RMBR: A Regularized Minimum Bayes Risk Reranking Framework for Machine Translation

Anonymous ACL submission

Abstract

Beam search is the most widely used decoding method for neural machine translation (NMT). In practice, the top-1 candidate with the highest log-probability among the n candidates is selected as the 'preferred' one. However, this top-1 candidate may not be the best overall translation among the n-best list. Recently, Minimum Bayes Risk (MBR) decoding has been proposed to improve the quality for NMT, which seeks for a consensus translation that is closest on average to other candidates from the n-best list. We argue that existing MBR decoding still suffers from the following problems: The utility function only considers the lexical-level similarity between candidates; The expected utility considers the entire n-best list which is time-consuming and inadequate candidates in the tail list may hurt the performance; Only the relationship between candidates is considered. To solve these issues, we design a regularized MBR reranking framework (RMBR), which considers semantic-based similarity and computes the expected utility for each candidate by truncating the list. We expect the proposed framework to further consider the translation quality and model uncertainty of each candidate. Thus the proposed quality regularizer and uncertainty regularizer are incorporated into the framework. Extensive experiments on multiple translation tasks demonstrate the effectiveness of our method.

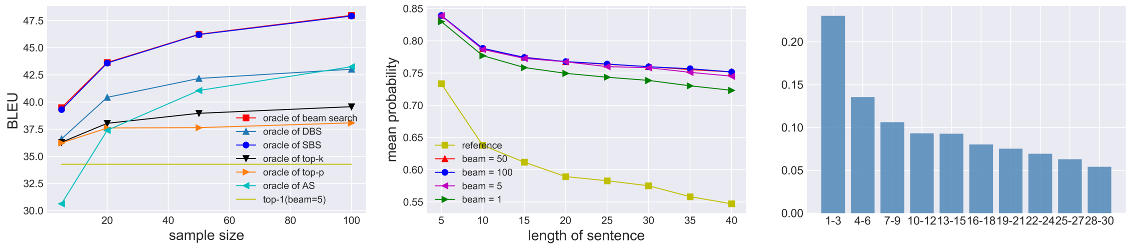
1 Introduction

Given a source sentence, neural machine translation (NMT) (Sutskever et al., 2014) models are trained to predict conditional probability distributions for candidate translations. In practice, it is desirable to output a single sentence, not a distribution. Therefore, a decision rule is required to rank the candidates and select the 'preferred' one. The most widely used decision rule is maximum-a-posteriori (MAP) decoding, which seeks the most probable translation under the conditional distribution. Due to the huge search space, beam search is proposed as an approximation. Given a pre-defined beam size n, beam search always keeps the top-n candidates based on the log-probability score. Then, the top-1 candidate, *i.e.*, the one with the highest log-probability among the n-best list, is selected as the 'preferred' one. Unfortunately, this top-1 candidate might not be the best translation on the n-best list.

We conduct oracle experiments to explore the performance gap between the oracle result[1] in the n-best candidates and top-1 candidate. Besides using beam search, we further use three stochastic decodings (ancestral search (AS) (Fu et al., 2021), top-k (Fan et al., 2018), top-p (Holtzman et al., 2020)), and two deterministic decodings (diverse beam search (DBS) (Vijayakumar et al., 2016), sibling beam search (SBS) (Li et al., 2016)) to obtain n candidates, respectively. The results are reported in Fig. 1a. The top-1 candidate of beam search with beam size 5 is used as baseline. Overall, all of the oracle results achieve *significantly* higher BLEU (Chen and Cherry, 2014) scores than baseline. For example, under the beam size 100, an oracle result of beam search achieves the high BLEU score of 47.98, while the baseline achieves only 34.28.

Furthermore, we observe that under the oracle experiment, using beam search to obtain n-best candidates still outperforms other decoding methods. These results suggest that beam search actually performs well, yet log-probability scores fail to select the best translation from the n-best list. Similar to our study, Blain et al. (2017) has observed that NMT model is capable of outputting high-quality candidate translations, but fails at picking them as the best one. Leblond et al. (2021) also points out that, NMT models are good at spreading probability mass over a large number of acceptable outputs,

[1] The oracle result is defined as $\text{argmax}_{Y \sim p_{\text{NMT}(Y|X)}}$ BLEU(Y, Y'), where (X, Y') is the pair of source and reference sentence.



(a) (b) (c)

Figure 1: An example of exploring candidate spaces on the IWSLT'14 De→En test set. (a) Oracle ranking of samples generated by multiple decoding strategies. (b) The token probabilities of sentences in different length intervals. The x-axis is the length interval, and the y-axis is the average token probability of the sentences within the same length range. (c) The distribution of oracle translations' rank index in the n-best list (n=30). The x-axis represents the index interval, and the y-axis represents the proportion of oracle translations indexed in an interval.

but they are not efficient at selecting the best one.

To further explore why the top-1 candidate is not the best translation, we compare the token probability between top-1 candidates and references. Specifically, the average probability of all the tokens in each sentence is firstly computed, which is defined as the token probability. To eliminate the effect of sentence length, the mean token probability of all candidates in the same length range is observed. As shown in Fig. 1b, we find that the token probability of top-1 candidates is much higher than that of references, especially when the result length is longer, suggesting that NMT models may over-confident about the top-1 candidates. During beam search decoding, assigning an excessively high probability to a suboptimal sequence in one step can lead to a chain reaction that eventually produces an unnatural candidate with high probability. Besides, we argue that the essence of the beam search curse (Meister et al., 2020) (large beam sizes hurt translation quality) is lying in the token probability gap between top-1 candidates and reference translations, as larger beam sizes lead to larger gaps from Fig. 1b.

