Unveiling the Power of Source: Source-based Minimum Bayes Risk Decoding for Neural Machine Translation

Anonymous ACL submission

Abstract

Maximum a posteriori decoding, a commonly used method for neural machine translation (NMT), aims to maximize the estimated posterior probability. However, high estimated probability does not always lead to high translation quality. Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) offers an alternative by seeking hypotheses with the highest expected utility.

In this paper, we show that Quality Estimation (QE) reranking (Fernandes et al., 2022), which uses a QE model as a reranker, can be viewed as a variant of MBR. Inspired by this, we propose source-based MBR (sMBR) decoding, a novel approach that utilizes synthetic sources (generated via back-translation or paraphrasing) as "support hypotheses" and a reference-free quality estimation metric as the utility function, marking the first work to solely use sources in MBR decoding. Experiments show that sMBR outperforms QE reranking and the standard MBR decoding. Our findings suggest that sMBR is a promising approach for NMT decoding.[1]

1 Introduction

Neural Machine Translation (NMT) models typically aim to select a hypothesis with the highest estimated posterior probability during decoding, an approach known as Maximum A Posteriori (MAP) decoding. Beam search (Graves, 2012; Sutskever et al., 2014), which balances computational cost and search accuracy, has become the standard approximate decoding method for MAP.

However, the underlying assumption of beam search - that estimated probability is a good proxy for translation quality - has been challenged by evidence showing that estimated probability and quality do not always correlate positively (Ott et al., 2018; Freitag et al., 2021). For example, Fig 1 illustrates a case where a human reference translation has a lower estimated probability than the hypothesis generated by beam search, and even lower than that of a poor translation. Furthermore, the true MAP output is sometimes an empty string or overly brief translation (Koehn and Knowles, 2017; Murray and Chiang, 2018; Ott et al., 2018; Stahlberg and Byrne, 2019). These suggest that solely searching for high estimated probability hypotheses may not be an effective strategy for improving translation quality.

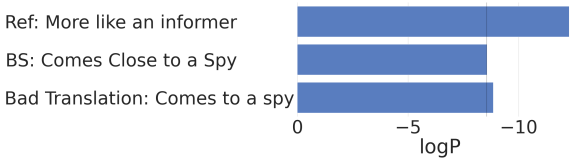

Figure 1: Example of De→En, with source *"Kommt einem Spitzel nahe"*. BS denotes beam search. The estimated log probability of a human reference is lower than that of the beam search output, and even lower than that of a bad translation.

Given the limitations of using estimated probability as a proxy for quality, an attractive alternative is to directly target translation quality during decoding (Freitag et al., 2022). Minimum Bayes Risk (MBR) decoding, proposed in the era of statistical machine translation (Kumar and Byrne, 2004; Tromble et al., 2008), aims to find the hypothesis with the highest expected utility compared to ideal translations. Traditionally, surface-based evaluation metrics like BLEU (Papineni et al., 2002) were used as the utility function in MBR decoding (Eikema and Aziz, 2020; Eikema and Aziz, 2022). However, these metrics have shown limited correlation with human judgments (Mathur et al., 2020; Freitag et al., 2023b), hindering the widespread adoption of MBR decoding based on them. Recent work has explored using state-of-the-art neural met-

[1]We will make the source code publicly available once the paper is accepted.

rics, such as COMET (Rei et al., 2022a), as utility functions for MBR decoding (Freitag et al., 2022; Fernandes et al., 2022; Freitag et al., 2023a), showing promising improvements in human evaluations.

Moreover, advances in reference-free evaluation metrics (Rei et al., 2021; Rei et al., 2022b; Rei et al., 2023) have enabled their direct application to hypothesis reranking, which we refer to as Quality-Estimation (QE) reranking (Fernandes et al., 2022). QE reranking selects the hypothesis with the highest reference-free quality estimation score among the candidate hypotheses. However, QE reranking remains understudied compared to MBR decoding.

In this work, we establish a connection between QE reranking and MBR decoding, showing that QE reranking can be viewed as a variant of MBR when using a source as "support hypotheses" and a QE metric as the utility function. Inspired by this finding, we propose a novel approach called **source-based MBR** (sMBR) decoding, which uses synthetic sources generated by back-translation or paraphrasing as support hypotheses and a QE metric as the utility function. This marks the first work to solely use sources as support hypotheses in MBR decoding, breaking the long-standing tradition of relying on using other hypotheses to approximate true utility for this purpose. See Fig 2 for a overview of our methodology.

In this paper, we first explain conceptually why QE reranking can be considered a variant of MBR decoding and introduce sMBR decoding. We then provide empirical evidence through comprehensive experiments on three translation directions in both classic (large transformer models trained from scratch, including high-resource and low-resource) and large language model (LLM) setups, demonstrating that sMBR outperforms QE reranking and the standard MBR decoding. These findings suggest that sMBR is a promising NMT decoding approach.

2 Decoding methods in NMT

Decoding can be viewed as two phases: *hypothesis generation* and *decision*. Specifically, in the hypothesis generation phase, a certain generation method, such as beam search, is used to generate N hypotheses from the model $\{h_0, h_1, \ldots, h_{N-1}\}$. Then, in the decision phase, N decision scores need to be computed for each of these N hypotheses $\{score_{h_0}, score_{h_1}, \ldots, score_{h_{N-1}}\}$. Finally, the hypothesis with the highest decision score is selected as the final output.

2.1 MAP decoding

Given the source space $\mathcal{X}$, a source sentence $x \in \mathcal{X}$, model parameters θ, and a hypothesis space $\mathcal{H}$, MAP decoding aims to find the hypothesis y^{MAP} that maximizes the conditional probability:

$$y^{MAP} = \underset{h \in \mathcal{H}}{\operatorname{argmax}}\, P(h \mid \theta, x), \tag{1}$$

where h is a hypothesis in $\mathcal{H}$.

In other words, MAP decoding simply takes the estimated probability as the decision score.

However, considering all possible hypotheses in $\mathcal{H}$ is computationally intractable. Therefore, beam search is widely used as an efficient approximation of MAP decoding, balancing the trade-off between computational cost and search accuracy.

Increasing the beam size leads to searching for hypotheses with higher estimated probabilities. However, in practice, when the beam size exceeds 5 or 10, it often leads to a performance degradation instead (Tu et al., 2017; Koehn and Knowles, 2017). This phenomenon is known as the *beam search curse*, considered one of the six challenges of NMT (Koehn and Knowles, 2017).

