Aligning Neural Machine Translation Models: Human Feedback in Training and Inference

Miguel Moura Ramos[1,2]   **Patrick Fernandes**[1,2,3]   **António Farinhas**[1,2]
André F. T. Martins[1,2,4]

[1]Instituto Superior Técnico, Universidade de Lisboa (ELLIS Unit Lisbon)
[2]Instituto de Telecomunicações   [3]Carnegie Mellon University   [4]Unbabel
`miguel.moura.ramos@tecnico.ulisboa.pt`

Abstract

Reinforcement learning from human feedback (RLHF) is a recent technique to improve the quality of the text generated by a language model, making it closer to what humans would generate. A core ingredient in RLHF's success in aligning and improving large language models (LLMs) is its *reward model*, trained using human feedback on model outputs. In machine translation (MT), where metrics trained from human annotations can readily be used as reward models, recent methods using *minimum Bayes risk* decoding and reranking have succeeded in improving the final quality of translation. In this study, we comprehensively explore and compare techniques for integrating quality metrics as reward models into the MT pipeline. This includes using the reward model for data filtering, during the training phase through RL, and at inference time by employing reranking techniques, and we assess the effects of combining these in a unified approach. Our experimental results, conducted across multiple translation tasks, underscore the crucial role of effective data filtering, based on estimated quality, in harnessing the full potential of RL in enhancing MT quality. Furthermore, our findings demonstrate the effectiveness of combining RL training with reranking techniques, showcasing substantial improvements in translation quality.

1 Introduction

Neural machine translation (NMT) models (Bahdanau et al., 2015; Vaswani et al., 2017) are typically trained with *maximum likelihood estimation* (MLE), maximizing the log-probability of the next word in a translation given the previous words and the source sentence. While this approach has been effective at training high-quality MT systems, the difference between the training and inference objective can lead to *exposure bias* (Bengio et al., 2015; Ranzato et al., 2016; Wiseman and Rush, 2016), which hinders the model's ability to recover from early mistakes. Furthermore, the suitability of model likelihood as a proxy for generation quality has been questioned in machine translation (Koehn and Knowles, 2017; Ott et al., 2018) and beyond (Perez et al., 2022). These challenges sparked interest in alternative training and decoding paradigms for MT, such as *reinforcement learning* (RL; Kreutzer et al. (2018)) or *minimum Bayes risk* decoding (MBR; Eikema and Aziz (2022)).

More recently, the widespread success of *reinforcement learning from human feedback* (Stiennon et al., 2022) has highlighted the importance of a good reward model that approximates well to human preferences for the task at hand. While, in general, this requires training a reward model from scratch for the specific problem, in the case of machine translation (MT), the evaluation community has achieved significant progress in developing automatic quality estimation and evaluation metrics *learned* from human quality annotations (e.g. COMET-QE (Rei et al., 2020), COMET (Rei et al., 2022a), BLEURT (Sellam et al., 2020), which can be repurposed as reward models. As a consequence, recent research integrating these metrics into the training (Gulcehre et al., 2023) or decoding (Fernandes et al., 2022) procedures has had considerable success in improving the quality of translations. However, none of the previous work has systematically compared the effect of integrating metrics at different stages of the MT pipeline or has attempted to combine these techniques in a unified approach.

In this work, we perform a comprehensive study

on the integration of MT quality metrics into the MT pipeline as reward models. As illustrated in Figure 1, we assess their use at different stages: as a means for data filtering, during the training process through RL, and at inference time by way of reranking techniques. Furthermore, we explore the results of combining these methods.

We attempt to answer the following research questions:

- *Can data filtering based on estimated quality help minimize RL training instability?*

- *Which metrics are more suitable as reward models in RL training? Are reference-free metrics competitive with reference-based ones?*

- *How does the quality of translations achieved through RL training compare with those produced through reranking approaches? Can these two approaches be effectively combined to further enhance translation quality?*

Our main contributions arise from the research questions mentioned above:

- Inspired by Bane and Zaretskaya (2021) where they use cross-lingual encoders to score translation representations in an aligned multilingual vector space, we propose an alternative data filtering method that uses COMET-QE (Rei et al., 2020), a more robust model, to curate a high-quality dataset that empirically helps to **minimize RL training instability**.

- We show that neural metrics such as COMET(-QE) (Rei et al., 2022a; Rei et al., 2020) are more suitable than BLEU (Papineni et al., 2002) for RL training. Contrary to what happens with MBR decoding, RL training results in improved scores across all types of metrics, not only neural ones. In particular, using a reward model based on QE works surprisingly well, possibly paving the way for unsupervised training of NMT systems.

- Experiments in EN→DE and EN→FR show that both RL training and reranking techniques enhance translation quality, with RL training often outperforming reranking methods. Furthermore, combining RL and MBR decoding results in more consistent improvements across various evaluation metrics.

- We quantify and discuss the trade-offs in running time at both training and inference, clarifying the efficiency and suitability of each approach.

2 Background

2.1 Neural Machine Translation

An NMT model has learnable parameters, θ, to estimate the probability distribution, $p_\theta(y|x)$ over a set of hypotheses $\mathcal{Y}$, conditioned on a source sentence x. MLE is the training principle of estimating θ, given parallel data, formalized as

$$\mathcal{L}(\theta, y_{1:L}) = -\frac{1}{L} \sum_{t=1}^{L} \log p_\theta(y_t|y_0, .., y_{t-1}). \quad (1)$$

NMT systems typically employ *maximum a posteriori* (MAP) decoding to generate translations,

$$\hat{y}_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} \log p_\theta(y|x), \quad (2)$$

where algorithms such as greedy decoding or beam search (Reddy, 1977) approximate the most *probable* translation given the source. An alternative approach is to sample translations according to $p_\theta(y|x)$, using techniques such as top-k or nucleus sampling (Fan et al., 2018; Holtzman et al., 2020).

In §3.3 of this paper, we also consider two distinct reranking approaches (Fernandes et al., 2022), namely N-best reranking and MBR decoding. While N-best reranking selects the candidate translation that maximizes a given (reference-free) metric, MBR decoding ranks candidates using reference-based metrics, maximizing the expected utility (or minimizing the risk).

2.2 MT Evaluation

Human evaluations are the most reliable way to assess the performance of MT systems, but they are time-consuming and costly. For that reason, the standard way to evaluate MT is through automatic evaluation metrics, which can be reference-based or quality estimation (QE) metrics.

Reference-based metrics compare the generated translation to human-written reference texts. Lexical reference-based metrics, such as the widely used BLEU (Papineni et al., 2002), rely on word overlap and n-gram matching, making them ineffective for translations that have the same meaning but are substantially different from the reference. On the other hand, neural metrics, such as COMET (Rei et al., 2022a), are a recent alternative that relies on neural networks trained on human-annotated
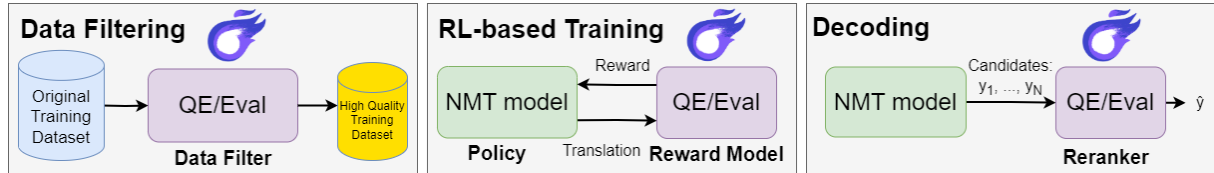
Figure 1: Preference models can have multifaceted roles within the MT pipeline. They can serve as effective data filters, refining datasets by incorporating user preferences. They can also assume a pivotal role in classic RL training by providing rewards to optimize the MT model performance. Finally, they can act as rerankers during the decoding phase, selecting the final translation by maximizing their scores derived from user preferences.

data and that leverages contextual embeddings to address semantic similarity.