In view of the above analysis, we expect to find a consensus candidate from the n-best list to avoid the "over-confident" candidates. Recently, a decision rule, Minimum Bayes Risk (MBR) decoding, which was first proposed in Goel and Byrne (2000) and Kumar and Byrne (2004), has received much attention in NMT. The main idea of this method is to find the translation that is closest to other candidate translations to minimize the expected risk for a given utility function. In Shu and Nakayama (2017) and Blain et al. (2017), MBR decoding are combined with beam search to improve the translation quality. Nevertheless, we argue that there are still some defects in MBR decoding: (a) The utility function only considers the lexical-based similarity between candidates, such as BLEU, METEOR (Denkowski and Lavie, 2011), CHRF (Popovic, 2016) etc.; (b) The expected utility for each candidate considers the entire n-best list, which requires a large computational cost, especially when n is large. Besides, inadequate candidates in the tail list may hurt the performance; (c) MBR only considers the similarity between candidates but completely ignore the model uncertainty and the translation quality of each candidate.

To solve above issues, we propose a **R**egularized **M**minimum **B**ayes **R**isk reranking framework (**RMBR**). For the first problem, we explore the use of semantic-based evaluation metrics (*e.g.*, COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020)) as the utility function. Aiming at the second issue, we conduct experiment to analyze the probability ranking of the oracle translations in the n-best list (n=30). As shown in Fig. 1c, the oracle translations are less likely to appear in the tail list. Therefore, we use only the top-l ($l \leq n$) candidates of the n-best list to calculate the MBR score (expected utility) for each candidate in the n-best list. In this way, the computational cost is reduced and the inadequate candidates in the tail list that is close to each other, are avoided. For the third problem, we incorporate two types of regularizers into the framework: quality regularizer and uncertainty regularizer. Quality regularizer allows RMBR framework to further consider the translation quality of a single candidate in addition to considering the similarity between candidates. To be concrete, we consider four regularization scores as the quality regularizer: language model score (Radford et al., 2019), back-translation score (Rapp, 2009), quality estimation score (Ranasinghe et al., 2020), and translation score (log-probability score). While the uncertainty regularizer aims to further consider the model uncertainty for each output. In this paper, we explore two kinds of uncertainty regularizers: Monte Carlo (MC) Dropout (Wang et al., 2019; Gal and Ghahramani, 2016) and the entropy of model output distributions.

We conduct extensive experiments to compare different settings of RMBR, as well as the previous MBR method (Shu and Nakayama, 2017; Blain et al., 2017) using BLEU as utility and several commonly used translation reranking methods. Experimental results show that after using COMET as utility function, our MBR outperforms previous MBR decoding methods (Shu and Nakayama, 2017; Blain et al., 2017). When the proposed quality regularizer or uncertainty regularizer is further introduced, the performance of RMBR can be further improved. Our method achieves consistent performance gains on the tasks of German-English from IWSLT'14, and German-English, English-German, and English-French tasks from WMT'14, which demonstrates the effectiveness of our method.

2 Preliminary

2.1 The Decoding Problem

Let $X = \{x_1, x_2, ..., x_{|X|}\}$ denote a source sequence, $Y = \{y_1, y_2, ..., y_{|Y|}\}$ denote a target sequence. A NMT model defines a distribution over outputs and sequentially predicts tokens using a softmax function as follows:

$$p(Y|X) = \prod_{t=1}^{|Y|} p_{\text{NMT}}(y_t|X, y_1, y_2, ..., y_{t-1}). \quad (1)$$

When $t = 1, y_0 = \text{BOS}$, which means that at the beginning of the decoding, an additional sequence start token is input. The decoding problem can be written as finding a sequence Y^* that maximizes the probability given input X:

$$Y^* = \arg\max_{Y} p(Y^*|X). \quad (2)$$

2.2 Beam Search

When decoding with the above distribution over sequences, it is not feasible to pick out the most probable sequence among all possible sequences. A common approximate decoding method is beam search, which maintains the top-n highly scoring candidates at each time step. n is known as beam size, and the log-probability of a sequence at time t is computed as:

$$S(Y_t|X) = S(Y_{t-1}|X) + \log p_{\text{NMT}}(y_t|X, Y_{t-1}), \quad (3)$$

where $S(Y_{t-1}|X) = \log p_{\text{NMT}}(y_1, y_2, ..., y_{t-1}|X)$. The decoding process is repeated until the stop condition is met. After that, we can obtain a list of n most promising candidates. Finally, the most likely sequence is selected as the 'preferred' translation by ranking the n candidates based on log-probability scores $S(Y|X)$.

3 Regularized MBR Reranking Framework

As discussed in Sec §1, picking the candidate with the highest log-probability score is unable to effectively obtain the best result. In this paper, we propose a regularized MBR reranking framework (RMBR) that adopts the semantic similarity evaluation metric as the utility function. Besides considering the similarity between the output candidates, we expect the proposed framework to further consider the translation quality of each candidate and the uncertainty of the model. Thus we incorporate two types of regularizers into the framework: Quality Regularizer (Sec §3.2) and Uncertainty Regularizer (Sec §3.3). The candidate with the highest reranked score is formally defined as the 1-best candidate.

Given a list of n most likely candidates generated by beam search with beam size n, which can be written as $\{H_1, H_2, ..., H_n\}$, the regularized score for H_i is computed as:

$$S_{\text{RMBR}}(H_i|X, H) = S_{\text{MBR}}(H_i|H) + \sum \lambda_j \mathcal{R}_j(H_i|X), \quad (4)$$

where S_{MBR} is the MBR score, which is introduced in the next section. Note that we introduce two types of regularizers, $\mathcal{R}_j$ is used to denote the j-th regularizer score. λ_j is a tradeoff parameter[2] to achieve a satisfying balance among multiple decoding objectives. Finally, the 1-best candidate is selected as the 'preferred' translation.

