Direct Preference Optimization for Neural Machine Translation with Minimum Bayes Risk Decoding

Guangyu Yang, Jinghong Chen, Weizhe Lin, Bill Byrne
Department of Engineering
University of Cambridge
{gy266, jc2124, wl356, wjb31}@cam.ac.uk

Abstract

Minimum Bayes Risk (MBR) decoding can significantly improve translation performance of Multilingual Large Language Models (MLLMs). However, MBR decoding is computationally expensive. We show how the recently developed Reinforcement Learning technique, Direct Preference Optimization (DPO), can fine-tune MLLMs to get the gains of MBR without any additional computation in inference. Our method uses only a small monolingual fine-tuning set and yields significantly improved performance on multiple NMT test sets compared to MLLMs without DPO.

1 Introduction

MBR decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2022; Suzgun et al., 2023) is a two-pass procedure that generates multiple translation hypotheses and selects a hypothesis based on Bayesian risk. Recent work (Garcia et al., 2023; Suzgun et al., 2023; Yang, 2023) has shown that MBR decoding can significantly boost the translation performance of MLLMs (Lin et al., 2022; Muennighoff et al., 2023; Zeng et al., 2023a), outperforming greedy decoding and beam search. However, MBR decoding is expensive, both in computation and in latency.

Our goal is to fine-tune a base MLLM so that it has the same single-pass decoding performance as MBR decoding. We propose a novel self-supervised fine-tuning method based on DPO (Rafailov et al., 2023). Our method uses MBR decoding on an MLLM to produce a preference dataset consisting of pairs of ranked translations. The DPO algorithm is used to fine-tune the MLLM to prefer the higher-ranked translations over lower-ranked ones. MLLMs optimized for MBR preference achieve significantly better translation performance when decoded with beam search, achieving translation quality on par with MBR decoding of the original model.

2 MBR and DPO

We follow the expectation-by-sampling approach to MBR (Eikema and Aziz, 2022). Given a set of sampled translations $H(\mathbf{x}) = \{\mathbf{y}' \sim P(\cdot|\mathbf{x})\}$ and a loss (or utility) function $L(\cdot, \cdot)$, the score (negative Bayes risk) of each translation is found as

$$S(\mathbf{y}) = -\frac{1}{|H(x)|} \sum_{\mathbf{y}' \in H(\mathbf{x})} L(\mathbf{y}', \mathbf{y}) \tag{1}$$

and the MBR hypothesis is then computed as

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in H(\mathbf{x})} S(\mathbf{y}) \tag{2}$$

This is simple but expensive. Our goal is to train a model that produces translations with scores consistent with MBR, but without multi-step decoding.

2.1 DPO Fine-Tuning Objective

DPO (Rafailov et al., 2023) reformulates the usual approach to Reinforcement Learning from Human Feedback (RLHF) so as to avoid a distinct reward modelling step. The typical RLHF criteria is

$$\begin{aligned} \max_{\pi_\theta} \; & \mathbb{E}_{\mathbf{x} \sim D, \mathbf{y} \sim \pi_\theta(\mathbf{y}|\mathbf{x})} \left[r_\phi(\mathbf{x}, \mathbf{y}) \right] \\ & - \beta \mathbb{D}_{KL} \left[\pi_\theta(\mathbf{y}|\mathbf{x}) \parallel \pi_{\text{ref}}(\mathbf{y}|\mathbf{x}) \right] \end{aligned} \tag{3}$$

where r_ϕ is a reward model trained from human feedback, π_θ is the model being trained, and π_{ref} is the reference model. DPO effectively replaces the reward model with a preference distribution based on π_θ, the model being trained; DPO also retains the KL regularization term with weighting β.

