High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics

Markus Freitag, David Grangier, Qijun Tan, Bowen Liang

Google Research, USA
{freitag, grangier, qijuntan, bowenl}@google.com

Abstract

In Neural Machine Translation, it is typically assumed that the sentence with the highest estimated probability should also be the translation with the highest quality as measured by humans. In this work, we question this assumption and show that model estimates and translation quality only vaguely correlate. We apply Minimum Bayes Risk (MBR) decoding on unbiased samples to optimize diverse automated metrics of translation quality as an alternative inference strategy to beam search. Instead of targeting the hypotheses with the highest model probability, MBR decoding extracts the hypotheses with the highest estimated quality. Our experiments show that the combination of a neural translation model with a neural reference-based metric, BLEURT, results in significant improvement in human evaluations. This improvement is obtained with translations different from classical beam-search output: These translations have much lower model likelihood and are less favored by surface metrics like BLEU.

1 Introduction

Neural sequence-to-sequence models constitute the state-of-the-art for machine translation. These models estimate the probability of a target sentence given a source sentence. At inference, it is commonplace to approximate the maximum-a-posteriori (MAP) hypothesis with beam search in order to output a sentence with (close to) the highest probability given the provided source.

This strategy assumes that the sentences with the highest estimated probabilities should also be the translations with the highest quality as measured by humans. This assumption can be questioned based on two observations: (i) Neural Machine Translations (NMTs) generated by beam search are ranked below human translations in professional evaluations (Freitag et al., 2021a)

while (ii) the NMT model itself considers human translations much less likely than its beam outputs (Ott et al., 2018). These observations clearly show that estimated probability and translation quality do not always correlate. An example is given in Table 1, where beam search generates a translation using mostly frequent words which results in inaccuracies. The two correct human translations contain infrequent words and phrases with low estimated probabilities based on the model.

These observations do not in themselves suggest an alternative to likelihood in selecting better hypotheses. For that, we look at recent progress in automated evaluation. Recently introduced utility metrics, such as BLEURT (Sellam et al., 2020a) or COMET (Rei et al., 2020), estimate human judgments $u(h, r)$ from a candidate translation h and a reference human translation r with a neural network. These learned metrics have shown higher correlation with human judgments compared with traditional metrics based on lexical overlap such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). BLEURT and COMET have also been shown by the WMT metric task (Freitag et al., 2021b) to perform better than YiSi (Lo, 2020), which measures overlap in a neural embedding space. BLEURT and COMET are able to evaluate hypotheses with different word choices, sentence structures, and lengths compared to the reference translations. Unlike overlap-based metrics like BLEU, these metrics do not necessarily prefer the most likely tokens to increase the chance of covering n-grams in the reference translations (Freitag et al., 2020). When comparing a model output h and an alternative human reference r', BLEU and BLEURT behave differently. While BLEU often estimates the quality of the model output h to be much higher than the alternative human translation r' ($B_{LEU}(h, r) > B_{LEU}(r', r)$), BLEURT and COMET typically prefer the human translation over the MT output ($B_{LEURT}(h, r) < B_{LEURT}(r', r)$).

811

system	translations	logP
source	Der Ausbruch sei ''mit Ansage'' gekommen.	
MAP/ beam	The outbreak came ''with announcement.''	−2.82
human-A	The outbreak occurred ''predictably.''	−18.1
human-B	The outbreak happened ''on cue.''	−18.74

Table 1: Example of De→En translations generated by NMT or humans. Human translations obtain a low model estimated probability (logP) as they do not generate the most frequent and direct translation.

This behavior generally agrees with professional raters (Toral, 2020).

These observations suggest that selecting model hypotheses likely to have a high quality score with respect to learned neural utility metrics should bring the quality of MT output closer to that of human translations. For that, we rely on Minimum Bayes Risk (MBR) decoding, in particular the sampling-based approximation recently introduced by Eikema and Aziz (2020). Sampling-based MBR starts with a set of unbiased samples drawn from an NMT model and finds the candidate which has the highest average utility when each hypothesis in the set is used as a pseudo-reference.