[2] λ_j is selected from the set $\{0.001, 0.01, 0.1, 1, 10\}$ with the best performance on the validation set. In theory, the performance could be further improved if using more advanced methods to search for weights, such as MERT (Fernandes et al., 2022), and Nelder-Mead (Singer and Nelder, 2009)

3.1 MBR Score

Given a utility function $\mathcal{U}$ (*e.g.*, BLEU) and a list of n-best candidates, the MBR score (expected utility) for each candidate is computed by comparing it to all candidates in the n-best list. Since only a few oracle translations appear at the tail list as we observed in preliminary experiment, we compute the MBR score for H_i by comparing it to top-l candidates:

$$S_{\text{MBR}}(H_i) = \frac{1}{l}\sum_{j=1}^{l}\mathcal{U}(H_i, H_j), \tag{5}$$

where $l \in \{1, 2, ..., n\}$ is tuned on the validation set and fixed for inference for all testing instances. The candidate with the highest MBR score S_{MBR} is the consensus translation in the n candidates. Besides using lexical-based method (BLEU) as utility function $\mathcal{U}$ which is called MBR_{BLEU}, we further explore two semantic-based evaluation methods BLEURT and COMET as utility functions $\mathcal{U}$ in our framework, which are called $\text{MBR}_{\text{BLEURT}}$ and $\text{MBR}_{\text{COMET}}$, respectively.

3.2 Quality Regularizer

MBR score only considers the similarity between the output candidates and ignores the translation quality of each candidate. To bridge this gap, we introduce a quality regularizer into MBR framework. In this work, we explore four kinds of scores as the quality regularizer: a) Language Model (LM) score; b) Back-Translation (BT) score; c) Quality Estimation (QE) score; and d) log-probability scores. The computation for candidate H_i is as follows:

$$\text{LM}(H_i) = \log p_{\text{LM}}(H_i), \text{QE}(H_i) = f_{\text{QE}}(X, H_i), \tag{6}$$

$$\text{BT}(H_i) = \log p_{\text{NMT}}(X|H_i), \tag{7}$$

where $p_{\text{LM}}(H_i)$ is calculated by a pre-trained language model, $p_{\text{NMT}}(X|H_i)$ is via a backward NMT model, and $f_{\text{QE}}(X, H_i)$ is by a off-the-shelf quality estimation model (*e.g.*, TransQuest (Ranasinghe et al., 2020)).

3.3 Uncertainty Regularizer

In this section, we introduce the uncertainty regularizer, which quantifies whether the current model is confident or hesitant on the candidate translation. For efficiency, we utilize widely used Monte Carlo (MC) dropout and entropy measures to compute model uncertainty.

MC Dropout. At test time, for a candidate H_i paired with input X, we perform m forward passes through the NMT model parameterized by $\hat{\theta}$, where the t-th pass randomly deactivates part of neurons. Then, m sets of sentence-level perturbed log-probability score are collected, which is written as:

$$\text{MC}_{\hat{\theta}_t}(H_i) = -\log p_{\text{NMT}}(H_i|X, \hat{\theta}_t). \tag{8}$$

Entropy Measures. We also consider using the entropy of model predicting probability distribution of each candidate as a measure of model uncertainty. Intuitively, given an output sample, if the model probability distribution entropy of each token is very small, it means that the model has a high degree of confidence in this output result. Let $\mathcal{V} = \{v_1, v_2, ..., v_{|V|}\}$ denote the target vocabulary of NMT, we compute the token entropy for each token in the candidate $H_i = \{h_{i_1}, h_{i_2}, ..., h_{i_{|H_i|}}\}$. Then $|H_i|$ sets of token entropy are collected, which is written as:

$$S_{\text{entropy}}(h_{i_t}) = -\sum_{j=1}^{|\mathcal{V}|}\log p_{\text{NMT}}(v_j|X, h_{i_0}, ..., h_{i_{t-1}}), \tag{9}$$

where h_{i_0} = BOS. Finally, the expectation of m sets of $\text{MC}_{\hat{\theta}_t}(H_i)$ and $|H_i|$ sets of $S_{\text{entropy}}(h_{i_t})$ are used as the uncertainty regularizer score.

4 Experiments

4.1 Experimental Settings

In this section, we describe the datasets, NMT models, and metrics used in our experiments to investigate the effect of the proposed reranking methods on the n-best candidate list.

4.1.1 Datasets and Models

To implement the NMT task, we use the German-English (De→En) from IWSLT'14 task, German-English (De→En), English-German (En→De), and English-French (En→Fr) from the WMT'14 translation task. For IWSLT'14 task, we use the data preprocessing scripts and hyperparameter settings provided by fairseq NMT repository[3]. For WMT'14 task, we train a Transformer base model (Vaswani et al., 2017) as the base NMT model and use the Newstest'14 dataset as the test set.

[3]https://github.com/pytorch/fairseq/tree/master/examples/translation.