The preference dataset D for DPO consists of triplets $(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)$ where $\mathbf{x}$ is the input prompt, $\mathbf{y}_w$ is the winnng (prefered) response, and $\mathbf{y}_l$ is the losing (disprefered) response. DPO uses the language model likelihood to approximate the reward as $\beta \log \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}$. During training, with π_θ typically

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 391–398
June 16-21, 2024 ©2024 Association for Computational Linguistics

initialized from π_{ref}, the objective is to maximize the expected reward margin between $\mathbf{y}_w$ and $\mathbf{y}_l$:

$$L_{\text{DPO}} = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim D}[\log\sigma(M(\mathbf{y}_w,\mathbf{y}_l,\mathbf{x},\theta))] \tag{4}$$

where the reward margin $M(\mathbf{y}_w,\mathbf{y}_l,\mathbf{x},\theta)$ is

$$\beta\,(\log\frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w|\mathbf{x})} - \log\frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})}) \tag{5}$$

2.2 Related Work in Translation

Previous work has explored the effectiveness of enhancing the translation performance of LLMs via Reinforcement Learning (RL) algorithms or supervised fine-tuning. Dong et al. (2023) proposed RAFT that iteratively generates samples and fine-tunes the model on the filtered samples ranked by a reward model. Gulcehre et al. (2023) proposed ReST that uses similar method for translation task, where they apply several fine-tuning steps on a sampled dataset, each time higher ranked samples.

Similar to our pairwise preference learning, Zeng et al. (2023b) introduced a framework TIM to enhance the translation performance of LLMs by learning to compare good translations and bad translations via a preference learning loss.

Contemporaneous with this work, Finkelstein et al. (2023) proposed MBR fine-tuning, which fine-tunes an NMT model on the MBR decoding outputs generated either by the model itself or by an LLM. However, their MBR fine-tuning utilizes only the final translations of MBR decoding whereas our fine-tuning method uses sets of sampled translations ranked by MBR, thus enabling the model to learn the same ranking preferences as MBR.

3 Methodology

Our method combines MBR decoding and DPO fine-tuning (Yang, 2023). We use the MBR procedure to calculate a score (Equation 1) for each of a set of translation hypothesis generated by the base model. We then fine-tune the base model using the DPO objective (Equations 4,5) where the winning and losing hypotheses provided to DPO are chosen based on their relative MBR scores. If successful, the fine-tuned model will have learned to rank translations consistently with MBR decoding under the base model.

3.1 Creation of the DPO Preference Sets

Following Eikema and Aziz (2022), we use sampling to generate the translation hypotheses that will be used in DPO. For a source sentence $\mathbf{x}$ we use simple ancestral sampling with a temperature of 0.7 to create a set of translations $H(x) = \{\mathbf{y} \sim \pi_{base}(\mathbf{y}|\mathbf{x})\}$ of size $|H(x)|$. We use this collection as both the MBR evidence and hypothesis spaces (Goel and Byrne, 2000).

The hypotheses in $H(x)$ are ordered by their MBR scores as $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{|H|}$ with the BLEURT metric (Sellam et al., 2020) as the utility function. The ordering reflects the MBR preference, i.e. $\mathbf{y}_1$ would be the most preferred MBR hypothesis.

Preference Selection Strategies DPO requires a set of preference triplets $\mathcal{D} = \{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\}$ where $\mathbf{y}_w$ has better MBR score than $\mathbf{y}_l$ and both of the hypotheses are selected from the hypothesis set $H(x)$. There are numerous strategies for selecting the preference pairs $(\mathbf{y}_w,\mathbf{y}_l)$ from the hypothesis set. We experimented with four selection schemes:

1. **BW** is a simple strategy that selects the **best** and **worst** translation hypotheses from the ranked sets. For each source sentence $\mathbf{x}$, we only have one preference triplet $(\mathbf{x},\mathbf{y}_1,\mathbf{y}_{|H(x)|})$.

2. **BMW** adds the **middle** hypothesis $\mathbf{y}_m$ from the ranked lists with index $m = \lceil |H(x)|/2 \rceil$. This gives two triplets per source sentence: $(\mathbf{x},\mathbf{y}_1,\mathbf{y}_m)$ and $(\mathbf{x},\mathbf{y}_m,\mathbf{y}_{|H(x)|})$.