2.2 MBR decoding

Unlike MAP decoding, which aims to find the highest estimated probability hypothesis, Minimum Bayes Risk (MBR) decoding seeks the hypothesis that minimizes the expected loss (or equivalently, maximizes the expected utility) with respect to the true translation distribution. Formally, let $\mathcal{Y}$ be the space consisting of all ideal translations. Then, for an ideal translation $y \in \mathcal{Y}$, MBR decoding selects the hypothesis y^{MBR} that has the highest MBR decision score $score_h^{MBR}$:

$$y^{MBR} = \underset{h \in \mathcal{H}}{\operatorname{argmax}}\, score_h^{MBR} \tag{2}$$

$$= \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\left[L(y, h) \mid \theta, x\right]. \tag{3}$$

In practice, it is common to use a utility function, correlated to human evaluation results, such as BLEU or COMET, as an alternative to the loss function $L(\cdot, \cdot)$. Thus, the purpose of MBR decoding is actually to select the hypothesis with the maximum expected utility. However, the hypothesis space is usually too large, hence it is impractical to traverse all the hypotheses to find a translation that satisfies the above conditions. Therefore, a

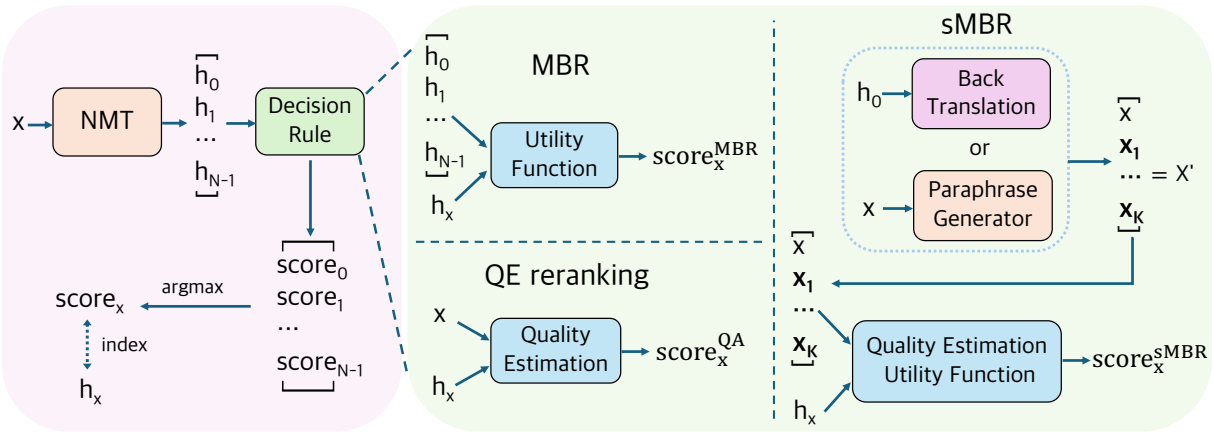

Figure 2: Overview of decoding methods in NMT. The diagram illustrates the process for MBR decoding, QE reranking, and the proposed sMBR decoding. It also shows two practices of sMBR: sMBR-BT and sMBR-PP. The figure demonstrates how the score used for selecting the final hypothesis is computed for each method.

set of hypotheses C from the hypothesis space $\mathcal{H}$, called "candidate hypotheses", is often used as a representative of the whole hypothesis space. Formally, for a given utility function $u(\cdot,\cdot)$, the MBR decoding objective can be reformulated as:

$$y^{MBR} = \underset{h \in \mathcal{C}}{\operatorname{argmax}} \, \mathbb{E}\left[u(y,h) \mid \theta, x\right]. \tag{4}$$

Since the ideal translations $\mathcal{Y}$ are often inaccessible in practice, approximations are needed. A common practice is to use a set of hypotheses, which are assumed to be equally probable, from the hypothesis space, called "support hypotheses", as an approximation to ideal translations.

Formally, let $S \subseteq \mathcal{H}$ be a set of support hypotheses and μ_u be the expected utility $\mathbb{E}\left[u(y,h) \mid \theta, x\right]$, then its approximation $\hat{\mu}_u(h;S)$ is defined as:

$$\hat{\mu}_u(h;S) := \frac{1}{M}\sum_{j=1}^{M} u(h_s^{(j)}, h), \tag{5}$$

where $h \in C$, $h_s^{(n)} \in S$, and $M = |S|$.

A common practice is to obtain the same set of hypotheses for both C and S, which can be obtained through, for example, beam search or ancestral sampling[2]. In other words, a common practice in MBR decoding is that the expected utility for a chosen hypothesis is approximated by averaging it with the utilities of other hypotheses.

3 Method

In this section, we first establish a connection between QE reranking and MBR decoding, showing that QE reranking can be viewed as a special case of MBR decoding. We then introduce our proposed approach, source-based MBR (sMBR) decoding, as a generalization of QE reranking. Finally, we introduce two practices of sMBR decoding: back-translation-based (sMBR-BT) and paraphrasing-based (sMBR-PP). See Figure 2 for a quick overview of how our approach to sMBR differs from MBR decoding and QE reranking.

3.1 QE reranking: A Closer Look

Quality estimation (QE) is a task that aims to assess the quality of a translated sentence without reference translations. Recently, QE models have been employed to develop a new decoding method called QE reranking (Fernandes et al., 2022), which leverages QE models to rerank the candidate hypotheses. The main idea behind QE reranking is to select the hypothesis with the highest estimated quality with the QE model, rather than relying on the estimated probability.

Formally, for a source x, a candidate hypothesis from the hypothesis space $h \in C$, and QE function f_{QE}, QE reranking aims at finding a y^{QE} that has the highest QE decision score $score_h^{QE}$:

$$y^{QE} = \underset{h \in C}{\operatorname{argmax}} \, score_h^{QE} \tag{6}$$

$$= \underset{h \in C}{\operatorname{argmax}} \, f_{QE}(x,h). \tag{7}$$

This approach can be viewed as a special case of MBR decoding. In this context, the quality estimation function $f_{QE}(x,h)$ serves as a utility function that estimates the quality of a hypothesis h given a source sentence x. This function acts as a proxy for

[2]In the context of NMT, ancestral sampling refers to sampling from the entire vocabulary without any pruning.

the expected utility of the hypothesis h with respect to the set of ideal translations. The justification for this interpretation will be further elaborated in the following subsection.

3.2 sMBR

By establishing this connection between QE reranking and MBR decoding, we lay the foundation for our proposed method, source-based MBR (sMBR) decoding, which generalizes QE reranking by considering multiple sources to estimate the expected utility of each hypothesis.

We hypothesize that for a chosen hypothesis, using only the hypothesis and the source's quality estimates is a good approximation of the utility of that hypothesis with ideal translations.

Formally, $P(X|Y)$ denotes the backward posterior distribution, then for a reference-free quality-estimated utility function $u(\cdot,\cdot)$, sMBR looks for y^{sMBR} that has the highest sMBR decision score $score_h^{sMBR}$ in the set of candidate hypotheses C:

$$y^{sMBR} = \operatorname*{argmax}_{h \in C} score_h^{sMBR} \tag{8}$$

$$= \operatorname*{argmax}_{h \in C} \mathbb{E}\left[u(x', h) \mid \theta\right], \tag{9}$$

where $x' \sim P(X|Y)$.

Similarly to definition (5), the approximation $\hat{\mu}_s(h)$ of the above expectation can be computed in a similar way. Let X' be a sample that obeys $P(X|Y)$ and $|X'| = K$, then $\hat{\mu}_s(h) := \hat{\mu}_u(h; X')$.

It is worth noting that QE reranking is a special case of sMBR when $K = 1$, i.e., when only the original source is used. Unlike QE reranking, sMBR considers multiple sources ($K > 1$), which are intended to serve as a more diverse and representative approximation of the ideal translations.

3.3 sMBR-BT and sMBR-PP

Next, we introduce two practices for sMBR. Since there is usually only one source available, in order to instantiate sMBR, we need to obtain methods for multiple sources.

Since x' comes from the back-translation distribution $P(X|Y)$, we can simply use a back-translation model to generate multiple sources.

Specifically, for an original source x, we first generated a translation h using the forward translation model and then used h as input to the back-translation model to generate K synthetic sources $\{x_1, x_2, \ldots, x_k\}$. We then use the set of sources $X' = \{x, x_1, x_2, \ldots, x_k\}$ for sMBR decoding. We call this **sMBR-BT**. Note that in sMBR-BT, the input to the back-translation model is a single hypothesis, where we simply use the highest estimated probability one. Then, we obtain K synthetic sources by beam search with beam size K.