	IWSLT'14 De→En			WMT'14 De→En		
Method	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU
Top-1 (beam=5)	34.79	16.16	34.28	42.35	21.90	32.70
Top-1 (beam=30)	34.22	15.99	34.17	41.80	21.60	32.54
LP+BT (Rapp, 2009)	40.63	18.57	35.11	45.94	23.42	33.06
LP+QE (Ranasinghe et al., 2020)	38.84	19.53	35.37	45.56	24.30	33.41
LP+LM (Radford et al., 2019)	36.33	16.58	35.14	44.48	22.48	33.49
Range Voting (Borgeaud and Emerson, 2020)	34.89	16.59	34.53	42.29	21.53	32.78
MBR$_{\text{BLEU}}$(full) (Blain et al., 2017)	33.76	15.91	34.38	41.66	20.96	32.68
MBR$_{\text{BLEU}}$	34.39	16.39	34.54	42.53	22.03	32.83
MBR$_{\text{BLEURT}}$	33.10	**22.00**	33.01	42.71	**25.31**	32.45
MBR$_{\text{COMET}}$	42.53	17.78	34.55	47.10	23.06	32.93
MBR$_{\text{COMET}}$+LP	41.60	17.89	34.91	46.69	22.89	33.08
MBR$_{\text{COMET}}$+LP+BT	**43.64**	18.86	35.24	**47.67**	23.57	33.17
MBR$_{\text{COMET}}$+LP+QE	42.04	19.96	35.62	46.89	23.57	33.76
MBR$_{\text{COMET}}$+LP+LM	41.75	18.40	35.49	47.56	23.91	33.85
MBR$_{\text{COMET}}$+LP+entropy	42.04	18.34	35.24	46.24	22.99	33.16
MBR$_{\text{COMET}}$+LP+dropout	41.47	17.90	34.95	47.43	22.91	33.10
MBR$_{\text{COMET}}$+LP+QE+LM	42.24	20.60	**36.19**	47.34	25.18	**34.29**

Table 1: BLEU, COMET, and BLEURT score comparison. All candidates are obtained by beam search.

4.1.2 Evaluation Metrics

In our experiments, three widely used automatic evaluation metrics are utilized to evaluate the machine translation: BLEU, an n-gram-based precision metric which measures the lexical similarly between translation and reference; COMET (Rei et al., 2020), a multilingual and adaptable MT evaluation model, which exploits information from both source sentence and target sentence to measures the semantic similarity between translation and reference; and BLEURT (Sellam et al., 2020), a learned evaluation metric based on BERT, which measures the semantic similarity between two sequences.

4.2 Baselines

We take the top-1 results of the beam search with beam size 5 as the baseline, which is the most widely used setting of NMT models. For all reranking methods, we follow previous work (Eikema and Aziz, 2020) using beam search with beam size 30 to generate the candidates (experimental results with varying beam size and different decoding method can be found in **Appendix D** and **Appendix A**, respectively). MBR$_{\text{COMET}}$ denotes use only MBR score to rank the candidate without any regularizer, where COMET is used as the utility function. Besides, we also compare MBR$_{\text{BLEU}}$ and MBR$_{\text{BLEURT}}$ which use BLEU and BLEURT as utility function, respectively. We further compare the performance of introducing different regularizer on MBR$_{\text{COMET}}$, including four kinds of quality regularizer scores: log-probability (LP) score, language model (LM) score, back-translation (BT) score, quality estimation (QE) score, and two uncertainty regularizer scores: entropy score and MC-dropout score. We use GPT-2$_{\text{base}}$ model (Radford et al., 2019) to calculate LM score. BT score and QE score is computed via backward NMT models and TransQuest (Ranasinghe et al., 2020), respectively. For the proposed method, we compute MBR score for each candidate by comparing it to partial top candidates, where the details are reported in Sec §5.3. We also compare the method Range Voting (Borgeaud and Emerson, 2020) and MBR$_{\text{BLEU}}$(full) (Blain et al., 2017), which using BLEU as utility function of MBR. The only difference between MBR$_{\text{BLEU}}$(full) (Blain et al., 2017) and our MBR$_{\text{BLEU}}$ is that MBR$_{\text{BLEU}}$(full) uses all candidates to calculate MBR score.

4.3 Results

We first report the results on IWSLT'14 De→En and WMT'14 De→En tasks. From Table 1, we can see that MBR$_{\text{COMET}}$ outperforms MBR$_{\text{BLEU}}$, top-1, and other baselines on all three evaluation metrics. Interestingly, we find that MBR$_{\text{BLEURT}}$ achieves the highest BLEURT score but low BLEU and COMET scores. To find out which utility function is the best, we further perform human evaluation (see Sec §5.1) to more quantitatively compare the reranked 1-best candidates. The human evaluation results show that $\text{MBR}_{\text{COMET}}$ outperforms MBR_{BLEU} and $\text{MBR}_{\text{BLEURT}}$, demonstrating that semantic-based MBR outperforms traditional lexical-based MBR. For the proposed regularizers, $\text{MBR}_{\text{COMET}}$+LP improves the scores in BLEU comparing to $\text{MBR}_{\text{COMET}}$. Besides, $\text{MBR}_{\text{COMET}}$+LP can be further improved in three metrics by adding other regularizers. For example, the $\text{MBR}_{\text{COMET}}$+LP+QE achieves higher scores on BLEU, COMET, and BLEURT. In addition, a similar trend is observed in $\text{MBR}_{\text{BLEURT}}$ and $\text{MBR}_{\text{COMET}}$. More results and details can be found in Sec §5.4. The regularized MBR reranking works better than beam search with sizes 5 and 30, bringing 8 points and 1.5 points of improvement on COMET and BLEU metrics, respectively.