3. **CP** selects **consecutive pairs** from the ranked list, yielding $|H(x)| - 1$ triplets per source sentence, as $(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)$, $(\mathbf{x},\mathbf{y}_2,\mathbf{y}_3)$, ...

4. **CPS** introduces a **stride** into the CP selection strategy so as to avoid requiring DPO to learn distinctions between translations that are similarly ranked. For example, with a stride of 2 we select triplets $(\mathbf{x},\mathbf{y}_1,\mathbf{y}_3)$, $(\mathbf{x},\mathbf{y}_3,\mathbf{y}_5)$, ...

3.2 DPO Fine-Tuning

With a set of preference triplets $\mathcal{D}$ selected by one of the schemes above, DPO fine-tuning proceeds as described in Section 2.1 and by Rafailov et al. (2023). The base model serves as the reference model in Equation 4. The base model is also used to initialise π_θ, which is the model being fine-tuned. The only DPO hyper-parameter we tune is β, which regulates how the fine-tuned model departs from the reference model Rafailov et al. (2023).

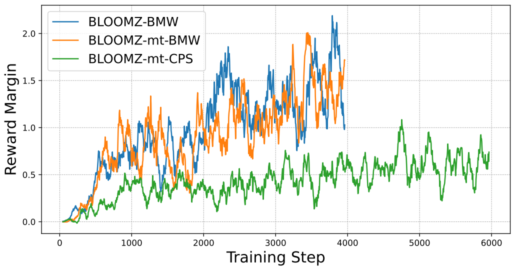

Figure 1: Reward margins for DPO MBR fine-tuning of BLOOMZ and BLOOMZ-mt with BMW and CPS (stride of 2) selection strategies. Margins are calculated on the Zh-En fine-tuning set (WMT20 test set) as fine-tuning proceeds over one epoch. Results are plotted as moving averages with a window size of 20. CPS yields more preference pairs than BMW.

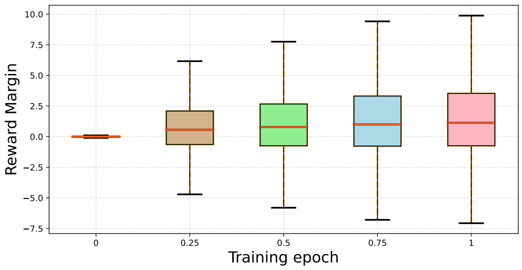

Figure 2: Reward margin distributions over all preference pairs extracted via the BMW scheme from a held-out dataset (WMT18 Zh-En test). Distributions are gathered over the entire held-out set at model checkpoints at the beginning, a quarter, middle, three quarters, and end of one epoch of DPO fine-tuning. $|H| = 8$ and $\beta = 0.7$. DPO fine-tuning generalises beyond its fine-tuning set and yields improved reward margins on held-out data.

4 DPO MBR Fine-Tuning and MT

Datasets: We evaluate translation on the WMT21 news translation test sets (Akhbardeh et al., 2021) and the WMT22 general translation for Chinese-English (Kocmi et al., 2022), and the IWSLT 2017 test set for French-English (Cettolo et al., 2017). For DPO fine-tuning we use the source language text in the WMT20 test sets for Chinese-English (Barrault et al., 2020) and IWSLT 2017 validation sets for French-English. We do not use the corresponding reference translations, as DPO MBR fine-tuning is unsupervised. The fine-tuning and test sets are distinct and do not overlap.

Models: We use the BLOOMZ and BLOOMZ-mt models (Muennighoff et al., 2023) with 7.1 billion parameters as our base model. BLOOMZ-mt was pre-trained on 366 billion tokens from monolingual texts and was fine-tuned for translation task on Flores-200 (NLLB Team et al., 2022) and Tatoeba (Tiedemann, 2020) datasets. To prompt the model for translation, we include two randomly selected translation examples from the fine-tuning set into the input prompt as demonstration examples; these prompts are kept fixed throughout. In addition, we also fine-tuned the BLOOMZ-mt model in a supervised fashion for each language pair and denote this third base model as BLOOMZ-mt-sft. We use previous WMT news translation test sets from 2017 to 2020 as supervised fine-tuning sets for Chinese-English, and the first 20000 translation pairs from the IWSLT 2017 training set for French-English. Training details can be found in Appendix B.