In addition to back-translation-based sMBR, we introduce another practice: paraphrasing-based. We hypothesize that the paraphrase of the original source x also obeys the distribution $P(X|Y)$. We use a paraphrase generator to generate K paraphrases of x as the synthetic source $\{x_1, x_2, \ldots, x_k\}$, and then the set $X' = \{x, x_1, x_2, \ldots, x_k\}$ is used for the sMBR decoding. In other words, the key idea of this approach is to directly model the mapping of the original source to the synthetic source from the distribution $P(X|Y)$. We call this **sMBR-PP**.

4 Experiments

4.1 Setup

In this subsection, we present the details of NMT systems, decoding methods, and evaluation.

4.1.1 Data and models for NMT

We consider two setups: the classic (encoder-decoder based transformer model trained from scratch on parallel corpora) and the LLM setup.

Our experiments were performed on the English to German (En→De), English to Russian (En→Ru), and Chinese to English (Zh→En), as COMET and COMET-QE on them proved to be highly correlated with human judgments at the segment level (Rei et al., 2022a).

Classic setup The classic setup does not include the Zh→En due to computing resource limitations. For both En→De and En→Ru in the classic setup, we use *newstest2017-2019* as the development set and *newstest2020* as the test set.

We further consider both high-resource and low-resource sub-setups. For the high-resource setup, we use Facebook FAIR's WMT19 news translation task submission (Ng et al., 2019). For low resource setup, training data consists of 0.44M and 0.38M parallel sentences for En→De and En→Ru, respectively. These systems allow us to assess the performance in a data-constrained scenario. See Appendix D for more details.

LLM setup We use *TowerInstruct-13B* (Alves et al., 2024), a state-of-the-art LLM for translation-related tasks as NMT model. We prompt LLM to

perform zero-shot translation. See Appendix E for more details.

For all language pairs in the LLM setup, considering that the training data of *TowerInstruct-13B* already contains some earlier WMT test sets artificially (Alves et al., 2024), we use *generaltest2023* as the test set for each translation direction.

4.1.2 Decoding

We employ four approaches for hypothesis generation: beam search, ancestral sampling, top-k sampling, and epsilon sampling. Beam search has been widely used for MBR decoding in the past (Kumar and Byrne, 2004; Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017), while ancestral sampling has gained popularity in recent work (Eikema and Aziz, 2020; Freitag et al., 2022; Eikema and Aziz, 2022). We include top-k sampling and epsilon sampling since we find that they yield better performance than ancestral sampling.

Both for the classic and LLM setups , we evaluate the following decision rules:

- MAP: A widely used rule in NMT that selects the hypothesis with the highest estimated probability.
- MBR: MBR decoding based on COMET, using the *Unbabel/wmt22-comet-da*[3] model. It calculates decision scores for candidate hypotheses using the approach described in 2.2, and then select the highest score one.
- QE reranking: A special form of MBR, which calculates decision scores with the quality estimation model $f_{QE}(\cdot,\cdot)$. Since COMET does not support reference-free quality estimation, we use COMET-QE (*Unbabel/wmt22-cometkiwi-da*[4]) as the utility function.
- sMBR: Source-based MBR decoding. We evaluate its two practices: sMBR-BT and sMBR-PP. The same QE model used for QE reranking is employed as the utility function. As the back-translation model of sMBR-BT, in the classic setup, we again use Facebook FAIR's WMT19 news translation task submission in the high resource setup; in the low resource setup, we use the same model architecture and data as in the forward-translation model. For the LLM setup, we reused the LLM (*TowerInstruct-13B*) as a back-translation model. For sMBR-PP, we fine-tuned a unique *T5-large* (Chung et al., 2024) model as a paraphrase generator for each source language, which was fine-tuned on paraphrase data created through back-translation. We detail the sMBR-PP implementation in Appendix A.

Our baseline is a MAP based on beam search with a beam size of 5 both for the classic and LLM setups, since we found that larger beams do not lead to better performance. Except for the baseline, we use **400** candidate hypotheses for the classic setup unless otherwise specified, as we find that more hypotheses have limited gains in performance but result in higher costs.[5] For the LLM setup, we only use **128** hypotheses because generating more hypotheses leads to too much computing.

For MBR decoding, we tried two setups: (1) using the same set of hypotheses as candidate hypotheses and support hypotheses; and (2) using QE reranking to filter the set of support hypotheses to a smaller size that matches the size of the set of support hypotheses for sMBR. For sMBR-BT and sMBR-PP, we study case using 16 synthetic sources, as adding more did not yield further gains.

More details are provided in Appendix B.

4.1.3 Evaluation metrics

We use automatic evaluation metrics, including BLEU (Post, 2018), XCOMET (XCOMET-XXL) (Guerreiro et al., 2023), and MetricX (MetricX-XXL) (Juraska et al., 2023) to evaluate different methods. Our choice of XCOMET and MetricX is motivated by two key factors: 1) they are state-of-the-art neural metrics (Freitag et al., 2023b); 2) they were trained on Multidimensional Quality Metrics (MQM) data (Guerreiro et al., 2023; Juraska et al., 2023), while the COMET series, on which MBR and sMBR were directly optimized, were trained only on Direct Assessments (DA) data (Rei et al., 2022a). Given the limited correlation between MQM and DA (Freitag et al., 2021), we expect XCOMET and MetricX to provide a more independent assessment, as they are less likely to be biased towards the COMET-optimized MBR and sMBR methods. We perform significance tests using paired bootstrap resampling (Koehn, 2004).

[3] huggingface.co/Unbabel/wmt22-comet-da
[4] huggingface.co/Unbabel/wmt22-cometkiwi-da

Decoding method			En→De			En→Ru		
w/ beam search	N	M	BLEU↑	XCOMET↑	MetricX↓	BLEU↑	XCOMET↑	MetricX↓
High Resource (55.4M and 52.0M training data)								
MAP	5	–	36.48	93.64	1.40	24.62	86.63	2.07
MBR	5	5	36.52	93.70	1.40	24.38	86.97	1.97
MAP	400	–	**36.30**††	93.70	1.44	**24.54**††	86.89	2.07
QE reranking	400	1	35.32	94.85	1.12	23.00	90.29	1.47
MBR	400	400	35.87††	94.29	1.29	23.95	88.44	1.71
MBR	400	17	36.39††	94.21	1.29	24.44††	88.64	1.73
sMBR-BT	400	17	34.75	94.46	1.26	22.81	89.16	1.62
sMBR-PP	400	17	35.30	**95.19**††	**1.05**††	23.07	**90.55**†	**1.40**††
Low Resource (0.44M and 0.38M training data)								
MAP	5	–	23.93	78.66	4.84	17.43	67.54	5.73
MBR	5	5	24.05	79.48	4.70	17.33	68.62	5.46
MAP	400	–	23.89	79.06	4.74	17.51	68.02	5.56
QE reranking	400	1	25.48	85.21	3.23	18.01	76.48	3.79
MBR	400	400	24.94	83.07	3.71	17.87	73.21	4.37
MBR	400	17	25.10	82.52	3.90	**18.04**	72.82	4.58
sMBR-BT	400	17	24.79	81.94	3.97	17.84	75.26	3.94
sMBR-PP	400	17	**25.57**	**85.36**†	**3.15**††	18.03	**76.62**†	**3.73**†

Table 1: Compares decision rules in the classic setup. N and M indicate the number of candidate hypotheses and supportive hypotheses, respectively. For sMBR, we used $M = 17$ support hypotheses (1 original source + 16 synthetic sources). We performed paired bootstrap resampling; † and †† indicate significantly better than QE reranking within groups ($p < 0.05$ and $p < 0.01$, respectively; Multiple testing correction is not applied). The best in each group is marked in bold.