QE assesses translation quality without human references, being particularly useful in dynamic, data-intensive environments, where references are costly and time-consuming to obtain. This paper focuses on sentence-level QE as a reward model, providing a single quality assessment for each translation. COMET-QE (Rei et al., 2020) is a state-of-the-art reference-free quality estimation metric derived from COMET used to evaluate MT performance.

Neural reference-based and QE metrics are valuable preference models because they offer a more accurate and contextually-aware measure of translation quality, aligning better with human preferences and judgments (Freitag et al., 2022b).

2.3 Reinforcement Learning Training in NMT

In MT, approaches based on reinforcement learning (RL; Sutton and Barto (2018)) cast the problem as a Markov decision process (MDP; Puterman (1990)), where a source sentence $x = (x_1, ..., x_n)$ is translated into a target sentence $y = (y_1, ..., y_m)$. Under this perspective, the NMT system can be viewed as the agent with a conditional probability distribution based on its parameters, $p_\theta(y_t|x, y_{<t})$. The states of the MDP are defined by the target sentence that has already been decoded, $s_t = (y_1, ..., y_{t<m})$, and the action corresponds to the selection of the next word, y_{t+1}. Based on the states and actions, all transitions are deterministic and the reward function, R, is provided by the MT evaluation model which returns a quality score for the generated translation $\hat{y}$. The main purpose of using RL in NMT is to provide learning signals that go beyond a single reference translation, by providing reward signals for arbitrary translations. MLE provides less robust learning signals that are more susceptible to the shortcomings of noisy references. However, it

is essential to note that if the reward model used relies on reference-based metrics, some vulnerability to noisy references may still persist. Accordingly, the goal of RL training is to maximize the expected reward, $L_{rl}(\theta) = \mathbb{E}_{p_\theta(\hat{y}|x)}[R(\hat{y})]$. Commonly used RL training procedures include REINFORCE (Williams, 1992), minimum risk training (Och, 2003; Shen et al., 2016), and proximal policy optimization (PPO; Schulman et al. (2017)).

3 Aligning MT with Reward Models

3.1 Data Filtering

The success of fine-tuning NMT models with MLE is highly dependent on the quantity and quality of the training dataset (Wang et al., 2018; Koehn and Knowles, 2017; Khayrallah and Koehn, 2018). This is because accurate references are crucial for computing meaningful learning signals that correctly guide the NMT model towards improved translations (Kong et al., 2018). Despite its recent successes, RL-based training can be unstable, so using only high-quality data could help mitigate this instability. This can be addressed via **data filtering**, by seeking a good balance between the aggressiveness of filtering and the resulting dataset size: if the original dataset is already small, too much filtering can be detrimental to the performance of NMT systems (Zoph et al., 2016; Jiao et al., 2020). Furthermore, when looking at the RL scenario, having a sufficiently large training dataset can help guarantee that the NMT model explores a wide range of scenarios for policy improvement.

We apply our data filtering method on the considerably large and noisy WMT datasets (Bojar et al., 2015; Bojar et al., 2016) since they have been reported to have less relevant and uncorrelated sentences that can lead to sub-optimal results when used during training (Koehn et al., 2020; Malli and Tambouratzis, 2022). We do not perform data filtering to the IWSLT2017 (Cettolo et al., 2012;

Cettolo et al., 2017) dataset due to concerns about its limited amount of available data. Further dataset filtering could potentially result in a too-small training dataset, which is not be desirable for training MT systems.

As illustrated in Figure 1, to perform the training dataset filtering, we use a filter that reranks the sentence pairs according to quality scores that indicate the correlation and relevance of each sentence and its given reference. This approach allows us to filter out low-quality sentence pairs, thereby improving the overall quality of the data. In our approach, we use a robust preference model called COMET-QE (Rei et al., 2020) as the data filter, which combines the use of encoders and a regression model trained on human-annotated data to estimate the quality score of each sentence pair. This reference-less model is expected to be more accurate in quality score estimation and have a superior alignment with human judgments than just resorting to the currently used cross-lingual encoders which only take into account vector-space mapping similarity (Bane and Zaretskaya, 2021). Furthermore, COMET-QE seems particularly suitable as our preference model during data filtering, as it is a multilingual reference-free neural-based metric trained on human annotations of translation quality, and therefore can be used to filter by thresholding on predicted quality or on the number of sentences in the training set. After scoring all sentence pairs, we select the threshold based on the number of high-quality sentence pairs to use as the filtered dataset for RL training. For that, we apply different thresholds and sizes to the reranked sentences. We, then, MLE fine-tune our baseline on these subsets and select the subset that gives the overall best-performing model on the dev. set. These best-performing models serve as baselines for our RL-based training and reranking methods during decoding.

In conclusion, it is worth noting that our data filtering method is, as shown in Figure 1, one of three methods we cover for employing a preference model in the MT pipeline. This filtering method can significantly increase the performance of MT systems by introducing feedback in an earlier stage of the pipeline.

3.2 Training Phase

The use of RL-based training has the potential to bridge the gap between MLE training objectives, MT evaluation metrics and human-like translations.

However, it faces challenges of instability and inefficiency, especially in gradient variance and reward computation. As illustrated in Figure 1, the RL training process is composed of an NMT model that generates translations that are evaluated by the reward model through rewards that represent the quality of the translation. This reward is used by the policy gradient algorithm to update the NMT model's **policy**. To address the problem of gradient variance, we employ PPO (Schulman et al., 2017) as our policy gradient algorithm since it is a stable and efficient algorithm that updates the policy parameters in a controlled way with a predetermined proximity bound, avoiding sudden changes that might destabilize the learning.

Reward computation is the most crucial part of this entire process as it guides the NMT model during training. Previous work on RL-based NMT systems predominantly used BLEU as the reward function. However, BLEU has several limitations, as discussed in §2.2. To address these shortcomings, we leverage robust preference models during RL training, such as the reference-based COMET (Rei et al., 2022a) and the reference-free COMET-QE (Rei et al., 2020), as highlighted in Figure 1. Since learning these models is a complex task, we incorporate these pre-trained preference models, which have already been shown to correlate well with human judgments (Freitag et al., 2022b; Rei et al., 2022a; Rei et al., 2020), to ensure that RL systems can better capture the nuanced preferences of the user by receiving human-like feedback as rewards. These models assign numerical quality scores to each translation hypothesis based on their desirability, making them similar to utility functions. Our study aims to demonstrate that training with RL can generate higher-quality NMT models using neural metrics and investigate the competitiveness of COMET-QE as a reward model.

Another crucial decision was related to the exploitation vs. exploration problem of RL in the context of MT (Wu et al., 2018). The beam search algorithm generates more accurate translations by exploiting the probability distribution/policy of the NMT model, while sampling aims to explore more diverse candidates. During generation, we observed that sampling techniques generally led to candidates of lower quality when compared to beam search, according to the preference models used. Therefore, all RL-based models used beam search during their training and inference.

3.3 Decoding Phase

Reranking methods (Ng et al., 2019; Bhattacharyya et al., 2021; Fernandes et al., 2022; Eikema and Aziz, 2022) are an alternative to MAP-based decoding that relies on reranking techniques and presupposes access to N candidate translations for each source sentence, generated by the NMT system through methods like beam search or sampling. The generated candidates are reranked according to their quality given an already determined metric/reward model.

We employ two reranking methods to select a final translation: N-best reranking (Ng et al., 2019; Bhattacharyya et al., 2021) and *minimum Bayes risk* decoding (MBR; Eikema and Aziz (2022)).