Evaluation Metrics: We use three evaluation metrics: BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), and COMET-22 (Rei et al., 2020, 2022). BLEU serves only as a safety check: Ideally DPO fine-tuning should not decrease BLEU.

Baselines and Targets: We take the base model and evaluate it on all the test sets with both beam search and MBR decoding. Our fine-tuned models, when decoded with beam search, should achieve similar performance as MBR decoding under the base model and show improvement over the base model. We investigate two questions:

(1) Can DPO teach MLLMs to learn their MBR translation preferences?

(2) Does preference learning with DPO lead to improved translation?

4.1 DPO Fine-Tuning Teaches a MLLM to Learn Its MBR Preferences

Figure 1 shows that the reward margins remain positive and, with some fluctuations, increase as fine-tuning proceeds, for all three models. This suggests that DPO MBR fine-tuned models learn to put more probability mass on the winning hypotheses. The larger the margins, the more the models prefer the winning over the losing hypotheses.

To further investigate DPO MBR fine-tuning, we plot the distribution of reward margins on a held-out set, shown in Figure 2. The median of the distributions increase consistently as fine-tuning proceeds, indicating that the MBR preferences learned in fine-tuning also generalize to unseen data.

#	Model (Decoding)	WMT21		WMT22		IWSLT17	
		zh-en	en-zh	zh-en	en-zh	fr-en	en-fr
1	BLOOMZ (Beam)	59.6 | 76.5	59.2 | 81.1	59.9 | 74.6	55.9 | 76.7	72.7 | 83.9	69.3 | 83.1
2	BLOOMZ (MBR $\lvert H\rvert = 8$)	60.0 | 76.4	62.5 | 82.3	62.1 | 75.8	62.7 | 80.0	73.6 | 84.2	70.4 | 83.3
3	BLOOMZ (MBR $\lvert H\rvert = 32$)	62.5 | 77.2	**64.7** | **83.0**	**64.0** | 76.4	**64.9** | 80.7	74.8 | 85.0	**72.6** | 84.3
4	**BLOOMZ-DPO-MBR** (Beam)	**62.3** | **77.9**	62.5 | 82.7	**64.0** | **77.2**	64.2 | **82.0**	**76.5** | **86.9**	72.2 | **84.8**
5	BLOOMZ-mt (Beam)	60.3 | 77.0	59.2 | 80.9	60.9 | 75.5	59.0 | 79.1	74.8 | 85.4	70.3 | 83.5
6	BLOOMZ-mt (MBR $\lvert H\rvert = 8$)	61.6 | 77.6	62.6 | 82.3	63.0 | 76.5	64.7 | 81.4	75.4 | 85.5	71.0 | 83.3
7	BLOOMZ-mt (MBR $\lvert H\rvert = 32$)	63.4 | 78.3	**64.9** | 82.9	64.8 | 77.2	66.8 | 82.1	76.3 | 86.0	**73.2** | 84.3
8	**BLOOMZ-mt-DPO-MBR** (Beam)	**63.9** | **78.7**	64.0 | **83.6**	**65.1** | **77.9**	**67.6** | **83.7**	**76.5** | **86.8**	71.9 | **84.6**
9	BLOOMZ-mt-sft (Beam)	64.3 | 79.4	62.6 | 83.0	62.6 | 76.5	65.6 | 83.1	76.9 | 86.6	71.2 | **83.8**
10	BLOOMZ-mt-sft (MBR $\lvert H\rvert = 8$)	65.3 | 79.8	64.8 | 83.9	65.4 | 78.2	69.1 | 84.2	77.3 | 86.7	72.6 | 83.6
11	BLOOMZ-mt-sft (MBR $\lvert H\rvert = 32$)	**66.8** | 80.4	**66.7** | **84.4**	**67.1** | 78.9	**71.0** | 85.1	**78.2** | **86.9**	**74.9** | 83.3
12	**BLOOMZ-mt-sft-DPO-MBR** (Beam)	66.0 | **80.8**	64.2 | 83.9	66.5 | **79.6**	69.5 | **85.6**	76.4 | 83.4	72.4 | **83.8**