4.2 Results

In this subsection, we analyze the performance by observing the results of the automatic evaluation metrics for both the classic and LLM setups. Due to space constraints, we show in the main text only the results of beam search based results in the classic setup and epsilon sampling based results in the LLM setup; other results are included in the Appendix C.

Classic setup Table 1 highlights the effectiveness of sMBR decoding in the classic setup with beam search. Regarding XCOMET and MetricX, sMBR-PP significantly outperforms QE reranking, proving the validity of our extension to QE reranking. The results of the experiments based on sampling methods are shown in Appendix C, where similar gains to those based on beam search can be observed. The sMBR-PP outperforms the standard MBR on the neural metric, regardless of whether the standard MBR can use the full 400 support hypotheses or only 17. Thus, we conclude that the sMBR-PP outperforms QE reranking and the standard MBR in the classic setup.

LLM setup Table 2 shows the results of epsilon sampling in the LLM setup. We observe that sMBR-PP can still significantly outperform QE reranking regarding XCOMET and MetricX. The XCOMET and MetricX of sMBR-PP are comparable to the standard MBR and sometimes outperform standard MBR. As with the results of the classic setup, a gain similar to that based on epsilon sampling can be observed in Appendix C. Thus, we conclude that sMBR-PP still significantly outperforms QE reranking and is competitive with the standard MBR in the LLM setup.

The performance of sMBR-PP relative to standard MBR differs between the two setups. The standard MBR shows performance similar to that of the sMBR-PP in the LLM setup. We believe this is due to the better quality of the support hypothesis generated by LLM, which leads to a higher approximation accuracy in eq. 5.

Compared to QE reranking, sMBR-BT shows gains regarding XCOMET and MetricX in the LLM setup, but even lower metrics than QE rerank-

[5]Appendix G shows the impact of candidate hypotheses number on the metrics.

Decoding method			En→De			Zh→En		
	N	M	BLEU↑	XCOMET↑	MetricX↓	BLEU↑	XCOMET↑	MetricX↓
w/ epsilon sampling				***TowerInstruct-13B***				
MAP	5	–	30.24	86.06	3.18	20.73	88.15	2.56
MBR	5	5	28.10	87.30	2.65	19.74	88.96	2.36
MAP	128	–	**32.64**††	86.43	3.15	23.12††	89.14	2.42
QE reranking	128	1	29.40	88.76	2.41	19.88	90.64	2.07
MBR	128	128	29.84	89.19†	2.36	22.01††	90.39	2.06
MBR	128	17	31.93††	88.83	2.63	**23.34**††	90.43	2.12
sMBR-BT	128	17	28.73	89.04	**2.34**	19.77	90.38	2.06
sMBR-PP	128	17	27.19	**89.47**††	2.36	19.87	**90.70**†	**1.95**††

Table 2: Compares decision rules in the LLM setup. The meaning of table elements is the same as Table 1.

M	1	6	11	17	33
XCOMET ↑	94.85	95.05	95.15	95.19	95.19
MetricX ↓	1.12	1.08	1.07	1.05	1.06

Table 3: Impact of increasing sources for sMBR-PP: The number of sources is positively correlated with the evaluation metrics. $M = 1 + K$, i.e., (# of the original source) + (# of synthetic sources). Candidate hypotheses were generated by beam search. When $M = 1$, sMBR-PP is QE reranking.

	Self-BLEU↓	Semantic Similarity↑
sMBR-BT	72.31	94.74
sMBR-PP	51.34	94.79

Table 4: Analyzing of synthetic sources: analyzed on the En→De *newstest2020*, high resource. Lower Self-BLEU means richer surface diversity; higher semantic similarity means closer semantics to the original source.

ing in the classic setup. We investigate this reason in section 5.2.

5 Discussion

In this section, we discuss the effectiveness of sMBR-PP and the mechanism behind it through some experiments. Due to space constraints, we discuss the efficiency of sMBR-PP in Appendix F, where we show that sMBR-PP is much slower than QE reranking and MBR with simple optimization.

5.1 Effects of increasing sources

Since sMBR is an extension to QE reranking by increasing the number of sources, we first investigate the impact of increasing the number of sources on the performance of sMBR-PP. We focus on the En→De high resource setup of the classic setup and evaluate with neural metrics. Table 3 presents the results, demonstrating a positive correlation between source number and evaluation metrics. This observation again shows that our extensions to QE reranking are effective. In addition, the increase in synthesis sources from 16 to 32 does not result in further gains, which we hypothesize is due to the inability of the paraphrase generator to achieve the generation of up to 32 generative high-quality synthesis sources.

5.2 Qualities of synthetic sources

To understand the properties of the synthetic sources of sMBR, we analyzed them in terms of surface diversity and semantic similarity with the original source. Surface diversity was measured using Self-BLEU (Zhu et al., 2018), while semantic similarity was assessed by cosine distance between sentence embeddings[6].

The results presented in Table 4 reveal that the synthetic sources generated by sMBR-PP exhibit much lower Self-BLEU scores compared to those produced by sMBR-BT, indicating greater surface diversity. On the other hand, the scores of the two in terms of semantic similarity are close, implying that both generated synthetic sources do not deviate too much from the original source's semantics. We hypothesize that the poor performance of sMBR-BT in the classic setup can be attributed to the limited surface diversity of its synthetic sources.

5.3 Why sMBR-PP works

The results shown in Section 4.2 demonstrate that sMBR-PP is significantly better than QE reranking

[6]We use huggingface.co/sentence-transformers/all-mpnet-base-v2

M	1	6	11	17
Ave. QE	86.13	85.82	85.78	85.80

Table 5: Average QE scores with the original source: QE scores are negatively correlated with source number. The analysis was performed in the En→De high resource setup on sMBR-PP. Candidate hypotheses were generated by beam search.

in terms of neural metrics. Next, we will discuss the mechanism behind producing these gains. We hypothesize that the QE model used in QE reranking is overly sensitive to the specific phrasing and structure of the original source, leading to an over-reliance on a single source that could negatively impact performance. In contrast, aggregating QE scores across multiple sources in sMBR decoding is expected to provide more robust QE. We observed that as the number of sources increases in sMBR decoding, the average QE score between the selected translations and the original source sentence decreases (Table 5). This suggests that sMBR decoding no longer relies solely on QE with respect to the original source. We conjecture that this is because sMBR decoding tends to select hypotheses that are more generally applicable to different source variants.

6 Related Work

MBR decoding has been used in speech recognition (Stolcke et al., 1997; Goel and Byrne, 2000), word alignment (Kumar and Byrne, 2002), and statistical machine translation (Kumar and Byrne, 2004; Tromble et al., 2008). Recently, some works have re-explored the application of MBR decoding in NMT and demonstrated promising results (Stahlberg et al., 2017; Shu and Nakayama, 2017; Eikema and Aziz, 2020; Eikema and Aziz, 2022). These works have shown that MBR decoding can help overcome some of the limitations of MAP.

