = 0.4$, $w_{QE} = 0.4$, $w_{BLEU} = 0.1$ and $w_{chrF} = 0.1$ for the submitted test set translation. We use a completely different NMT system than in the dev set experiments to create the initial population for the submission, as described in 3.3.

We show the automatic scores of all the submissions on the test set in Table 4. The *CUNI-GA* submission outperforms both the base submissions *CUNI-Transformer* and *CUNI-DocTransformer* across all metrics. It ranks comparably to the best unconstrained system using COMET, but lags behind in chrF and BLEU.

We analyzed the percentages of the final submitted translated sentences that were present in some of the initial n-best lists and the percentage of novel sentences, created by the GA. We show these results in Table 5. We see that 21.7% of the final submitted sentences are new, not contained in any of the initial n-best lists, but rather created by the GA mutation and crossover operations.

4.2 Czech to Ukrainian

We also ran the GA on a concatenation of n-best lists produced by the two $cs \rightarrow uk$ models, see Popel et al. (2022) for details on these systems. We used beam size 10 for the *CUNI-Transformer-inca-roman* model and beam size 25 for the *CUNI-Transformer* model, resulting in 35 initial candidate sentences. We did not perform any parameter tuning on the validation set, we used the same parameters as for the $en \rightarrow cs$ submission. We present the automatic metrics results on the test set in Table 6. Our submissions outperform the only other constrained system and are competitive with the unconstrained systems, scoring best in COMET and 2nd in chrF and BLEU. For COMET and chrF, GA outperforms the unmodified baseline translation, while in BLEU, the baseline scores slightly better.

Again, we show what is the percentage of final best translations selected for submission contained in either of the initial n-best lists and the percentage of new translations, created by GA operations, in Table 7. 35.1% of the final submitted translations are novel.

[7]Means metrics not used as a part of the fitness function. Note that these metrics are not completely independent, they can be still linked to the fitness metrics by spurious correlations caused by data and model architecture similarity

System	COMET
ONLINE-W	91.8
CUNI-GA	90.8
ONLINE-B	89.9
GPT4-5shot	89.4
ONLINE-A	88.4
CUNI-DocTransformer	88.3
GTCOM_Peter	87.7
ONLINE-M	87.4
Lan-BridgeMT	87.3
CUNI-Transformer	87.2
NLLB_Greedy	87.1
ONLINE-Y	87.0
NLLB_MBR_BLEU	86.9
ONLINE-G	85.9
ZengHuiMT	85.4

System	chrF
ONLINE-W	76.3
ONLINE-B	70.4
ZengHuiMT	67.5
ONLINE-A	66.3
CUNI-GA	65.9
GTCOM_Peter	65.4
CUNI-DocTransformer	65.1
ONLINE-Y	64.6
CUNI-Transformer	63.9
Lan-BridgeMT	63.8
ONLINE-G	63.7
ONLINE-M	63.2
GPT4-5shot	62.3
NLLB_Greedy	60.0
NLLB_MBR_BLEU	59.1

System	BLEU
ONLINE-W	59.4
ONLINE-B	50.1
ONLINE-A	43.4
CUNI-GA	43.3
ZengHuiMT	43.1
CUNI-DocTransformer	42.5
GTCOM_Peter	42.3
CUNI-Transformer	41.4
ONLINE-Y	40.8
Lan-BridgeMT	40.7
ONLINE-G	39.6
ONLINE-M	39.6
GPT4-5shot	37.8
NLLB_Greedy	35.9
NLLB_MBR_BLEU	35.1

Table 4: Results of automatic evaluation on $en \rightarrow cs$ testset. Unconstrained systems are indicated with a grey background. Coincidentally, all three $en \rightarrow cs$ unconstrained systems are our submissions described in this paper. CUNI-GA is better than the two baselines according to all three metrics.

	doc-4	doc-10	sent-4	sent-10	new
contains	36.3%	42.1%	31.4%	52.8%	21.7%
unique	2%	6.8%	0.7%	17.5%	
merge	50.0%		53.5%		
u-merge	24.7%		33.3%		

Table 5: Percentages of final best scoring in CUNI-GA English to Czech submission sentences by the initial n-best list they are contained in (doc-4 denotes document-level, beam size 4 and so on). The first row shows how many sentences from the final translation were present in the respective n-best list, while the last column shows the percentage of completely new sentences, that were not present in any of the lists. The second row looks at the percentages of final sentences that are uniquely in exactly one of the lists. The last two rows show the same for merged doc-level and sent-level lists, i.e. we concatenated both beam sizes for each into one list.