We additionally explore the performance of combining more regularizers on $\text{MBR}_{\text{COMET}}$. We collectively tune the λ value for each of the regularizers on validation sets. We observe the results of $\text{MBR}_{\text{COMET}}$+LP+QE+LM (we use $\text{RMBR}_{\text{COMET}}$ to denote this setting latter) that achieves the highest BLEU score among all the combinations, improving the BLEU score by more than 2 points. We also find that combining quality and uncertainty regularizers with $\text{MBR}_{\text{COMET}}$ can not lead to further performance gains. Also, we conduct re-reanking experiment on WMT'19 De→En and En→De tasks, which can be found in **Appendix C**. Moreover, we carry experiment to evaluate the effectiveness of larger beam size on our proposed method. More experimental results are reported in **Appendix D**. The results suggest that our proposed reranking method can alleviate the beam search curse and generate better translations as beam size increases.

Method	Score
$\text{MBR}_{\text{COMET}}$	0.281
$\text{MBR}_{\text{BLEURT}}$	0.129
MBR_{BLEU}	0.125
Top-1 (beam=5)	0.120

Table 2: Results of the human evaluation. The score column represents the percentage of times each reranking method is judged better across its competitors.

5 Analysis

5.1 Human Evaluation

From the previous results, we observe that $\text{MBR}_{\text{COMET}}$ outperforms MBR_{BLEU} and $\text{MBR}_{\text{BLEURT}}$ in BLEU and COMET metrics, but not in BLEURT metric. This motivate us to perform human evaluation to more quantitatively compare the reranked results. We randomly select a subset of 500 source sentences from the test sets of IWSLT'14 De→En. Reranking is also based on the beam search results of beam size 30. We request 3 human annotators to rank the four translations from the best to the worst. Specifically, we first set a guideline for evaluating, which includes the task background, key points, detailed descriptions, and 5 examples. Then, we set an entry barrier for annotators. In detail, we organize a training program and a preliminary annotating examination (50 examples for each baseline) to select appropriate annotators with an approval rate higher than 95%. All the annotators are highly educated, and the cost of the evaluation is about 0.05$ for each word by one annotator. Table 2 reports the ranking results according to the Expected Wins method (Sakaguchi et al., 2014). Our observation is that the 1-best candidates reranking by $\text{MBR}_{\text{COMET}}$ outperforms the other three methods. We provide some examples in **Appendix B**.

	WMT'14 En→De		WMT'14 En→Fr	
Methods	COMET	BLEU	COMET	BLEU
Top-1 (beam=5)	27.24	27.09	55.11	38.74
Top-1 (beam=30)	20.32	26.50	50.31	38.22
LP+QE	28.10	27.80	55.39	39.60
LP+LM	27.92	28.04	56.10	39.62
LP+BT	27.50	27.75	56.06	39.70
$\text{MBR}_{\text{COMET}}$	34.25	27.37	59.85	39.18
MBR_{BLEU}	26.15	27.30	53.81	39.17
$\text{MBR}_{\text{COMET}}$+LP	31.98	27.93	57.88	39.58
$\text{MBR}_{\text{COMET}}$+LP+BT	32.53	28.01	**60.33**	39.83
$\text{MBR}_{\text{COMET}}$+LP+QE	32.71	28.00	59.83	39.84
$\text{MBR}_{\text{COMET}}$+LP+LM	**34.97**	28.19	59.80	39.87
$\text{MBR}_{\text{COMET}}$+LP+QE+LM	32.51	**28.40**	59.71	**40.15**

Table 3: BLEU and COMET score comparison on WMT'14 En→De and WMT'14 En→Fr tasks.

5.2 Results on non-English Target Translation Tasks

To further verify the effectiveness of the proposed model on non-English target translation tasks, we conduct experiments on WMT'14 En→Fr and En→De, where we follow the same settings in Sec §4.2. Since the evaluation metric BLEURT only supports evaluation the language of English, we only report BLEU and COMET scores for En→Fr and En→De tasks. The results are shown in Table 3, which are consistent with those in Table 1.

5.3 N-by-L

The number of candidates used to compute expected utility is defined as l in Sec §3.1. To explore the effectiveness of l on BLEU score of the reranked 1-best candidates, we use $\text{MBR}_{\text{COMET}}$ and MBR_{BLEU} to rank the 30 candidates decoded by beam search with beam size of 30. We compute the expected utility for each candidate by comparing it to top-l candidates of the 30 candidates. The results are shown in Fig. 2. As l increases, the BLEU scores of the 1-best candidates reranked by both $\text{MBR}_{\text{COMET}}$ and MBR_{BLEU} go up and then down. The reason may be that partial candidates near the end of the list is extremely close to each other, but of poor quality. When l increases, this part of candidates are more likely to be selected. When l is around 21, BLEU scores of $\text{MBR}_{\text{COMET}}$ and MBR_{BLEU} are close to the optimal. For the proposed reranking method, l is tuned on the validation set and fixed for inference for all testing instances.

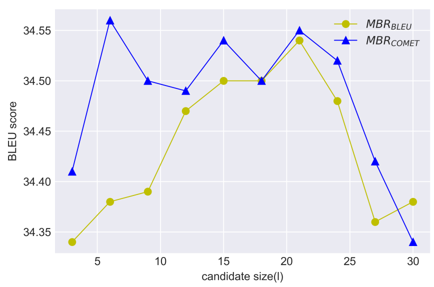


Figure 2: The reranking results using partial candidates to compute expected utility on the IWSLT'14 De→En dev sets. y-axis is the BLEU score. x-axis is the number of candidates used to compute MBR scores.