In past work, MBR decoding is usually based on beam search to generate candidate hypotheses (Stahlberg et al., 2017; Shu and Nakayama, 2017). Recently, Eikema and Aziz (2020) proposed sampling-based MBR decoding and found that the samples from the model are faithful to the training data statistics, while the beam search is not. Freitag et al. (2022) further explored the impact of the generation method on the performance.

In terms of utility functions, past work has primarily used surface-based metrics such as BLEU and BEER (Stanojević and Sima'an, 2014). However, these metrics have limited correlation with human judgments (Mathur et al., 2020; Freitag et al., 2023b). Recently, a trend has been to combine advanced neural metrics with MBR decoding, such as COMET and BLEURT. These works demonstrate that neural metrics-based MBR can improve performance in human evaluations (Freitag et al., 2022; Fernandes et al., 2022; Freitag et al., 2023a). However, they are also limited by the high cost, as MBR decoding has a secondary cost for the number of candidate hypotheses. Eikema and Aziz (2022) investigated decoupling candidate and support hypotheses, enabling the exploration of more potential candidate hypotheses within a limited computational budget. In addition, some recent work has focused on improving the efficiency of the MBR decoding (Cheng and Vlachos, 2023; Vamvas and Sennrich, 2024; Deguchi et al., 2024).

On the other hand, some work has found that models used to assess the quality of NMT systems (i.e., quality estimation) can perform well even in the absence of a reference (Rei et al., 2021; Rei et al., 2022b; Rei et al., 2023). Fernandes et al. (2022) explored the direct use of quality estimation models as rerankers for NMT and showed promising results, referred to as QE reranking.

7 Conclusions and Future Work

In this work, we explored the connection between QE and MBR decoding, showing that QE reranking is a special case of MBR decoding. Inspired by this finding, we proposed source-based MBR (sMBR) decoding, the first practical method to solely rely on sources as "support hypotheses" in MBR decoding. Experimental results (Section 4.2) show that sMBR decoding outperforms QE reranking and the standard MBR decoding.

Despite its limitations, such as the challenge of generating synthetic sources, sMBR represents a significant step forward in MBR decoding. By breaking with the tradition of approximating true utility using only the average of utilities with respect to other hypotheses, sMBR opens up new possibilities for future research.

Future work will explore more efficient methods for generating synthetic sources and investigate broadening the boundaries of "support hypotheses" to include sentences in languages other than the source and target.

8 Limitations

While our proposed sMBR decoding approach shows promising results, it has some limitations.

Firstly, generating high-quality synthetic source sentences remains a challenge. We explored two methods based on back-translation and paraphrasing, but the back-translation approach did not consistently improve reranking performance. This suggests that further research is needed to identify more effective techniques for generating diverse, representative synthetic sources.

Secondly, another potential issue is that reranking methods that directly optimize evaluation metrics may "overfit" to those metrics, causing the metrics that are directly optimized to become unreliable. We mitigate this problem by using automatic metrics that are likely to have low correlation with the metrics that are directly optimized. However, since they still have common parts in the training data, this measure may not completely avoid the problem of unreliable metrics. Human evaluation can mitigate this issue but is costly and time-consuming.

Finally, we have only tested the proposed method in a limited number of translation directions and domains. However, not all language pairs have well-performing quality estimation models available. In the case of some language pairs, this may lead to a questioning of one of our basic hypothesis, i.e., the quality estimation model is a good proxy for the true utility. Therefore, the effectiveness of sMBR in a wider range of settings remains an open question.

Acknowledgements

References

Duarte M. Alves, José P. Pombal, Nuno M. Guerreiro, Pedro H. Martins, Joao Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, Jos'e G. C. de Souza, and André Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *ArXiv*, abs/2402.17733.

Frédéric Blain, Lucia Specia, and Pranava Madhyastha. 2017. Exploring hypotheses spaces in neural machine translation. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 282–298, Nagoya Japan.

Julius Cheng and Andreas Vlachos. 2023. Faster minimum Bayes risk decoding with confidence-based pruning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12473–12480, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. 2024. Centroid-based efficient minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11009–11018, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023a. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

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

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom

Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023b. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

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

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André Martins. 2023. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *ArXiv*, abs/2310.10482.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

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

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

Shankar Kumar and Bill 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.

Shankar Kumar and William Byrne. 2002. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Nathan Ng, 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. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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

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

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova,

Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 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). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Raphael Shu and Hideki Nakayama. 2017. Later-stage minimum bayes-risk decoding for neural machine translation. *CoRR*, abs/1704.03169.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Andreas Stolcke, Yochai Konig, and Mitch Weintraub. 1997. Explicit word error minimization in n-best list rescoring. In *EUROSPEECH*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3097–3103. AAAI Press.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Additional details of sMBR-PP

For sMBR-PP, we generally use the same paraphrase generator for both the classic and LLM setups for the same source language.

When the source is English, for the paraphrase generator used in sMBR-PP, its specific model is *google/flan-t5-large*[7]. This model is trained for instruction following and thus works out-of-the-box for paraphrase generation. However, we found its performance to be rather poor, thus we chose to fine-tune it.

The fine-tuning training data consists of a publicly available paraphrase generation dataset, PAWS (Zhang et al., 2019), concatenated with a dataset we created. The dataset we created is based on En-De's News-Commentary parallel corpus[8] and uses machine translation to create paraphrased sentences. Specifically, we first input German sentences from the parallel corpus into the De→En NMT model, and then paired its output with English sentences from the original parallel corpus to compose the samples in the paraphrase generation dataset. We use the De→En model from Facebook FAIR's WMT19 news translation task submission (Ng et al., 2019). We use a semantic similarity-based approach to estimate the quality of the dataset we created, and then filter out sentence pairs with low similarity. We use the *sentence-transformers/all-mpnet-base-v2*[9] model to compute the similarity between paraphrase pairs and filter out sentence pairs with a similarity of 0.88 or less. In the end, the training data for the model consisted of a total of about 339.2K paraphrased sentence pairs, of which about 317.4K came from the data we created and about 21.8K came from PAWS.

For the training of this model, we used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3e-4, weight decay of 0.0, and a batch size of 1536 examples, trained with fp32 full precision (This is because we found that the flan-t5 series is prone to training failure at fp16 precision). We set the maximum number of training epochs to 10. We randomly separate 3K sentence pairs from the dataset as the development set, and then select the checkpoints with the lowest loss on the development set.

When the source is Chinese, we follow a similar procedure as above. The difference is that the base model is *mT5-large* (Xue et al., 2020) and the training data only includes the dataset created by using *TowerInstruct-13B* to perform reverse translation on the News-Commentary parallel corpus. We use *lier007/xiaobu-embedding-v2*[10] to calculate the cosine similarity between the rephrase and the original sentence, and filter out samples with a similarity below 0.925.

In the inference phase of generating paraphrases, we used epsilon sampling (Hewitt et al., 2022) (epsilon = 0.02), as we found that this setup balances the diversity and quality of the synthesized sources well. Training and inference were done on a single NVIDIA H100.

B Additional details on decoding and training of low-resource NMT models

We completed training of the NMT low-resource model, and all decoding experiments on a machine with 4 NVIDIA RTX A6000. In the hypothesis generation phase, we used CTranslate2[11] to generate hypotheses because of its efficiency.