N-best reranking (3) employs a reference-free metric, M_{QE}, to reorder a set of N candidate translations, denoted as $\bar{\mathcal{Y}}$, and selects the candidate with the highest estimated quality score as the final translation, $\hat{y}_{\text{RR}}$,

$$\hat{y}_{\text{RR}} = \arg\max_{y \in \bar{\mathcal{Y}}} M_{\text{QE}}(y). \tag{3}$$

Considering the previous equation, and assuming $C_{M_{\text{QE}}}$ as the computational cost of evaluating a candidate translation with QE metric, M_{QE}, we obtain the final computational cost of finding the best translation from N candidate translations as $O(N \times C_{M_{\text{QE}}})$.

MBR decoding, in contrast, relies on a reference-based metric and chooses the candidate that has the highest quality when compared to other possible translations (in expectation). We define $u(y^*, y)$ as the utility function, quantifying the similarity between a hypothesis $y \in \mathcal{Y}$ and a reference $y^* \in \bar{\mathcal{Y}}$. In our context, the utility function is represented by either BLEU or COMET. Therefore, MBR decoding can be mathematically expressed as

$$\hat{y}_{\text{MBR}} = \arg\max_{y \in \bar{\mathcal{Y}}} \underbrace{\mathbb{E}_{Y \sim p_\theta(y|x)}[u(Y, y)]}_{\approx \frac{1}{N} \sum_{j=1}^{N} u(y^{(j)}, y)}, \tag{4}$$

where in Eq. 4 the expectation is approximated as a Monte Carlo sum using model samples $y^{(1)}, \ldots, y^{(N)} \sim p_\theta(y|x)$. These samples may be obtained through biased sampling (e.g., nucleus-p or top-k) or beam search. Knowing that the utility function is a reference-based metric M_{REF} with computational cost, $C_{M_{\text{REF}}}$, and that to find the best translation we need to do pairwise comparisons between hypotheses, we obtain the final computational cost as $O(N^2 \times C_{M_{\text{REF}}})$. These reranking

methods become particularly effective when N is not excessively large, making the process computationally more manageable.

Preference models capture the preferences of human evaluators and can be used during the decoding stage to influence MT systems, as shown in Figure 1. By doing this, the MT system will prioritize translations that are more aligned with human judgments, therefore reducing the chances of generating severely incorrect translations. We believe that incorporating preference models during the decoding stage can lead to even better translation quality, even if the underlying model has already been RL-trained using the same or a different preference model. The benefits we expect to see include improved fluency, adequacy, and consistency compared to the respective baselines since our preference models have been trained on annotations that aim to optimize these linguistic aspects.

4 Experiments

4.1 Setup

During the training phase, we investigate the advantages of RL training (with and without data filtering §3.1) for enhancing the performance of NMT systems. We employ a T5 model[1], pre-trained on the C4 dataset (Raffel et al., 2019). First, we fine-tune the models using MLE training with Adam (Kingma and Ba, 2017) as the optimization algorithm, learning rate decay starting from 5×10^{-6} and early stopping. For RL training[2], we use PPO with learning rate set as 2×10^{-5}, γ set as 0.99, trajectory limit set as $10,000$, beam search size set as 5 and mini-batch updates were conducted using stochastic gradient descent with a batch size of 32, gathered over 4 PPO epochs. In the inference phase, our emphasis shifts towards reranking techniques and their impact on the performance of NMT systems. As for the candidate generation method used, early experiments, omitted for relevancy, show that the best configuration is to generate 100 candidates per source sentence and then use sampling with $p = 0.6$ and $k = 300$ to select the best translation. Consequently, the evaluation encompasses all the baseline and RL-trained models, both with and without N-best reranking and MBR decoding. These evaluations are conducted across the follow-

---

[1] We leverage the T5-Large model available in Huggingface's *Transformers* framework (Wolf et al., 2020).
[2] Our RL implementation relies on the Transformer Reinforcement Learning X framework (Castricato et al., 2023, trlX).

ing datasets:

- The small IWSLT2017 datasets (Cettolo et al., 2012; Cettolo et al., 2017) for English to German (EN → DE) and English to French (EN → FR), featuring 215k and 242k training examples, respectively.

- The large WMT16 dataset (Bojar et al., 2016) for English to German (EN → DE) with 4.5M training examples.

- The large WMT15 dataset (Bojar et al., 2015) for English to French (EN → FR) with over 40M training samples.

We assess the performance of each NMT system using well-established evaluation metrics, which include BLEU, chrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005), COMET, COMET-QE, and BLEURT. Additionally, for certain experiments executed on a single NVIDIA RTX A6000 GPU, we provide wall clock time measurements to offer insights into computational efficiency.

4.2 Finding the Optimal Quality Subset Size

In this section, we discuss our approach to quality-aware data filtering as a stabilizing strategy (§3.1), for the WMT datasets. Figure 2a summarizes our findings for the WMT16 EN→DE dataset (Bojar et al., 2016) on the influence of a high-quality subset on translation performance as we vary the subset size, based on various evaluation metrics and COMET-QE sentence filtering. Across all metrics, a consistent trend emerges: after reaching training sizes of 500 000, there is a notable decline in performance. Particularly, this decline is less prominent for lexical metrics, possibly due to their inherent limitations (Freitag et al., 2022b). A similar analysis for WMT15 EN→FR that can be found in Figure 2b results in an optimal training size of 300 000 examples.

While the data filtering process has led to remarkable improvements in performance, it is important to note that the effectiveness of this process is dependent on the selected reranking metric. Using metrics that are not closely aligned with human judgments can result in poorly correlated and misaligned sentences, which can make the training process more unstable. Therefore, it is recommended to use robust QE models, such as COMET-QE. The more recent COMETKIWI (Rei et al., 2022b) model may offer even greater performance improvements.

4.3 Impact of Quality-aware Data Filtering

After obtaining the best configuration for our data filtering process, we experiment with the use of the curated high-quality training subset from COMET-QE and assess its impact on the MLE and RL training performance. We compare our filtering method with no filtering by using the original full training dataset, random filtering and cross-lingual embedding similarity filtering using MUSE (Lample et al., 2017) and XLM-R (Conneau et al., 2019).

Table 1 provides a comprehensive overview of the experimental results using BLEU, COMET and COMET-QE as reward models. Both MT tasks demonstrate the same tendency when trained using MLE. COMET-QE and MUSE high-quality subsets have enough reduced noise to provide more stable training, as evidenced by the overall increase in performance across all metrics compared to the baseline training on the full original dataset. Moreover, a randomly selected subset fine-tuned with MLE performs worse or at most not significantly better than the baseline trained on the original dataset, as expected. Furthermore, in accordance with our expectations (Bane and Zaretskaya, 2021), XLM-R filtering does not improve training and is actually the worst-performing model.

Regarding RL-based training on both MT tasks, we observe that most RL-trained models outperform their MLE-trained baseline counterparts across various metrics. Notably, the best-performing models are the ones that were MLE fine-tuned and then RL-trained on the COMET-QE high-quality subset using both COMET and COMET-QE as reward models. On top of that, we can see that in some cases RL training solely does not yield significant improvements, but when combined with high-quality training subsets, it results in substantial enhancements and a competitive edge over the normal, random and XLM-R baselines. Additionally, we see impressive BLEU scores with RL training with COMET(-QE) as reward model. This finding underscores that optimizing for COMET(-QE) yields superior BLEU scores compared to direct optimization for BLEU. This phenomenon is likely attributed to COMET(-QE) providing **more effective** reward signals during training, thus highlighting the limitations of BLEU.