5.4 Utility Functions

To further verify the effectiveness of different utility functions, we also compare the performance of introducing the quality regularizers that performs well in previous experiments on MBR_{BLEU} and $\text{MBR}_{\text{BLEURT}}$. We follow the same settings

Method	COMET	BLEURT	BLEU
Top-1 (beam=5)	34.79	16.16	34.28
Top-1 (beam=30)	34.22	15.99	34.17
MBR_{BLEU}	34.39	16.39	34.54
MBR_{BLEU}+LP	34.75	16.64	34.56
MBR_{BLEU}+LP+BT	**42.48**	19.03	35.17
MBR_{BLEU}+LP+QE	38.68	19.75	35.44
MBR_{BLEU}+LP+LM	38.89	19.91	35.41
MBR_{BLEU}+LP+QE+LM	39.82	19.92	35.81
$\text{MBR}_{\text{BLEURT}}$	33.10	**22.00**	33.01
$\text{MBR}_{\text{BLEURT}}$+LP	35.83	19.86	34.55
$\text{MBR}_{\text{BLEURT}}$+LP+BT	42.46	19.20	35.18
$\text{MBR}_{\text{BLEURT}}$+LP+QE	38.91	20.19	35.42
$\text{MBR}_{\text{BLEURT}}$+LP+LM	36.79	18.04	35.25
$\text{MBR}_{\text{BLEURT}}$+LP+QE+LM	40.65	20.49	36.14
$\text{MBR}_{\text{COMET}}$+LP+QE+LM	42.24	20.60	**36.19**

Table 4: Comparison results of $\text{MBR}_{\text{BLEURT}}$ and MBR_{BLEU} with the proposed quality regularizers on IWSLT'14 De→En.

in Sec §4.2. As shown in Table 4, similar to $\text{RMBR}_{\text{COMET}}$, $\text{RMBR}_{\text{BLEU}}$ and $\text{RMBR}_{\text{BLEURT}}$ also outperform beam search with sizes 5 and 30, which is consistent with the results shown in Table 1 and Table 3. Overall, $\text{RMBR}_{\text{BLEURT}}$ variants achieve better scores than $\text{RMBR}_{\text{BLEU}}$ variants, and $\text{RMBR}_{\text{COMET}}$ variants perform best. These results show that semantic-based MBR leads to better translation options.

5.5 Inference Time

We further compare the inference time of the proposed reranking variants and baseline. For reranking, we still use 30 candidates obtained by beam search on the IWSLT'14 De→En test sets. To compare the inference time, all experiments are performed on single Tesla V100 16GB GPU. Note that, in practice we can further reduce inference time by using more GPUs to compute utility functions in parallel. The results are shown in Table 5. n represents the number of candidates used to rerank, l represents the number of candidates used to compute expected utility ($n = 30, l_1 = 21, l_2 = 3$). For $\text{RMBR}_{\text{COMET}}$(C2F), which is a coarse-to-fine MBR procedure proposed in Eikema and Aziz (2021), we use BLEU as the proxy utility to select 15 candidates and then use COMET as the target utility to select the 1-best candidate. From the results we can see that $\text{RMBR}_{\text{COMET}}$(n-by-$l_1$) achieves the best performance with about 3.6 times more inference time than top-1 (beam=5). Both $\text{RMBR}_{\text{COMET}}$(n-by-$l_2$)

Methods	COMET	BLEURT	BLEU	Time
Top-1 (beam=5)	34.79	16.16	34.28	x1
$\text{RMBR}_{\text{COMET}}$(n-by-n)	42.52	20.47	36.01	x4.7
$\text{RMBR}_{\text{COMET}}$(n-by-$l_1$)	42.24	20.60	36.19	x3.6
$\text{RMBR}_{\text{COMET}}$(n-by-$l_2$)	40.93	20.26	35.90	x1.4
$\text{RMBR}_{\text{COMET}}$(C2F)	41.50	19.41	35.93	x1.9

Table 5: Comparison results of inference time. Reranking uses $n = 30$ candidates per sample.

and $\text{RMBR}_{\text{COMET}}$(C2F) can further reduce inference time and outperform the baseline, which can be used as a trade-off between time cost and performance.

6 Related Work

In NMT, reranking is a way of improving translation quality by scoring and selecting a 'preferred' translation from a list of candidates generated by a source-to-target model. MBR decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004) is one of the effective methods. The goal of MBR decoding is to find a consensus translation that is closest to other candidates. Some studies rerank the n candidates directly sampled from the model. Eikema and Aziz (2020) is the first to use unbiased samples from the model by ancestral sampling, to approximate hypotheses space. Aiming at keeping computational cost of estimating expected utility tractable, a coarse-to-fine MBR procedure is proposed in Eikema and Aziz (2021). Other studies tend to rerank the n candidates decoded by beam search. In Shu and Nakayama (2017), both MBR scores and log-probability scores are considered at each step of decoding. Blain et al. (2017) investigates some automatic MT evaluation metrics (BLEU, BEER, and CHRF), and observes that evaluation metric plays a major role in the n-best reranking approach. Borgeaud and Emerson (2020) designs some similarity functions to make more informative candidates receive stronger votes, thus selecting the most representative candidate.

These previous studies only use MBR score to rank each candidate without considering source sentence and model score. In the proposed RMBR, some regularizers are utilized to rank candidates in an overall way. Different from previous works which select candidates based on only lexical similarity, we also explore the semantic similarity between candidates. The other difference is that MBR score is computed using top-l candidates of the n-best list to avoid candidates with poor quality in the tail list and reduce the computation cost.

Besides MBR, there are some studies focus on MT reranking. For example, Ng et al. (2019) describes using language model to rank candidates. In Bhattacharyya et al. (2021), an energy based model is trained to rank samples drawn from NMT. Lee et al. (2021) predicts the observed distribution of a desired metric, *e.g.*, BLEU, over the n-best list by training a large transformer architecture. Note that these methods are orthogonal to our method, and they can be theoretically used as the quality regularizer in our framework.