For the training of low-resource NMT models, we use the fairseq (Ott et al., 2019) tool. We use base size transformer (Vaswani et al., 2017) architectures with a dropout rate of 0.3. And train for a maximum of 100 epochs at full fp16 precision. we select the checkpoints with the highest BLEU on the development set. We use adam (Kingma and Ba, 2015) optimizer with an initial learning rate of 1e-3, weight decay of 1e-4, a warm-up step of 4000, and batch size is 1e5 tokens. We build vocabulary of size 32000 with Byte-Pair Encoding (Sennrich et al., 2016) using the sentencepiece (Kudo and Richardson, 2018) tool. The vocabulary is shared between source and target languages.

C Additional experimental results

In this section, we present additional experimental results that could not be included in the main text due to space constraints. For the classic setup, this section includes the results of the generation methods other than beam search; for the LLM setup, this section includes not only the results of the generation methods other than Epsilon sampling, but also the complete experimental results of En->Ru.

[7] huggingface.co/google/flan-t5-large
[8] data.statmt.org/news-commentary/v18.1/
[9] huggingface.co/sentence-transformers/all-mpnet-base-v2
[10] https://huggingface.co/lier007/xiaobu-embedding-v2
[11] https://github.com/OpenNMT/CTranslate2

In addition to beam search and epsilon sampling, we attempted to use top-k sampling and ancestral sampling to generate hypotheses. Unlike top-k sampling, ancestral sampling the entire vocabulary for each time step in autoregressive decoding without any pruning. The results of the experiments are shown in Table 6, 8, 7, 9, and 10. Compared to other generation methods, ancestral sampling performs poorly on both surface-based and neural metrics. Among the sampling-based methods, epsilon sampling performs best, which is consistent with the findings of Freitag et al., 2023a.

We used $k = 10$ for top-k sampling and epsilon $= 0.02$ for epsilon sampling (Freitag et al., 2023a). Due to implementation issues with some CUDA programming, we do not consider epsilon sampling with low resource setup.

In conclusion, similar to the experimental results based on beam search and epsilon sampling, the significantly boosted neural metrics demonstrate that sMBR-PP significantly outperforms QE reranking. However, improvements in neural metrics do not always lead to gains in surface-based metrics and even lead to deterioration compared to the baseline, especially when using sampling-based hypothesis generation. One possible explanation is that sampling leads to more diverse hypotheses, making it easier to generate candidates hypotheses that would lead to higher neural metrics but not favored by BLEU. Unfortunately, sMBR decoding does not consistently mitigate this issue compared to QE reranking, suggesting potential limitations in the utility functions.

D Details of the classic setup

For the high-resource setup, training data consists of 27.7M and 26.0M parallel sentences for En→De and En→Ru, respectively, augmented with an equal amount of back-translation sentences. We use a single model without ensembling or language model reranking to focus on the impact of the proposed methods. For the low resource setup, we train two base Transformer models using the News-Commentary dataset[12].

E Details of the LLM setup

We use the target language to prompt the model to perform zero-shot translation. We completed all decoding on a server with four NVIDIA H100s with bfloat16 precision. For sMBR-PP on Zh→En, we trained a mT5-based (Xue et al., 2020) model for the paraphrase generator. See Appendix A for details.

[12] data.statmt.org/news-commentary/v18.1/

F Efficiency comparison

In this section, we discuss the efficiency of sMBR-PP and compare it with QE reranking and MBR. For the latter, we compare the average decision time required for MBR, an optimized implementation of MBR (MBR-fast), and sMBR-PP to translate a single sentence. We ran each method five times on a single NVIDIA H100 with batch size 256 examples and then report the means.

As expected, the decision time required for sMBR-PP to translate a sentence is much larger than that of QE reranking. Specifically, the decision time of sMBR-PP consists of two parts: the time required to generate the synthetic source and the time required to calculate the quality-aware utility function. In fact, we find that the time required to generate the synthetic source is only a small part of the overall decision time, which is about 0.13 seconds for each sentence, while the large number of quality-aware utility functions requires 3.56 seconds. In contrast, the decision time for QE reranking is 0.21 seconds per sentence, which is much faster than sMBR-PP.

Obviously, compared to QE reranking, sMBR-PP uses the quality perception utility function times (number of synthetic sources $+1$).

Next, we compare the decision time of sMBR-PP with that of the standard MBR. In MBR, the COMET model can be decomposed into a sentence encoder f_{emb} for computing sentence embeddings, and a simple estimator $f_{est}(\cdot,\cdot,\cdot)$ based on a multi-layer perceptron. For a source x, candidate hypothesis h, and support hypothesis h_s, COMET-based MBR first computes the embeddings x^{emb}, h^{emb}, and h_s^{emb} using f_{emb}. Then, the approximated utility $\hat{\mu}_u(h;S)$ can then be computed as:

$$\hat{\mu}_u(h;S) = \frac{1}{M}\sum_{j=1}^{M} f_{est}(x^{emb}, {h_s^{(j)}}^{emb}, h^{emb}). \tag{10}$$

When $S = C$, the cost of computing utility for all candidate hypotheses in a naive MBR implementation is $\mathcal{O}(N^2)$, implying a quadratic cost for both f_{emb} and $f_{est}(\cdot,\cdot,\cdot)$.

However, MBR-fast optimizes embedding computation by recognizing that the embedding any

Decoding method			En→De			En→Ru		
w/ top-k sampling	N	M	**BLEU↑**	**XCOMET↑**	**MetricX↓**	**BLEU↑**	**XCOMET↑**	**MetricX↓**
High Resource (55.4M and 52.0M training data)								
MAP	5	–	28.23	91.88	1.89	20.63	84.25	2.35
MBR	5	5	27.21	92.40	1.60	19.17	85.51	2.01
MAP	400	–	**32.83**††	93.21	1.66	**24.03**††	86.37	2.07
QE reranking	400	1	29.07	94.70	1.11	19.15	90.81	1.33
MBR	400	400	30.47††	94.40	1.22	21.39††	88.94	1.54
MBR	400	17	32.75††	94.43	1.18	23.70††	89.41	1.56
sMBR-BT	400	17	28.22	94.23	1.36	19.16	88.89	1.56
sMBR-PP	400	17	27.64	**94.99**††	**1.03**††	18.93	**91.14**†	**1.25**††
Low Resource (0.44M and 0.38M training data)								
MAP	5	–	18.71	71.95	6.35	15.04	61.29	7.01
MBR	5	5	18.51	73.66	5.60	14.38	62.57	6.38
MAP	400	–	21.10	76.45	5.37	**16.80**††	66.57	5.93
QE reranking	400	1	21.19	84.23	3.32	15.96	74.50	3.93
MBR	400	400	21.13	82.29	3.70	16.07	71.25	4.38
MBR	400	17	**22.51**††	82.13	3.84	17.03††	71.50	4.60
sMBR-BT	400	17	20.24	80.27	4.22	15.28	72.94	4.17
sMBR-PP	400	17	20.91	**84.65**††	**3.21**†	15.75	**74.78**†	**3.86**†

Table 6: Compares sMBR with other decision rules for En→De and En→Ru in the **classic setup**. N and M indicate the number of candidate hypotheses and supportive hypotheses, respectively. For sMBR, we used $M = 17$ "support hypotheses" (1 original source + 16 synthetic sources). We performed paired bootstrap resampling; † and †† indicate significantly better than QE reranking within groups ($p < 0.05$ and $p < 0.01$, respectively; Multiple testing correction is not applied). The best in each group is marked in bold.

sentence in a triple (x, h, h_s) is independent of the other elements. By pre-computing sentence embeddings independently for all sources and hypotheses, MBR-fast avoids duplicate f_{emb} computations and reduces its cost to $\mathcal{O}(N)$ when $S = C$. The estimator $f_{est}(\cdot, \cdot, \cdot)$ still has a quadratic cost $\mathcal{O}(N^2)$ since the order of elements within the triple affects the output, but it is computationally cheaper compared to a f_{emb} consisting of multiple transformer blocks. Note that this optimization is not universal due to the fact that it takes advantage of the particular architecture of COMET.