Table 1: Translation performance in BLEURT and COMET (BLEURT | COMET) for models with beam search (beam width of 4) and MBR decoding on two language pairs from WMT21 news translation test sets, WMT22 general translation test sets, and IWSLT 2017 test sets. DPO-MBR indicates our translation performance with our fine-tuning method. All the DPO MBR models were fine-tuned using the BMW strategy and $\beta = 0.7$ except for BLOOMZ-mt-sft on IWSLT 2017, which used the BW strategy. We set $|H| = 32$ to fine-tune BLOOMZ-mt-DPO-MBR on English-Chinese direction, $|H| = 16$ on the French-English direction for BLOOMZ and BLOOMZ-mt, and set $|H| = 8$ to fine-tune other DPO MBR models. DPO-MBR improves both BLEURT and COMET whenever MBR itself improves substantially over the baseline.

4.2 DPO MBR Translation

Table 1 gives our main translation results. Comparing Rows 3 & 4, 7 & 8, and 11 & 12, we can see that DPO MBR fine-tuned models, when decoded with beam search, achieve similar performance in BLEURT and COMET as the base model decoded with MBR. The first two configurations (BLOOMZ-DPO-MBR and BLOOMZ-mt-DPO-MBR) outperform the base model's beam search results by ≈ 4 BLEURT and ≈ 2 COMET scores, and the third configuration outperforms the base mode by ≈ 3 BLEURT and ≈ 2 COMET on four out of six test sets. DPO MBR improves the translation ability of BLOOMZ, BLOOMZ-mt across a range of test sets. BLOOMZ-mt shows a notable improvement after DPO MBR fine-tuning, achieving the best performance in BLEURT on four out of six test sets and the best performance in COMET on all six test sets. We note that MBR decoding does not yield consistent improvement on the BLOOMZ-mt-sft model for IWSLT2017, and therefore does not provide a strong signal for DPO fine-tuning. We provide translation performance in BLEU in Appendix A for reference.

4.2.1 KL-Divergence Regularization

We investigated the role of β, the KL-divergence regularization factor, in DPO. Table 2 shows that fine-tuning with small β values yields high BLEURT score (exceeding 64), but also a degradation in BLEU and COMET. Anecdotally, we find that small values of β lead to repetitive outputs that are penalised heavily under BLEU and COMET. Gains in BLEU, BLEURT, and COMET are readily found, but we conclude that DPO MBR fine-tuning requires some care in regularization.

#	β	BLEU	BLEURT	COMET
1	(Baseline)	16.4	60.3	77.0
2	0.1	9.9	64.5	71.3
3	0.3	11.8	64.8	73.5
4	0.5	14.3	64.0	76.1
5	0.7	16.4	63.3	77.7
6	0.9	17.6	61.8	77.9

Table 2: Effect of regularization parameter β for DPO MBR fine-tuning of BLOOMZ-mt using CPS with $|H| = 8$. Models are fine-tuned on WMT20 zh-en and evaluated on WMT21 zh-en.