The excellent performance gains with COMET-QE as a data filter and also as a reward model emphasize the potential of RL-based NMT models trained with a QE reward model (which *does not* re-

(a) Impact of Data Filtering on WMT16 En→De
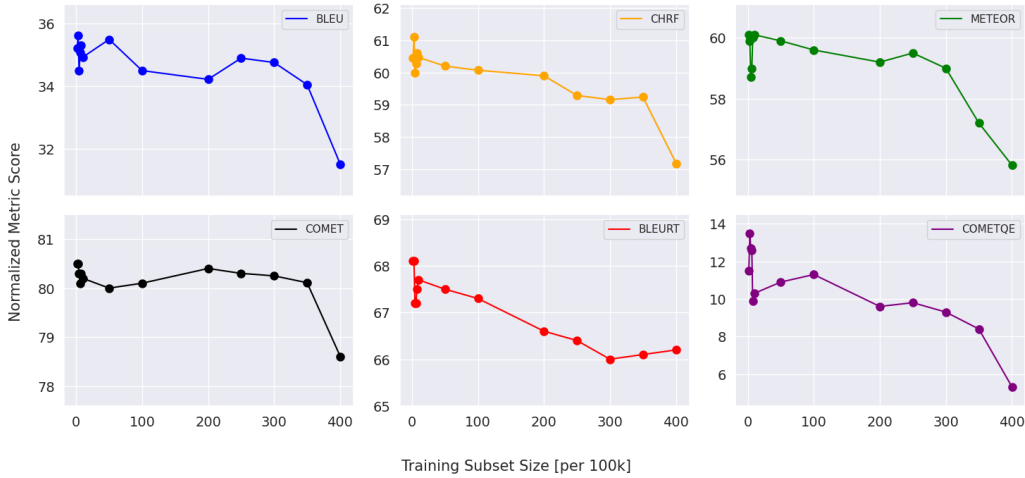
(b) Impact of Data Filtering on WMT15 En→FR

Figure 2: These models were fine-tuned by progressively increasing the size of the high-quality subset, obtained with COMET-QE sentence reranking and denoted in increments of 100,000.

quire a corpus with references) to outperform other RL-trained models, offering promising opportunities for unsupervised NMT training with monolingual data, especially for low-resource languages, by eliminating the need for reference translations in evaluation and reward signal generation.

In conclusion, we highlight the importance of thoughtful data selection for achieving better translation quality, showing that COMET-QE can consistently outperform the remaining filtering methods. Furthermore, the top-performing models were RL-trained with **neural** metrics, showing once again that **human-aligned** preference models can constantly outperform simpler metrics, such as BLEU.

4.4 Impact of preference-based MT alignment

Table 2 presents the performance scores of the best baseline model, across various MT tasks, focusing on the comparison between RL training, reranking

methods during inference and the potential synergies between RL training and reranking techniques in improving the translation quality of MT systems.

Our analysis reveals consistent improvements across all evaluation metrics and reward models, with RL training consistently achieving top scores, especially when using COMET-QE as the reward model. [3] MBR decoding with COMET and *N*-best reranking with COMET-QE outperformed RL training in COMET and COMET-QE metrics but had difficulty improving other evaluation metrics, while RL training exhibited better generalization with slightly less consistent improvements in COMET and COMET-QE scores. This phenomenon of increased COMET and COMET-QE scores comes at the cost of worse performance according to the

---

[3]We also provide additional fine-grained quality analysis in Appendix A to better illustrate and address specific research questions.

Training Data		Lexical Metrics			Neural Metrics		
SL Data	RL Data	BLEU	ChrF	METEOR	COMET	COMET-QE	BLEURT
MLE							
Original	-	35.04	61.30	61.91	84.40	39.50	74.70
Random	-	34.43	61.00	61.36	83.90	39.10	74.30
XLM-R	-	33.24	60.35	60.20	84.80	41.80	72.60
MUSE	-	35.10	61.90	62.20	85.10	40.40	74.30
COMET-QE	-	<u>35.45</u>	<u>62.00</u>	<u>62.75</u>	<u>85.50</u>	<u>42.00</u>	<u>75.90</u>
RL w/ BLEU							
Original	Original	34.70	60.90	61.45	85.60	42.20	74.60
Random	Random	34.49	61.10	61.49	85.60	42.20	74.40
XLM-R	XLM-R	33.21	60.41	60.10	85.10	42.70	73.10
MUSE	MUSE	35.34	62.10	62.73	85.60	40.80	74.50
Original	COMET-QE	35.37	61.70	62.04	85.40	41.00	74.20
COMET-QE	COMET-QE	<u>35.55</u>	<u>62.10</u>	<u>62.77</u>	**86.80**	**45.00**	**76.10**
RL w/ COMET							
Original	Original	35.05	61.30	61.82	85.60	41.80	74.40
Random	Random	34.96	61.40	61.80	85.60	41.80	74.20
XLM-R	XLM-R	33.60	60.74	60.40	85.00	42.00	72.90
MUSE	MUSE	35.18	61.90	62.56	85.50	41.90	74.60
Original	COMET-QE	35.58	61.80	62.20	85.70	41.70	74.50
COMET-QE	COMET-QE	<u>35.90</u>	<u>62.20</u>	<u>63.06</u>	<u>86.70</u>	<u>44.10</u>	<u>75.70</u>
RL w/ COMET-QE							
Original	Original	34.21	60.50	61.10	85.60	42.40	74.80
Random	Random	34.88	61.30	61.69	85.50	41.80	74.10
XLM-R	XLM-R	33.57	60.73	60.40	85.10	42.20	73.20
MUSE	MUSE	35.03	61.90	62.57	85.70	41.30	74.70
Original	COMET-QE	35.48	61.70	62.10	85.70	41.70	74.50
COMET-QE	COMET-QE	**<u>35.96</u>**	**<u>62.30</u>**	**<u>63.07</u>**	<u>86.70</u>	<u>44.70</u>	<u>75.90</u>
Training Data		Lexical Metrics			Neural Metrics		
SL Data	RL Data	BLEU	ChrF	METEOR	COMET	COMET-QE	BLEURT
MLE							
Original	-	31.49	57.18	55.80	78.60	5.30	66.20
Random	-	31.27	57.07	60.01	80.00	12.80	65.20
XLM-R	-	25.04	48.78	48.60	77.40	12.10	57.10
MUSE	-	35.49	59.10	60.55	80.10	13.10	67.50
COMET-QE	-	<u>35.62</u>	<u>59.90</u>	<u>61.11</u>	<u>80.50</u>	<u>13.50</u>	<u>68.10</u>
RL w/ BLEU							
Original	Original	35.47	59.90	61.03	80.20	16.90	67.10
Random	Random	32.75	58.10	60.20	80.03	14.10	66.35
XLM-R	XLM-R	25.78	49.69	49.30	77.70	13.30	57.80
MUSE	MUSE	35.55	60.10	60.56	81.90	17.10	67.50
Original	COMET-QE	35.67	60.10	61.01	81.20	17.10	67.30
COMET-QE	COMET-QE	<u>36.26</u>	<u>60.40</u>	<u>61.51</u>	<u>82.10</u>	<u>17.50</u>	<u>67.70</u>
RL w/ COMET							
Original	Original	35.50	59.90	61.00	80.40	16.80	67.00
Random	Random	34.15	59.50	60.93	80.50	15.50	67.10
XLM-R	XLM-R	25.08	48.84	48.60	77.50	12.40	57.20
MUSE	MUSE	36.00	60.10	61.20	80.80	17.00	67.30
Original	COMET-QE	35.98	60.00	61.09	81.80	17.10	67.20
COMET-QE	COMET-QE	**<u>36.62</u>**	**<u>60.60</u>**	**<u>61.79</u>**	<u>82.20</u>	<u>17.40</u>	<u>67.60</u>
RL w/ COMET-QE							
Original	Original	35.50	60.00	61.10	82.20	17.50	68.00
Random	Random	32.10	58.30	60.50	81.00	14.40	66.70
XLM-R	XLM-R	24.67	48.38	48.10	77.60	12.60	56.80
MUSE	MUSE	35.62	60.45	59.30	82.22	17.45	67.80
Original	COMET-QE	35.90	60.10	61.22	82.27	17.53	68.02
COMET-QE	COMET-QE	<u>36.25</u>	<u>60.50</u>	<u>61.58</u>	**82.40**	**17.70**	**68.10**