In contrast, the COMET-QE model used in sMBR consists of an encoder f^{QE}_{emb} that takes the concatenated source and hypothesis as input and outputs their joint embedding, and an estimator f^{QE}_{est}. The joint embeddings must be computed separately for each source-hypothesis pair, resulting in a cost of $\mathcal{O}(K \times N)$ for both f^{QE}_{emb} and f^{QE}_{est}.

Table 11 shows the measurement results. sMBR is faster than the naive implementation of MBR because it uses a smaller number of support hypotheses. However, it is much slower than the optimized implementation of MBR due to the difficulty of further optimizing its utility function itself.

In summary, while sMBR-PP significantly improves translation quality compared to QE reranking and has competitive performance to MBR, there is still room for improving its efficiency to match optimized COMET-based MBR decoding.

	MBR	MBR-fast	sMBR-PP
Decision time	135.29 s	0.32 s	3.56 (+ 0.13) s

Table 11: Decision time for translating a sentence: measured on *newstest2020* in En→De. For sMBR-PP, the number in parentheses is the synthetic source generation time. The batch size is 256.

G Impact of the number of candidate hypotheses

We explored the impact of the number of candidate hypotheses on the evaluation metrics in an En→De high resource setting. Figure 3 shows the results. We find that 400 is an appropriate number, as more candidate hypotheses bring small performance gains and lead to higher costs.

Decoding method			**En→De**			**En→Ru**		
w/ ancestral sampling	N	M	**BLEU↑**	**XCOMET↑**	**MetricX↓**	**BLEU↑**	**XCOMET↑**	**MetricX↓**
			High Resource (55.4M and 52.0M training data)					
MAP	5	–	18.79	66.15	8.35	14.03	57.87	8.12
MBR	5	5	17.90	63.89	8.92	12.91	55.81	8.44
MAP	400	–	**26.60**††	88.71	2.89	**20.87**††	82.70	2.74
QE reranking	400	1	25.50	90.03	**2.38**	18.51	**84.96**	2.23
MBR	400	400	25.05	88.23	2.76	18.01	81.47	2.48
MBR	400	17	25.98†	89.68	2.48	20.13††	84.01	2.27
sMBR-BT	400	17	25.37	89.56	2.73	18.79	83.74	2.41
sMBR-PP	400	17	25.29	**90.33**††	**2.38**	18.15	**84.96**	**2.18**†
			Low Resource (0.44M and 0.38M training data)					
MAP	5	–	13.88	50.28	11.78	11.03	43.64	12.20
MBR	5	5	13.90	48.56	12.05	9.62	42.35	12.80
MAP	400	–	17.26	69.99	7.01	**16.52**††	66.55	5.87
QE reranking	400	1	17.81	**74.98**	5.76	15.68	75.11	3.82
MBR	400	400	17.95	71.06	6.45	16.07	71.25	4.38
MBR	400	17	**18.37**††	73.47	5.91	16.14††	71.25	4.38
sMBR-BT	400	17	17.45	72.32	6.36	15.21	72.65	4.17
sMBR-PP	400	17	17.69	74.97	**5.69**†	15.80	**75.15**	**3.80**

Table 7: Compares sMBR with other decision rules for En→De and En→Ru in the **classic setup**. N and M indicate the number of candidate hypotheses and supportive hypotheses, respectively. For sMBR, we used $M = 17$ "support hypotheses" (1 original source + 16 synthetic sources). We performed paired bootstrap resampling; † and †† indicate significantly better than QE reranking within groups ($p < 0.05$ and $p < 0.01$, respectively; Multiple testing correction is not applied). The best in each group is marked in bold.

Decoding method			**En→De**			**En→Ru**		
w/ epsilon sampling	N	M	**BLEU↑**	**XCOMET↑**	**MetricX↓**	**BLEU↑**	**XCOMET↑**	**MetricX↓**
			High Resource (55.4M and 52.0M training data)					
MAP	5	–	28.64	91.93	2.01	18.69	79.36	2.86
MBR	5	5	28.65	93.18	1.46	19.60	82.86	2.28
MAP	400	–	29.92	90.91	2.93	18.50	79.40	2.84
QE reranking	400	1	29.82	94.76	1.09	19.22	88.19	1.54
MBR	400	400	31.63††	94.58	1.17	**21.66**††	85.70	1.85
MBR	400	17	**33.42**††	94.11	1.44	20.47††	85.42	1.84
sMBR-BT	400	17	26.09	91.73	2.68	18.54	84.03	2.00
sMBR-PP	400	17	29.09	**95.10**††	**1.01**††	18.83	**88.52**†	**1.47**†

Table 8: Compares sMBR with other decision rules for En→De and En→Ru in the **classic setup** with **epsilon sampling**. N and M indicate the number of candidate hypotheses and supportive hypotheses, respectively. For sMBR, we used $M = 17$ "support hypotheses" (1 original source + 16 synthetic sources). We performed paired bootstrap resampling; † and †† indicate significantly better than QE reranking within groups ($p < 0.05$ and $p < 0.01$, respectively; Multiple testing correction is not applied). The best in each group is marked in bold.

Decoding method			En→De			Zh→En		
	N	M	BLEU↑	XCOMET↑	MetricX↓	BLEU↑	XCOMET↑	MetricX↓
w/ beam search				***TowerInstruct-13B***				
MAP	5	–	39.75	87.11	1.93	24.94	89.13	2.43
MBR	5	5	39.84	87.29	1.90	25.00	89.14	2.42
MAP	128	–	40.04	87.07	1.84	24.96	89.12	2.44
QE reranking	128	1	40.11	87.35	1.81	24.96	89.22	2.40
MBR	128	128	40.07	87.29	1.82	**25.00**	89.17	2.41
MBR	128	17	40.14	87.19	1.81	24.96	89.13	2.44
sMBR-BT	128	17	**40.18**	87.39	1.81	24.90	89.21	2.41
sMBR-PP	128	17	40.15	**87.45**†	**1.78**	24.93	**89.28**†	**2.39**
w/ top-k sampling								
MAP	5	–	28.12	84.85	3.31	19.94	88.14	2.62
MBR	5	5	25.52	86.12	2.69	18.52	88.82	2.44
MAP	128	–	**32.13**††	86.17	3.30	21.90††	88.56	2.52
QE reranking	128	1	27.56	88.09	2.53	19.07	90.10	2.12
MBR	128	128	28.28†	88.68††	**2.21**†	21.19††	90.23	2.08†
MBR	128	17	30.66††	88.76††	2.38	**22.40**††	**90.47**††	2.08†
sMBR-BT	128	17	27.13	88.48	2.49	18.87	89.82	2.10
sMBR-PP	128	17	25.20	**89.05**††	2.29†	18.85	90.19	**2.03**†
w/ ancestral sampling								
MAP	5	–	26.29	83.20	3.35	19.07	87.21	2.83
MBR	5	5	24.20	84.67	2.97	17.97	87.79	2.62
MAP	128	–	**30.81**††	85.76	3.24	**21.86**††	88.57	2.58
QE reranking	128	1	27.10	87.57	2.59	18.67	89.69	2.24
MBR	128	128	26.88	87.59	**2.40**	20.65††	89.65	2.19
MBR	128	17	28.64††	87.50	2.57	21.72††	**90.02**†	2.17†
sMBR-BT	128	17	26.78	87.79	2.57	18.40	89.41	**2.17**†
sMBR-PP	128	17	25.78	**87.91**†	2.44	18.29	89.70	2.18†
w/ epsilon sampling								
MAP	5	–	30.24	86.06	3.18	20.73	88.15	2.56
MBR	5	5	28.10	87.30	2.65	19.74	88.96	2.36
MAP	128	–	**32.64**††	86.43	3.15	23.12††	89.14	2.42
QE reranking	128	1	29.40	88.76	2.41	19.88	90.64	2.07
MBR	128	128	29.84	89.19†	2.36	22.01††	90.39	2.06
MBR	128	17	31.93††	88.83	2.63	**23.34**††	90.43	2.12
sMBR-BT	128	17	28.73	89.04	**2.34**	19.77	90.38	2.06
sMBR-PP	128	17	27.19	**89.47**††	2.36	19.87	**90.70**†	**1.95**††