4.2.2 Effects of Pair Selection Strategy

Table 3 shows that models trained on preference datasets constructed with the BW, BMW, and CPS pair selection strategies achieve similar performance on WMT21 Zh-En, with BLEURT scores in the range 62.9-63.9. DPO MBR appears robust to the selection of preference pairs. In terms of training efficiency, the BW and BMW strategies require fewer preference pairs (1 and 2 per source sentence, resp.) compared to the CPS strategy. However, these results show that some selection strategy is

Selection Strategy	|H|=8	|H|=16	|H|=32
BW	63.3	63.9	63.9
BMW	63.9	64.2	63.6
CP	62.5	62.4	60.4
CPS (strides of 2, 4, and 8)	62.3	63.5	62.9

Table 3: WMT21 Zh-En BLEURT scores for BLOOMZ with DPO MBR fine-tuning with different preference pair selection strategies and hypothesis set sizes. The CP strategy results in lower performance in BLEURT compared to other strategies.

necessary since simply including all the pairs as in the CP strategy leads to degradation.

4.2.3 Effects of Size of Hypothesis Set

Table 3 shows that the number of hypotheses needed in the training preference dataset is less than that needed for MBR decoding (Rows 3 & 7 in Table 1). The best performance (BLEURT of 63.9) can be achieved with 16 hypotheses for the BW strategy and 8 hypotheses for the BMW strategy, an improvement over MBR decoding of the base model with $|H| = 8$ (Row 2 & 6 in Table 1).

5 Conclusion

We introduce DPO MBR fine-tuning, an unsupervised preference optimization algorithm that leverages the ranked lists from MBR decoding to teach MLLMs the preference of MBR decoding. Our method enables MLLMs to achieve significant performance improvement when decoded with beam search in one pass, on par with the performance gained from two-pass MBR decoding[1].

6 Acknowledgement

Jinghong Chen is supported by the Warwick Postgraduate Studentship from Christ's College and the Huawei Hisilicon Studentship for the undertaking of the PhD in Engineering at the University of Cambridge.

Weizhe Lin was supported by a Research Studentship funded by Toyota Motor Europe (RG92562(24020)).

Prof. Bill Byrne holds concurrent appointments as a Professor of Information Engineering at Cambridge University and as an Amazon Scholar. This publication describes work performed at Cambridge University and is not associated with Amazon.

We would also like to thank all the reviewers for their knowledgeable reviews.

7 Limitations

Our method was evaluated on WMT 2021 and WMT 2022 and IWSLT 2017 test sets, with high-resource languages only (English, Chinese, and French). While our fine-tuned models performed well on these diverse test sets, behaviour may be different on medium-resource or low-resource languages or on other domains.

Our experiments focus on BLOOMZ and BLOOMZ-mt due to the ease of working with them and because BLOOMZ-mt is fine-tuned for translation. Other (M)LLMs may yield different results.

We report MBR results using simple ancestral sampling. Other work (Freitag et al., 2023) has found that there may be advantages in using other sampling schemes, such as epsilon sampling, for MBR. Those other sampling methods potentially offer further gains beyond what we have already shown.

We do not report human assessments of translation quality to verify improvements, but we note that Freitag et al. (2022) have reported extensive results showing that MBR decoding under BLEURT leads to improvements in translation quality as assessed by human judges. We therefore take improvement in BLEURT as our main measurement of improved translation quality.

8 Risks

Our unsupervised fine-tuning technique could potentially amplify undesirable biases or language already present in the baseline systems. This could possibly happen if the MBR utility function, in our case BLEURT, somehow encourages consensus amongst similar translations that are also undesirable. Mitigation should be straightforward, in that any monitoring of the baseline models could also be applied after DPO MBR fine-tuning to reject fine-tuned models that exhibit any increase in bad behaviour. Although it is not a focus of this work, DPO MBR could possibly be used as a strategy for risk mitigation by penalizing undesirable behaviour through introduction of specific penalties into the MBR utility function.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatter-

[1] Codes are available at `https://github.com/BruceYg/DPO-MBR`

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

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

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: Reward ranked finetuning for generative foundation model alignment.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mara Finkelstein, 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.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation.

Markus Freitag, 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.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation.

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

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (ReST) for language modeling.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings* *of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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.

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. Implemented in SacreBLEU: `https://github.com/mjpost/sacrebleu`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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. 2022. 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, 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.