Table 1: Automatic evaluation metrics for the MLE and RL-trained models on the WMT16 EN→DE (top) and WMT15 EN-FR (bottom) original datasets, quality subsets obtained from COMET-QE, XLM-R and MUSE and a randomly selected subset. The training data used for MLE and RL can be found in the SL and RL Data, respectively. We experimented with BLEU, COMET and COMET-QE as reward models for the RL training. The best overall values are **bolded** and the best for each specific group are <u>underlined</u>.

other MT evaluation metrics, showing a potential of **overfitting** effect for these reranking techniques that occur across all datasets. These findings underscore the potential of neural metrics as reward signals in training and inference, as discussed in Deutsch et al. (2022) and Freitag et al. (2022b). While combining RL training and MBR decoding occasionally led to top performance, it did not consistently outperform other strategies, making it

a method that distributes gains across all evaluation metrics without exceptional generalization as RL training but provides better overall scores than reranking methods alone.

RL training and MBR decoding in MT exhibit distinct computational efficiency profiles, as shown in Table 3. RL training is computationally demanding but typically entails a one-time, resource-intensive training process (though less resource-

MODEL	WMT16 EN→DE						WMT15 EN→FR					
	BLEU	METEOR	ChrF	COMET	COMET-QE	BLEURT	BLEU	METEOR	ChrF	COMET	COMET-QE	BLEURT
High-Quality Subset Baseline (HQSB)	35.45	62.00	62.75	85.50	42.00	75.90	35.62	59.90	61.11	80.50	13.50	68.10
BLEU												
HQSB + RL	<u>35.55</u>	62.10	62.77	<u>86.80</u>	<u>45.00</u>	<u>76.10</u>	36.26	60.40	61.51	<u>82.10</u>	<u>17.50</u>	<u>67.70</u>
HQSB + MBR	35.53	**62.30**	<u>62.80</u>	86.70	44.20	75.90	35.73	60.40	61.42	81.60	15.60	67.20
HQSB + RL + MBR	35.22	61.90	62.62	86.20	43.10	75.50	**36.72**	**60.80**	**61.89**	82.00	16.30	67.20
COMET												
HQSB + RL	<u>35.90</u>	<u>62.20</u>	<u>63.06</u>	86.70	44.10	75.70	<u>36.62</u>	<u>60.60</u>	<u>61.79</u>	82.20	17.40	67.60
HQSB + MBR	33.58	60.70	61.48	88.00	<u>47.90</u>	76.50	34.89	59.60	60.94	**85.00**	<u>27.00</u>	<u>69.80</u>
HQSB+ RL + MBR	34.92	61.80	62.84	<u>88.10</u>	47.60	<u>76.90</u>	35.97	60.20	61.45	84.40	24.50	69.20
COMET-QE												
HQSB + RL	**35.96**	**62.30**	**63.07**	86.70	44.70	75.90	<u>36.25</u>	<u>60.50</u>	<u>61.58</u>	82.40	17.70	68.10
HQSB + N-RR	31.46	58.70	60.41	87.10	<u>53.80</u>	75.90	29.99	54.80	56.87	82.80	**39.10**	66.20
HQSB + RL + N-RR	32.73	59.80	61.32	87.30	53.20	76.30	32.61	57.40	58.96	83.40	36.10	67.60
HQSB + N-RR + MBR w/ COMET	33.73	60.90	61.79	88.10	49.60	76.70	34.34	59.40	60.69	84.80	29.40	69.50
HQSB + RL + MBR w/ COMET	34.61	61.60	62.72	**88.20**	50.10	**77.20**	35.47	59.90	61.26	<u>84.90</u>	28.80	**70.00**

MODEL	IWSLT2017 EN→DE						IWSLT2017 EN→FR					
	BLEU	METEOR	ChrF	COMET	COMET-QE	BLEURT	BLEU	METEOR	ChrF	COMET	COMET-QE	BLEURT
Normal Baseline (NB)	32.75	62.40	60.04	84.80	38.30	74.80	41.47	68.40	66.20	84.40	21.70	73.30
BLEU												
NB + RL	<u>34.48</u>	**62.90**	**60.51**	85.20	<u>39.70</u>	74.40	**44.58**	68.60	<u>66.76</u>	85.20	<u>24.70</u>	72.70
NB + MBR	33.87	62.20	60.05	85.00	38.90	<u>74.50</u>	44.08	**68.70**	66.52	85.20	24.40	<u>73.20</u>
NB + RL + MBR	34.46	62.50	60.22	85.00	39.00	74.10	44.25	68.30	66.50	85.00	24.20	72.40
COMET												
NB + RL	<u>34.17</u>	62.20	59.88	85.10	39.30	74.40	<u>44.48</u>	**68.70**	66.74	85.20	24.60	72.80
NB + MBR	33.33	62.10	<u>59.97</u>	**86.70**	<u>43.80</u>	<u>75.60</u>	39.04	65.30	63.32	86.80	<u>37.40</u>	<u>75.00</u>
NB + RL + MBR MBR	33.75	61.90	59.72	86.10	41.80	74.90	44.24	68.50	66.62	86.30	28.30	73.60
COMET-QE												
NB + RL	**34.53**	**62.90**	<u>60.49</u>	85.30	40.00	74.70	<u>44.56</u>	**68.70**	**66.87**	85.30	24.90	72.90
NB + N-RR	32.31	60.70	59.06	86.40	**50.00**	75.60	42.48	67.20	65.38	86.60	38.30	74.00
NB + RL + N-RR	32.98	61.50	59.48	86.40	48.70	75.40	43.29	67.50	65.90	86.50	36.00	73.70
NB + N-RR + MBR w/ COMET	33.53	61.90	59.95	**86.70**	46.00	**75.80**	39.41	65.40	63.42	**87.00**	**40.00**	**75.30**
NB + RL + MBR w/ COMET	34.18	62.50	60.27	86.60	43.50	75.40	44.07	68.20	66.55	86.70	32.50	74.00

Table 2: Automatic evaluation metrics for the best baseline in each dataset and its variations with RL training, reranking (N-RR) and MBR decoding. BLEU, COMET, and COMET-QE serve as reward models in the context of RL training and are subjected to comparison with respect to both reranking strategies employed as the optimization metric (reranker). Best-performing values are **bolded** and best for each specific group are <u>underlined</u>.

intensive than MLE training), involving iterative fine-tuning of NMT models, making it suitable for capturing nuanced quality improvements from the reward models. In contrast, MBR decoding, focused on optimizing translation during inference, requires recomputation for each input sentence, allowing for computational efficiency when performed infrequently. However, it may not fully utilize the capabilities of the NMT model and can be computationally demanding in high-throughput scenarios. The choice between RL training and MBR decoding depends on specific MT system requirements, considering computational resources, translation quality objectives, and the need for real-time adaptability.

In summary, the results demonstrate that integrating RL training consistently improves translation quality in both EN→DE and EN→FR tasks across various metrics. It consistently outperforms the MLE baseline and is superior in lexical metrics scores compared to reranking strategies, which perform well according to COMET and COMET-QE.

Additionally, most top-performing models incorporate RL training, highlighting its effectiveness in complementing reranking strategies to further improve translation quality.