Table 9: Compares sMBR with other decision rules for En→De and Zh→En in the **LLM setup**. N and M indicate the number of candidate hypotheses and supportive hypotheses, respectively. For sMBR, we used $M = 17$ "support hypotheses" (1 original source + 16 synthetic sources). We performed paired bootstrap resampling; † and †† indicate significantly better than QE reranking within groups ($p < 0.05$ and $p < 0.01$, respectively; Multiple testing correction is not applied). The best in each group is marked in bold.

Decoding method			En→Ru		
	N	M	BLEU↑	XCOMET↑	MetricX↓
w/ beam search			***TowerInstruct-13B***		
MAP	5	–	29.46	89.51	1.59
MBR	5	5	29.33	89.61	1.55
MAP	128	–	29.50†	89.58	1.58
QE reranking	128	1	29.33	89.65	1.56
MBR	128	128	29.38	89.69	1.54
MBR	128	17	**29.52**††	89.57	1.58
sMBR-BT	128	17	29.39	89.69	1.56
sMBR-PP	128	17	29.36	**89.77**†	**1.50**
w/ top-k sampling					
MAP	5	–	23.22	86.79	1.85
MBR	5	5	21.23	87.02	1.69
MAP	128	–	**28.07**††	89.30	1.61
QE reranking	128	1	22.02	91.37	1.18
MBR	128	128	24.24††	91.26	1.18
MBR	128	17	27.47††	91.36	1.19
sMBR-BT	128	17	21.27	90.80	1.29
sMBR-PP	128	17	20.24	**91.49**††	1.11
w/ ancestral sampling					
MAP	5	–	23.28	86.79	1.87
MBR	5	5	21.12	87.18	1.70
MAP	128	–	**28.32**††	89.27	1.60
QE reranking	128	1	21.35	90.96	1.20
MBR	128	128	23.77††	90.98	1.21
MBR	128	17	26.55††	**91.23**††	1.21
sMBR-BT	128	17	21.63	90.78	1.27
sMBR-PP	128	17	20.52	90.99	**1.13**†
w/ epsilon sampling					
MAP	5	–	26.19	88.53	1.68
MBR	5	5	23.92	89.56	1.42
MAP	128	–	**29.61**††	89.68	1.56
QE reranking	128	1	22.35	91.67	1.13
MBR	128	128	25.90††	91.33	1.16
MBR	128	17	28.94††	91.54	1.22
sMBR-BT	128	17	22.47	91.68	1.16
sMBR-PP	128	17	21.26	**92.12**††	**1.04**††

Table 10: Compares sMBR with other decision rules for **En→Ru**, the NMT model is ***TowerInstruct-13B***. N and M indicate the number of candidate hypotheses and supportive hypotheses, respectively. For sMBR, we used $M = 17$ support hypotheses (1 original source + 16 synthetic sources). We performed paired bootstrap resampling; † and †† indicate significantly better than QE reranking within groups ($p < 0.05$ and $p < 0.01$, respectively; Multiple testing correction is not applied). The best in each group is marked in bold.

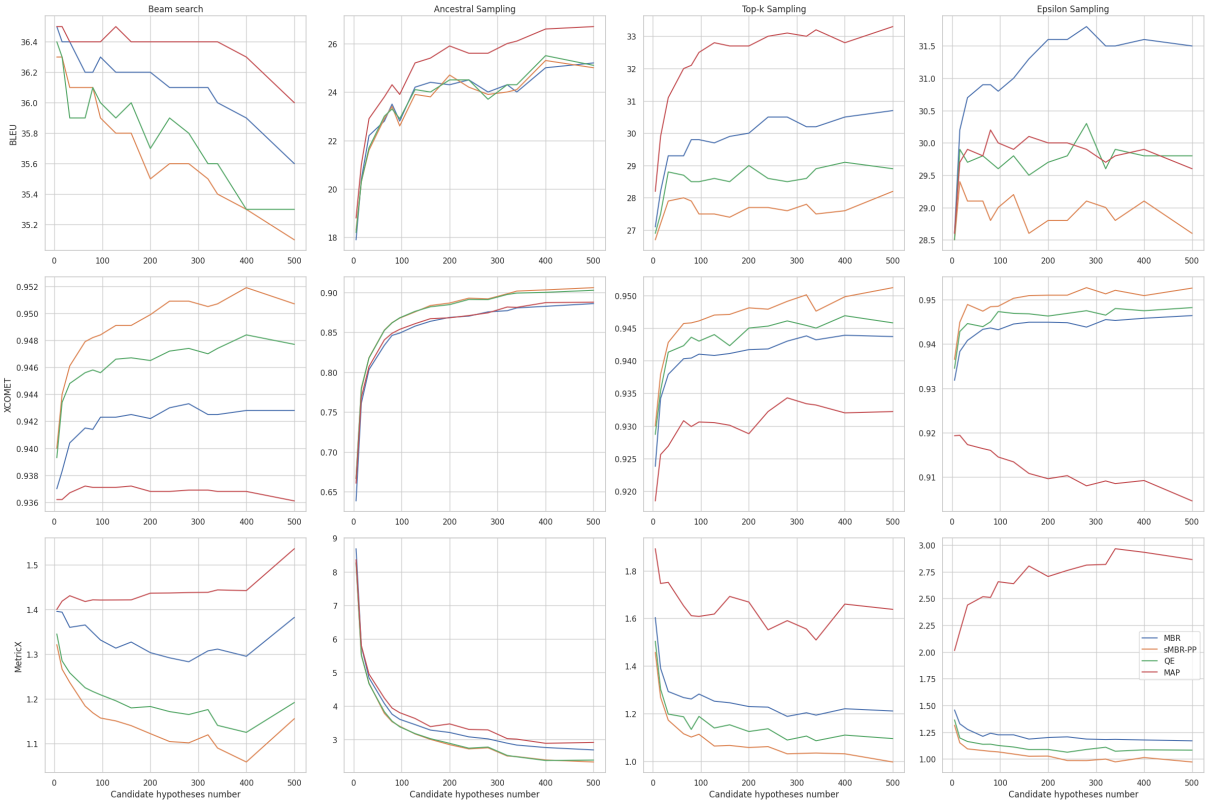

Figure 3: Impact of the number of candidate hypotheses on the evaluation metrics: in the En→De high resource setting. The horizontal axis indicates the number of candidate hypotheses and the vertical axis indicates the evaluation indicators.