Uncertainty quantification (Hüllermeier and Waegeman, 2021) have been widely used in neural networks, which is usually solved by Bayesian frameworks. Because the high training cost brought by Bayesian neural networks, various approximations, such as Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) and model ensembling (Lakshminarayanan et al., 2017) have been developed. In NMT, the MC dropout is used at test time, by performing several stochastic forward passes through the model. Then, the expectation or variance of the output which reflect whether the current model is confident or hesitant on the translation, is used to evaluate machine translation quality (Fomicheva et al., 2020). On the other hand, in the image classification task, entropy based measures are used to address uncertainty quantification (Smith and Gal, 2018). Our uncertainty regularizers adopt similar uncertainty quantification strategies.

7 Conclusion

In this paper, we introduce a RMBR to choose adequate translations from the candidates decoded by beam search. Based on MBR, we adopt semantic-based similarity and compute the expected utility by truncating the list. The proposed quality and uncertainty regularizers are further incorporated into the framework. Extensive experimental results show that RMBR outperforms several MBR-based variants and other reranking baselines on MT tasks: +1.9 BLEU points, +7.5 COMET points, +4.4 BLEURT points over the results of beam search with sizes 5 on IWSLT'14 German→English. To get a better insight into RMBR, we also conduct the in-depth ablation study and analytical experiments to show the performance improvement brought by each component of RMBR.

References

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *ACL/IJCNLP*.

Frédéric Blain, Lucia Specia, and Pranava Madhyastha. 2017. Exploring hypotheses spaces in neural machine translation. In *AAMT*.

Sebastian Borgeaud and Guy Emerson. 2020. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In *NGT@ACL*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *WMT@ACL*.

Michael J. Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *WMT@EMNLP*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *COLING*.

Bryan Eikema and Wilker Aziz. 2021. Sampling-based minimum bayes risk decoding for neural machine translation. *arXiv preprint arXiv: 2108.04718*.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José Guilherme Camargo de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. 2022. Quality-aware decoding for neural machine translation. In *NAACL*, pages 1396–1412. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguistics*.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *AAAI*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*.

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

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

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506.

Shankar Kumar and William J. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *HLT-NAACL*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*.

Rémi Leblond, Jean-Baptiste Alayrac, Laurent Sifre, Miruna Pislar, Jean-Baptiste Lespiau, Ioannis Antonoglou, Karen Simonyan, and Oriol Vinyals. 2021. Machine translation decoding beyond beam search. In *EMNLP*.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *ACL/IJCNLP*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv: 1611.08562*.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *EMNLP*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. In *WMT*.

Maja Popovic. 2016. chrf deconstructed: beta parameters and n-gram weights. In *WMT*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. In *COLING*.

Reinhard Rapp. 2009. The backtranslation score: Automatic mt evalution at the sentence level without reference translations. In *ACL*.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP*.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *WMT@ACL*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *ACL*.

Raphael Shu and Hideki Nakayama. 2017. Later-stage minimum bayes-risk decoding for neural machine translation. *arXiv preprint arXiv: 1704.03169*.

Sasa Singer and John A. Nelder. 2009. Nelder-mead algorithm. *Scholarpedia*, 4(7):2928.

Lewis Smith and Yarin Gal. 2018. Understanding measures of uncertainty for adversarial example detection. In *UAI*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv: 1610.02424*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *EMNLP-IJCNLP*.

Source	Wir erwarten ein paar außergewöhnliche Jahrzehnte.
Reference	We are living into extraordinary decades ahead.
Top-1 (beam=5)	We expect some extraordinary years.
$\text{MBR}_{\text{COMET}}$	We are looking forward to extraordinary decades.
$\text{MBR}_{\text{BLEURT}}$	We expect some extraordinary decades.
MBR_{BLEU}	We expect for several remarkable decades.

Table 7: Examples of 1-best candidates chosen by the proposed reranking methods from n-best list (with n = 30). Underline represents the main differences between the reference, the top-1 candidates, and the reranked 1-best candidates.

	WMT'19 De→En			WMT'19 En→De	
Method	COMET	BLEURT	BLEU	COMET	BLEU
Top-1 (beam=5)	44.82	25.00	40.02	41.53	41.23
Top-1 (beam=30)	44.77	24.71	39.86	41.50	41.14
LP+BT	45.77	26.97	40.33	40.73	41.46
LP+QE	45.61	25.47	40.20	41.88	41.52
LP+LM	45.18	2509	40.13	41.44	41.44
MBR_{BLEU}	45.05	24.56	39.89	41.19	41.35
$\text{MBR}_{\text{BLEURT}}$	44.70	**28.05**	37.91	\	\
$\text{MBR}_{\text{COMET}}$	49.03	25.74	39.88	**45.49**	41.38
$\text{MBR}_{\text{COMET}}$+LP	48.05	26.00	40.21	45.02	41.49
$\text{MBR}_{\text{COMET}}$+LP+BT	49.47	26.71	40.39	45.03	41.69
$\text{MBR}_{\text{COMET}}$+LP+QE	48.83	27.80	40.51	42.96	41.63
$\text{MBR}_{\text{COMET}}$+LP+LM	46.34	25.39	40.24	43.69	41.56
$\text{MBR}_{\text{COMET}}$+LP+QE+BT	**50.28**	27.92	**40.56**	45.15	**41.73**

Table 8: Reranking results on WMT'19 De→En and WMT'19 En→De tasks.