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. Implemented in `https://github.com/google-research/bleurt`.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Guangyu Yang. 2023. Multilingual models in neural machine translation. Dissertation submitted for the MPhil in Machine Learning and Machine Intelligence, University of Cambridge.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023a. GLM-130B: An open bilingual pre-trained model.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023b. TIM: Teaching large language models to translate with comparison.

#	Model (Decoding)	WMT21		WMT22		IWSLT17	
		zh-en	en-zh	zh-en	en-zh	fr-en	en-fr
1	BLOOMZ (Beam)	15.8	22.3	14.0	22.2	38.1	37.6
2	BLOOMZ (MBR $\|H\| = 8$)	11.3	19.7	11.6	20.3	34.2	32.6
3	BLOOMZ (MBR $\|H\| = 32$)	12.6	20.2	12.4	21.2	36.3	34.1
4	**BLOOMZ-DPO-MBR** (Beam)	**17.2**	**23.7**	**15.6**	**26.5**	**40.6**	**38.9**
5	BLOOMZ-mt (Beam)	16.4	22.5	14.7	26.2	38.7	37.8
6	BLOOMZ-mt (MBR $\|H\| = 8$)	13.5	20.2	12.2	23.3	35.2	31.8
7	BLOOMZ-mt (MBR $\|H\| = 32$)	14.3	20.8	13.0	24.0	36.9	33.8
8	**BLOOMZ-mt-DPO-MBR** (Beam)	**18.0**	**22.7**	**15.9**	**26.9**	**40.4**	**38.3**
9	BLOOMZ-mt-sft (Beam)	23.5	**27.5**	19.7	34.9	**44.2**	**40.7**
10	BLOOMZ-mt-sft (MBR $\|H\| = 8$)	20.2	24.0	17.7	30.1	40.7	34.2
11	BLOOMZ-mt-sft (MBR $\|H\| = 32$)	21.1	25.0	18.4	31.2	41.3	32.1
12	**BLOOMZ-mt-sft-DPO-MBR** (Beam)	**23.8**	26.3	**22.1**	**35.4**	27.3	38.5

Table 4: Translation performance in BLEU for models with beam search and MBR decoding on two language pairs from WMT21 news translation test sets, WMT22 general translation test sets, and IWSLT 2017 test sets. DPO-MBR indicates our translation performance with our fine-tuning method. All the DPO MBR models were fine-tuned using the BMW strategy and $\beta = 0.7$ except for BLOOMZ-mt-sft on IWSLT 2017, which used the BW strategy. We set $|H| = 32$ to fine-tune BLOOMZ-mt-DPO-MBR on English-Chinese direction, $|H| = 16$ on the French-English direction for BLOOMZ and BLOOMZ-mt, and set $|H| = 8$ to fine-tune other DPO MBR models.

A Translation Performance in BLEU

In Table 4, we provide the translation performance measured in BLEU score. The BLOOMZ-DPO-MBR and BLOOMZ-mt-DPO-MBR models achieve the best BLEU scores on all six test sets. The BLOOMZ-mt-sft model achieves lower BLEU score after DPO MBR fine-tuning on WMT21 English-to-Chinese, IWSLT17 French-to-English and English-to-French due to over-generation.

B Training Details

B.1 DPO MBR Fine-tuning Details

For DPO MBR fine-tuning, we trained each model for one epoch using the RMSProp optimizer. The learning rate is set to $5e^{-7}$ with 150 warmup steps. All fine-tuning experiments were done on two Nvidia A100-80G GPUs. We set the effective batch size to 4. We used FP32 and FP16 for the trained policy and the reference model in DPO fine-tuning, respectively.

B.2 Supervised Fine-tuning

We supervised fine-tuned the BLOOMZ-mt model on Chinese-to-English and English-to-Chinese directions for two epochs using previous WMT test sets. For French-to-English and English-to-French, we used the 20K translation pairs and trained for one epoch. Other hyper-parameters for SFT training are the same as DPO MBR fine-tuning.