5 Related Work

RL-based NMT. Extensive research has been conducted on RL algorithms to improve MT. Studies by Wu et al. (2018) and Kiegeland and Kreutzer (2021) have explored the impact of RL training on large-scale translation tasks and demonstrated the effectiveness of policy gradient algorithms in mitigating exposure bias and optimizing beam search in NMT. However, both studies were limited to the use of BLEU as a reward model. Our research differs in that we explore the benefits of employing more robust preference models to improve translation quality. Additionally, other researchers have made progress in advancing reward-aware training methods. For instance, Donato et al. (2022) introduced a distributed policy gradient algorithm using mean

	WMT16 EN→DE		WMT15 EN→FR		IWSLT2017 EN→DE		IWSLT2017 EN→FR	
Method	Training	Inference	Training	Inference	Training	Inference	Training	Inference
MLE	480	5	373	3	1020	13	905	16
RL	288	5	242	3	354	13	403	16
MBR	0	212	0	55	0	500	0	660
N-RR	0	183	0	50	0	455	0	625

Table 3: Wall-clock time values, in minutes, that represent the efficiency of MLE, RL, MBR decoding and N-best reranking. The training was performed on the WMT16 EN→DE and WMT15 EN→FR high-quality subsets and on IWSLT2017 EN→DE and EN→FR entire datasets with 500 000, 300 000, 215 000 and 242 000 sentence pairs, respectively. The inference was conducted on WMT16 EN→DE, WMT15 EN→FR, IWSLT2017 EN→DE and IWSLT2017 EN→FR official test set partitions with 2999, 1500, 8079 and 8597 sentence pairs, respectively. This assessment was done with COMET as the reward model for RL and as a reranker for the reranking methods.

absolute deviation (MAD) for improved training, excelling with BLEU rewards and generalizing well to other metrics. Moreover, Ouyang et al. (2022) pioneered reinforcement learning from human feedback (RLHF) for a human-based reward model, while Gulcehre et al. (2023) proposed Reinforced Self-Training (ReST) for more efficient translation quality improvement using offline RL algorithms.

Reranking methods for NMT. Shen et al. (2004) initially introduced the concept of discriminative reranking for Statistical Machine Translation, which was later adopted by Lee et al. (2021) to train a NMT model through a reranking strategy based on BLEU. Extending this concept, MBR decoding (Kumar and Byrne, 2004) has regained popularity for candidate generation during decoding, with Müller and Sennrich (2021) finding it more robust than MAP decoding, mitigating issues like hallucinations. Furthermore, Freitag et al. (2022a) showed that coupling MBR with BLEURT, a neural metric, enhances human evaluation results when compared to lexical metrics. Fernandes et al. (2022) conducted a comprehensive study comparing various reranking strategies, including reranking and MBR decoding, with both reference-based and quality estimation metrics, concluding that these strategies lead to better translations despite the increased computational cost. In our work, we build on these foundations and show that reranking methods can be coupled with RL training to provide better translation quality to MT systems.

Data filtering for NMT. In their study, Taghipour et al. (2011) explored the use of outlier detection techniques to refine parallel corpora for MT. Meanwhile, Cui et al. (2013) proposed an unsupervised method to clean bilingual data using a random walk

algorithm that computes the importance quality score of each sentence pair and selects the higher scores. Xu and Koehn (2017) presented the Zipporah system, which is designed to efficiently clean noisy web-crawled parallel corpora. Carpuat et al. (2017) focused on identifying semantic differences between sentence pairs using a cross-lingual textual entailment system. Wang et al. (2018) proposed an online denoising approach for NMT training by using trusted data to help models measure noise in sentence pairs. Artetxe and Schwenk (2019) introduced LASER based on a BiLSTM encoder that can handle 93 different languages. Our work builds on these previous studies as we implement a data filtering method based on COMET-QE, a preference model trained on human preferences. Our approach is similar to that of Bane and Zaretskaya (2021) but is significantly more robust as preference models are much more closely aligned to human judgments compared to cross-lingual encoders.

6 Conclusion

Our thorough analysis of feedback integration methods underscores the importance of meticulous data curation for enhancing MT reliability and efficiency. Our findings demonstrate the consistent improvement in translation quality when employing neural metrics, such as COMET(-QE), during training and/or inference. RL training with data filtering stands out as significantly superior to both MLE and reranking methods. Additionally, coupling RL training with reranking techniques can further enhance translation quality. While computational efficiency remains a concern due to the added overhead of RL and reranking methods on top of MLE-trained models, their adoption should be tailored to specific task and environmental requirements.

Acknowledgments

References

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, November.

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Bane, Fred and Anna Zaretskaya. 2021. Selecting the best data filtering method for NMT training. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 89–97, Virtual, August. Association for Machine Translation in the Americas.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Bengio, Samy, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

Bhattacharyya, Sumanta, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. 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 4528–4537, Online, August. Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Carpuat, Marine, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver, August. Association for Computational Linguistics.

Castricato, Louis, Alex Havrilla, Shahbuland Matiana, Duy V. Phung, Aman Tiwari, Jonathan Tow, and Maksym Zhuravinsky. 2023. trlx: A scalable framework for rlhf, jun.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, May 28–30. European Association for Machine Translation.

Cettolo, Mauro, 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, December 14-15. International Workshop on Spoken Language Translation.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cui, Lei, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for SMT using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–345, Sofia, Bulgaria, August. Association for Computational Linguistics.

Deutsch, Daniel, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab

Emirates, December. Association for Computational Linguistics.

Donato, Domenic, Lei Yu, Wang Ling, and Chris Dyer. 2022. Mad for robust reinforcement learning in machine translation.

Eikema, Bryan and Wilker Aziz. 2022. Sampling-based approximations to minimum bayes risk decoding for neural machine translation.

Fan, Angela, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July. Association for Computational Linguistics.

Fernandes, Patrick, António Farinhas, Ricardo Rei, José GC de Souza, Perez Ogayo, Graham Neubig, and André FT Martins. 2022. Quality-aware decoding for neural machine translation. *arXiv preprint arXiv:2205.00978*.

Freitag, Markus, David Grangier, Qijun Tan, and Bowen Liang. 2022a. 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, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. 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, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Jiao, Wenxiang, Xing Wang, Shilin He, Irwin King, Michael R. Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation.

Khayrallah, Huda and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia, July. Association for Computational Linguistics.

Kiegeland, Samuel and Julia Kreutzer. 2021. Revisiting the weaknesses of reinforcement learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1673–1681.

Kingma, Diederik P. and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, Philipp, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online, November. Association for Computational Linguistics.

Kong, Xiang, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2018. Neural machine translation with adequacy-oriented learning.

Kreutzer, Julia, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning.

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.

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lee, Ann, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for 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 7250–7264.

Malli, Marilena and George Tambouratzis. 2022. Evaluating corpus cleanup methods in the WMT'22 news translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 335–341, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Müller, Mathias 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, August. Association for Computational Linguistics.

Ng, Nathan, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy, August. 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.

Ott, Myle, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In Dy, Jennifer and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR, 10–15 Jul.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Papineni, Kishore, 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.

Perez, Ethan, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *CoRR*, abs/2202.03286.

Popović, Maja. 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, September. Association for Computational Linguistics.

Puterman, Martin L. 1990. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*.

Reddy, Raj. 1977. Speech understanding systems: A summary of results of the five-year research effort at carnegie mellon university.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online, November. Association for Computational Linguistics.

Rei, Ricardo, 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), December. 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. 2022b. 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), December. Association for Computational Linguistics.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Sellam, Thibault, 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, July. Association for Computational Linguistics.

Shen, Libin, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for 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 177–184.

Shen, Shiqi, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692.

Stiennon, Nisan, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,

Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback.

Sutton, Richard S. and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Taghipour, Kaveh, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Wei, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection.

Williams, Ronald J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Wiseman, Sam and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Wu, Lijun, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621.

Xu, Hainan and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation.

A Additional Results

To gain deeper insights into the effectiveness of both training and inference techniques, we also conducted a small fine-grained study evaluating the translation quality of models. Specifically, we compared translations produced by the High-Quality Subset Baseline using three different methods: MBR with COMET, RL training with COMET-QE as a reward model and a hybrid approach combining both. This complementary evaluation primarily relies on BLEURT, a neural metric highly correlated with human judgments and independent from the used reward models.