5 Future work

Our setting allows many straightforward modifications to potentially improve the results of our method. First of all, MBR decoding works well on a large, diverse set of initial candidates, obtained for example by sampling. In our experiments, we only use short n-best lists produced by beam search. An additional benefit stemming from the diversity of the initial candidates is a more dive diverse set of possible tokens for replacement mutations.

Second, we did not run any search for the parameters of the GA process (crossover and mutation rates, number of generations, population size, selection method), due to the large computational costs of this approach. We believe a set of better parameters could be found easily by, for example, a grid search. Finally, the metrics used for the fitness function are combined by a simple weighted sum. Multi-criterion genetic algorithms can be explored for a better approach to combine multiple evaluation scores for the translations.

Also, reranking and modifying the translations on a sentence level can introduce inconsistencies previously mitigated by using document-level MT, losing the advantages of document-level processing. Deutsch et al. (2023) show that using sentence-level metrics for whole document-level segments might be a viable option for avoiding this issue.

6 Conclusion

We confirm that using MBR decoding in combination with a genetic algorithm can improve scores in selected evaluation metrics, while creating original novel translations. We show that our systems are competitive in both submitted language pairs, winning among constrained systems based on automated evaluation metrics.

7 Acknowledgements

This work was partially supported by GAČR EXPRO grants NEUREM3 (19-26934X) and LUSyD (20-16819X), TAČR grant EdUKate (TQ01000458), by the Grant Agency of Charles University in Prague (GAUK 244523) and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254). It has been using data and tools 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. LM2023062).

System	COMET
CUNI-GA	90.9
GPT4-5shot	90.8
ONLINE-W	89.4
GTCOM_Peter	88.9
ONLINE-B	88.8
ONLINE-A	88.2
CUNI-Transformer	88.0
ONLINE-G	87.7
MUNI-NLP	87.0
ONLINE-Y	86.5
NLLB_Greedy	86.3
NLLB_MBR_BLEU	86.3
Lan-BridgeMT	86.0

System	chrF
GPT4-5shot	61.0
CUNI-GA	57.9
GTCOM_Peter	57.6
CUNI-Transformer	57.4
MUNI-NLP	57.0
Lan-BridgeMT	55.7
ONLINE-W	55.0
ONLINE-B	54.7
ONLINE-A	54.4
ONLINE-G	53.7
ONLINE-Y	53.4
NLLB_Greedy	52.5
NLLB_MBR_BLEU	52.3

System	BLEU
GPT4-5shot	32.8
CUNI-Transformer	30.2
GTCOM_Peter	29.8
CUNI-GA	29.5
MUNI-NLP	28.3
Lan-BridgeMT	27.5
ONLINE-W	26.8
ONLINE-B	25.7
ONLINE-A	25.4
NLLB_MBR_BLEU	25.1
NLLB_Greedy	24.9
ONLINE-G	24.8
ONLINE-Y	24.2

Table 6: Results of automatic evaluation on $cs \rightarrow uk$ testset. Unconstrained systems are indicated with a grey background. CUNI-GA is better than CUNI-Transformer according to COMET and chrF, but worse according to BLEU.

	CT-inca-roman-10	CT-25	new
contains	17%	58.5%	35.1%
unique	6.4%	48%	

Table 7: Percentages of final best scoring sentences by the initial n-best list they are contained in, the meaning of the rows is the same as in Table 5. CT=CUNI-Transformer.

References

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

Hans J Bremermann. 1958. *The evolution of intelligence: The nervous system as a model of its environment*. University of Washington, Department of Mathematics.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level.

Alex S Fraser. 1957. Simulation of genetic systems by automatic digital computers ii. effects of linkage on rates of advance under selection. *Australian Journal of Biological Sciences*, 10(4):492–500.

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.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

John H. Holland. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI. Second edition, 1992.

Josef Jon and Ondřej Bojar. 2023. Breeding machine translations: Evolutionary approach to survive and thrive in the world of automated evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2191–2212, Toronto, Canada. Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2022a. CUNI-bergamot submission at WMT22 general translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 280–289, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2022b. CUNI-Bergamot Submission at WMT22 General Translation Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 280–289, Abu Dhabi. 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, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme

Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

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.

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

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

Martin Popel, Jindřich Libovický, and Jindřich Helcl. 2022. CUNI systems for the WMT 22 Czech-Ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 352–357, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy. Association for Computational Linguistics.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, 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 *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

Ricardo Rei, 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. 2022c. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. 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.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F. Wong, and Lidia S. Chao. 2022. UniTE: Unified Translation Evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.