Methods	COMET	BLEURT	BLEU
Beam Search (beam=30)			
Top-1 (beam=30)	34.22	15.99	34.17
$\text{MBR}_{\text{COMET}}$	**42.53**	17.78	**34.55**
MBR_{BLEU}	34.39	16.39	34.54
$\text{MBR}_{\text{BLEURT}}$	33.10	**22.00**	33.01
Siblings Beam Search (beam=30)			
Top-1 ($n = 30$)	34.11	15.67	34.09
$\text{MBR}_{\text{COMET}}$	**41.44**	17.16	34.39
MBR_{BLEU}	33.83	16.04	**34.42**
$\text{MBR}_{\text{BLEURT}}$	31.78	**21.68**	32.95
Ancestral Sampling (n=30)			
Top-1 ($n = 30$)	21.37	10.62	29.33
$\text{MBR}_{\text{COMET}}$	**30.44**	13.71	28.27
MBR_{BLEU}	9.67	8.99	**30.62**
$\text{MBR}_{\text{BLEURT}}$	9.12	**19.74**	22.81

Table 6: The reranking results from 30 candidates decoded by beam search, SBS, and AS on the test sets of IWSLT'14 De→En.

A Diverse Candidate Spaces

From the oracle experiments (see Fig.1a), we observe that deterministic decoding performs better than stochastic decoding, and sibling beam search (SBS) performs as well as beam search. To further explore the effect of diverse candidate spaces, we rerank the 30 top candidates by SBS and 30 candidates sampled by AS. As shown in Table 6, the reranking results of the candidates decoded by SB perform slightly worse than that of beam search. For AS decoding, the scores of both top-1 candidates and reranked 1-best candidates are significantly low compared to other reranking methods.

B Qualitative Analysis

In Table 7, we illustrate some examples from the reranking approach. Although, the word overlap between the 1-best candidates by regularized MBR ranker and the top-1 candidates is high, the proposed reranking methods produce accurate and fluent translation with asyntactic re-orderings, new words, morphological variations.

C Experiments on WMT'19 Translation tasks

To further verify the effectiveness of the proposed model on the newly translation tasks, we conduct experiments on WMT'19 De→En and En→De. For baseline, we use the best performing single

Method	COMET	BLEURT	BLEU
Top-1 (beam=5)	34.79	16.16	34.28
Top-1 (beam=30)	34.22	15.99	34.17
Top-1 (beam=50)	33.84	15.87	34.10
beam=50			
MBR$_{\text{COMET}}$	43.50	18.27	34.57
MBR$_{\text{COMET}}$+LP	42.35	18.11	34.94
MBR$_{\text{COMET}}$+LP+BT	**44.42**	18.97	35.31
MBR$_{\text{COMET}}$+LP+QE	42.74	20.26	35.62
MBR$_{\text{COMET}}$+LP+LM	42.87	18.96	35.58
MBR$_{\text{COMET}}$+LP+QE+LM	42.62	**21.54**	**36.24**
beam=30			
MBR$_{\text{COMET}}$	42.53	17.78	34.55
MBR$_{\text{COMET}}$+LP	41.60	17.89	34.91
MBR$_{\text{COMET}}$+LP+BT	43.64	18.86	35.24
MBR$_{\text{COMET}}$+LP+QE	42.04	19.96	35.62
MBR$_{\text{COMET}}$+LP+LM	41.75	18.40	35.49
MBR$_{\text{COMET}}$+LP+QE+LM	42.24	20.60	36.19
beam=5			
MBR$_{\text{COMET}}$	36.65	16.03	34.19
MBR$_{\text{COMET}}$+LP	36.44	16.47	34.40
MBR$_{\text{COMET}}$+LP+BT	38.99	17.38	34.69
MBR$_{\text{COMET}}$+LP+QE	38.09	18.20	34.86
MBR$_{\text{COMET}}$+LP+LM	36.73	16.81	34.78
MBR$_{\text{COMET}}$+LP+QE+LM	37.67	18.70	35.28

Table 9: Comparison results of beam size 5, 30, and 50 on IWSLT'14 De→En.

pre-trained model and data pre-processing method provided by fairseq NMT repository. Reranking follows the same settings in Sec §4.2. Since the evaluation metric BLEURT only supports evaluation the language of English, we only report BLEU and COMET scores for En→De. The results are shown in Table 8, which is consistent with the conclusion in Table 1.

D Beam Sizes

In this section, we explore the performance of the proposed RMBR$_{\text{COMET}}$ reranking in large beam sizes, which performs best on average of three metrics on the IWSLT'14 De→En test sets. As shown in Table 9 and Fig. 3, the translation quality of beam search decreases with increased beam sizes. Notably, RMBR$_{\text{COMET}}$ achieves higher score in COMET, BLEU, and BLEURT with larger beam sizes, which suggests that RMBR benefits from larger beam sizes. Moreover, the 1-best candidates of RMBR$_{\text{COMET}}$ far outperforms the top-1 candidates of beam search with sizes 5, 30, and 50. The results means that the proposed reranking method can improve upon beam search.

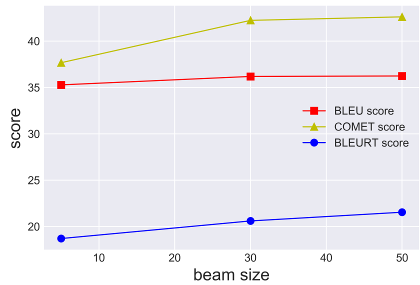


Figure 3: The results of the 1-best candidates reranked by the RMBR$_{\text{COMET}}$ using beam of sizes 5, 30, and 50.