The overall BLEURT scores for these systems can be obtained from Table 2, with HQSB, HQSB + MBR w/ COMET, HQSB + RL w/ COMET-QE and HQSB + RL w/ COMET-QE + MBR w/ COMET having 75.90, 76.50, 75.94 and 77.20, respectively. Figure 4 illustrates a discernible trend: across varying lengths of source sentences, the model trained with RL and employing MBR during inference consistently yields translations of higher quality. Additionally, there is a noticeable decline in translation quality when MBR alone is employed for exceptionally long sentences, a phenomenon seemingly linked to specific hallucinations evident in Figure 3. Furthermore, Table 4 showcases the most critical examples of hallucinations obtained during this analysis.
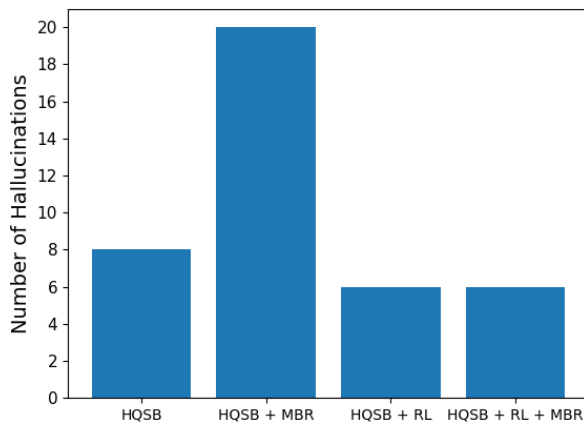

Figure 3: Number of hallucinations on the WMT16 EN→DE test set with 3000 sentences.

Examining Figures 5 and 6, depicting sentence counts across various ranges of BLEU and BLEURT scores, respectively, reveals the trend that the HQSB + RL + MBR system consistently outperforms the remaining systems across both metrics. Once again, the prevalence of low BLEU scores

---

Source: *Posted by TODAY on Monday, September 14, 2015*

Reference: *Geschrieben von TODAY am Montag, 14. September 2015*

MBR Hallucination: *Posted by TODAY am Montag, 14. September 2015, 14:45 Uhr Posted by TODAY am Montag, September 14, 2015, 14:40 Uhr Posted by TODAY am Montag, September 14, 2015, 14:00 Uhr Posted by TODAY am Montag, September 14, 2015, 14:30 Uhr Posted by TODAY am Montag, September 14, 2015, 14:30 Uhr Posted by TODAY am Montag, September 14, 2015, 14:30 Uhr Posted by TO*

RL + MBR Translation: *Veröffentlicht von TODAY am Montag, 14. September 2015*

Source: *Seehofer: "Borders will not be cordoned off"*

Reference: *Seehofer: "Grenzen werden nicht abgeriegelt"*

MBR Hallucination: *Seehofer: "Grenzen werden nicht abgeschottet" Seehofer: "Grenzen werden nicht abgeschottet" Seehofer: "Grenzen werden nicht abgeschottet" Seehofer: "Grenzen werden nicht abgeschottet" Seehofer: "Grenzen werden nicht abgeschottet" Seehofer: "Grenzen werden nicht abgeschottet" Seehofer*

RL + MBR Translation: *Seehofer: "Grenzen werden nicht abgeriegelt"*

Source: *Croatia: "We are letting the refugees through"*

Reference: *Kroatien: "Wir lassen die Flüchtlinge durch"*

MBR Hallucination: *Kroatien: "Wir lassen die Flüchtlinge durch" "Wir lassen die Flüchtlinge durch" Kroatien: "Wir lassen die Flüchtlinge durch" Kroatien: "Wir lassen die Flüchtlinge durch" Kroatien: "Wir lassen die Flüchtlinge durch" Kroatien: "Wir lassen die Flüchtlinge durch"*

RL + MBR Translation: *Kroatien: "Wir lassen die Flüchtlinge durch"*

Table 4: Instances of oscillatory hallucinations generated by the HQSB + MBR model.

underscores the issue of hallucinations associated with MBR. Furthermore, HQSB and HQSB + RL systems are quite competitive but a slight edge must be given to RL in enhancing the performance of the models

The bucketed word accuracy analysis aims to evaluate how effectively each system is at generating different types of words. Figure 7 shows that all four systems demonstrate robustness across all word frequencies but perform significantly better with higher-frequency words. Notably, among these systems, the one integrating reinforcement learning (RL) emerges as the top performer, emphasizing its effectiveness in word generation tasks.
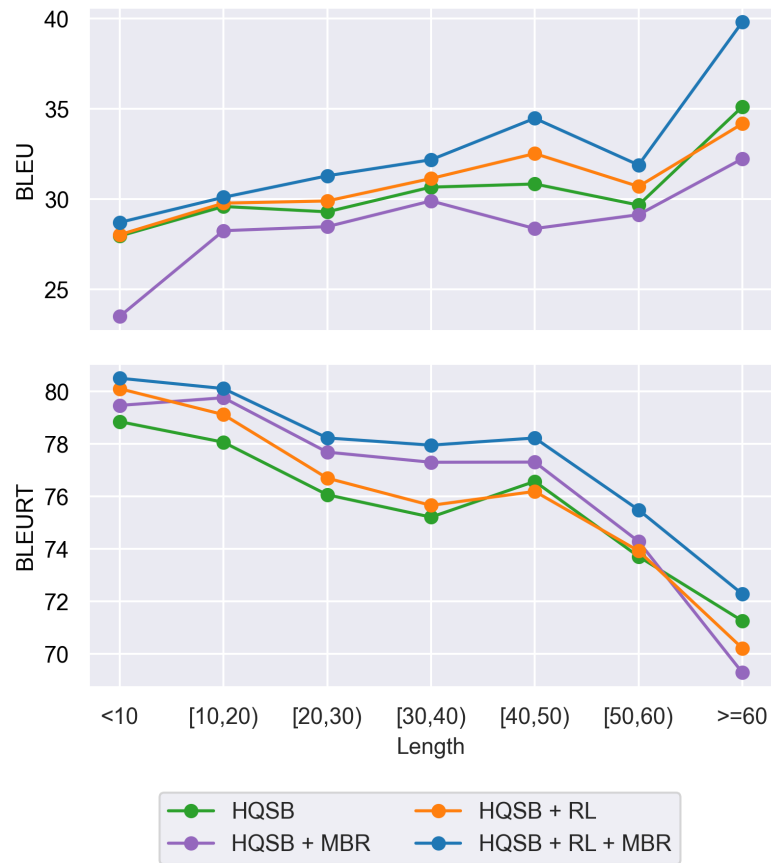
Figure 4: Comparison of BLEU (top) and BLEURT (bottom) scores for WMT16 EN→DE translations across diverse source sentence lengths, highlighting the influence of sentence length on translation quality.
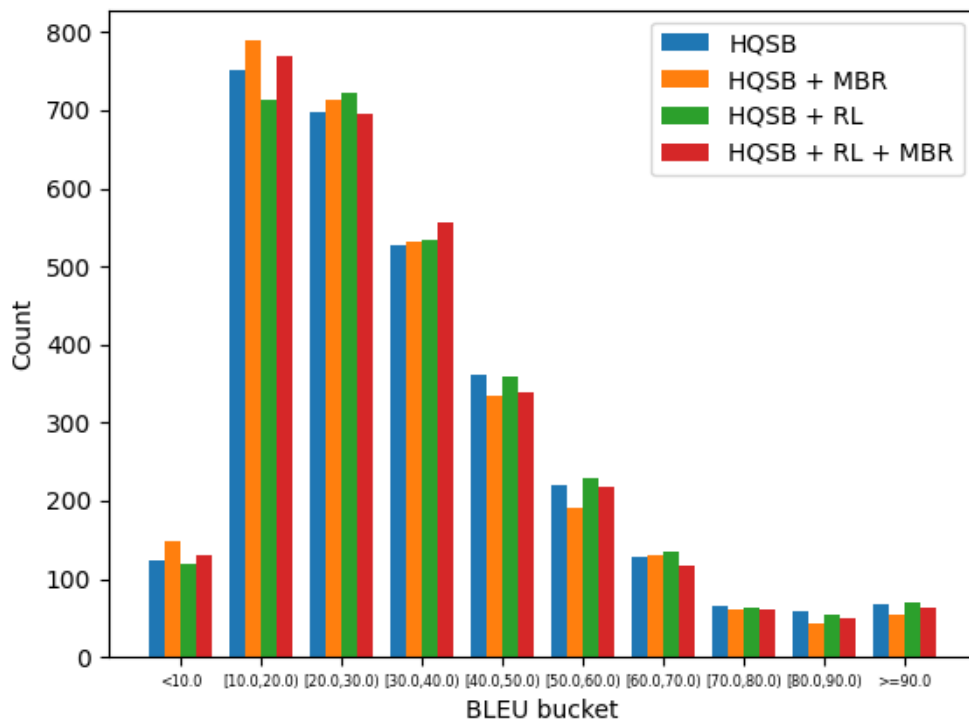

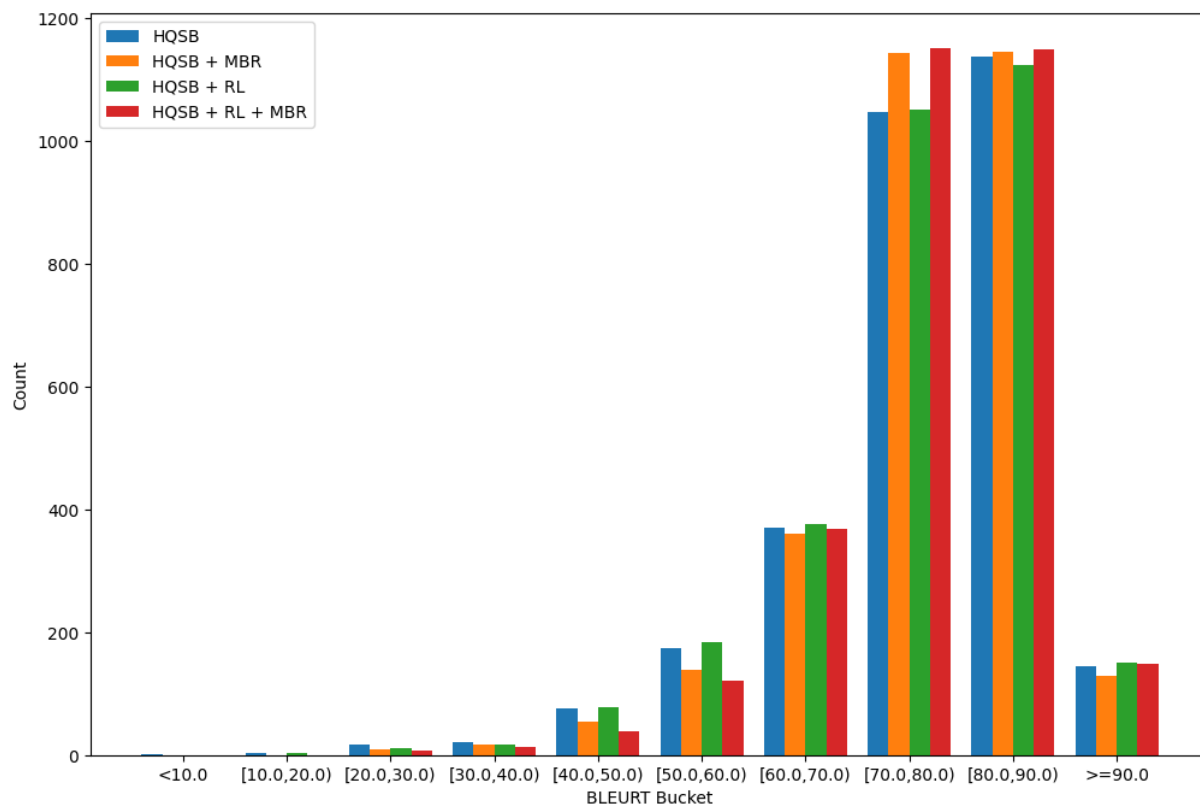Figure 5: Histograms of sentence BLEU scores for the specific systems on WMT16 EN→DE.

Figure 6: Histograms of sentence BLEURT scores for the specific systems on WMT16 EN→DE.
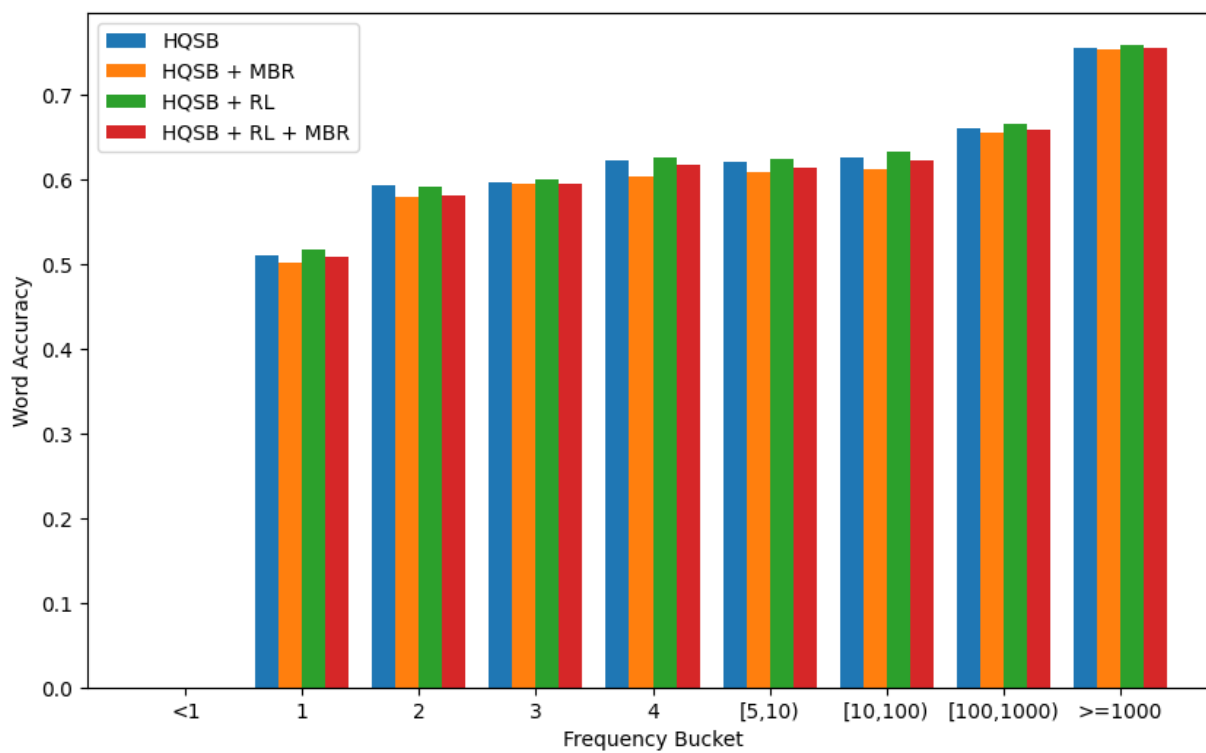

Figure 7: Word F-Measure Bucketed by Frequency for the specific systems on WMT16 EN→DE.