This MBR strategy has several potential pitfalls. First, the expectation of utility under the model distribution is used as a proxy to the expectation under the true underlying (human translator) distribution. This means that a high divergence between these two distributions will affect MBR (Pitfall 1: model quality). Second, the utility metric might be unreliable in areas of the space where it has not been evaluated (e.g., with low quality, low probability pseudo-references). This might cause its expectation to be very different from single point evaluations with high quality human references (Pitfall 2: utility validity over the reference space). Third, even if MBR discovers hypotheses with high utility with respect to actual human references, there is no guarantee that these hypotheses will receive high human judgments because these hypotheses are not necessarily close to the conditions for which the utility metrics have been designed (Pitfall 3: utility validity over the hypothesis space).

This paper evaluates MBR decoding for multiple utility functions and measures whether their predictions indeed improve the actual utility with respect to human references. We show that an NMT model based on the transformer-big architecture and BLEU, CHRF, YISI, and BLEURT successfully avoid Pitfalls 1 and 2. We also study the robustness of these conclusions with respect to the number of considered samples and model size. We then conduct a human evaluation of MBR hypotheses with high estimated utility according to different metrics to assess Pitfall 3. We show that MBR decoding using BLEU as a utility metric slightly improves over beam search decoding, even though the difference between these two translations are minor. In contrast, MBR using BLEURT as a utility metric generates translations further away from beam output. These translations are given significantly higher human quality ratings compared with beam search and the other MBR hypotheses.

Our contributions are:

- We are the first to use neural metrics—YISI and BLEURT—as utility functions during MBR decoding.

- We run a human evaluation with professional translators to assess the quality of MBR decode using different utilities.

- We show that MBR using BLEURT outperforms beam search decoding according to human judgments from experts.

- We further demonstrate that MBR decoding with BLEURT results in less likely translations which are lexically different from both beam output and MBR output relying on overlap-based utilities.

- We release all model hypotheses, candidate lists and human ratings as part of this paper.[1]

---

[1] https://www.kaggle.com/datasets/google /machine-translation-mbr-with-neural-metrics.

2 Related Work

Minimum Bayes Risk (MBR) decoding stems from statistical decision theory from the principal of maximization of expected utility (Bickel and Doksum, 1977; Berger, 1985). MBR has been applied to parsing (Goodman, 1996; Sima'an, 2003) and speech recognition (Stolcke et al., 1997; Goel and Byrne, 2000). The same idea was later applied to bilingual word alignment (Kumar and Byrne, 2002) and machine translation (Kumar and Byrne, 2004). MBR was used to maximize overlap metrics such as BLEU (Papineni et al., 2002) with statistical MT systems (Kumar and Byrne, 2004; Smith and Eisner, 2006; Tromble et al., 2008).

After the advent of neural machine translation (Sutskever et al., 2014), most methods relied on beam search to approximate MAP decoding (Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017). The question of optimizing utility metrics of interest such as BLEU was also explored. Approaches based on structured risk minimization (Edunov et al., 2018) or reinforcement learning (Bahdanau et al., 2017; Leblond et al., 2021) considered modifying the training procedure.

MBR decoding has recently gained attention in MT as a decision rule with the potential to overcome some of the biases of MAP decoding in NMT (Eikema and Aziz, 2020; Müller and Sennrich, 2021; Eikema and Aziz, 2021). While most prior work on MBR decoding for MT is based on k-best lists obtained via beam search, Eikema and Aziz (2020) proposed to use an approximation of MBR decoding based on unbiased sampling to overcome the shortcomings of MAP decoding. They demonstrated that samples from the NMT model are faithful to the training data statistics, while beam search is not. We adopt their sampling-based MBR decoding approximation in all our experiments.

The application of MBR to neural MT has focused on maximizing classical overlap-based metrics like BLEU, METEOR, CHRF, or BEER (Stanojević and Sima'an, 2014). Our work builds upon recent advances in the automatic evaluation of MT (Mathur et al., 2020), which has shown the emergence of learned utility metrics based on neural networks. We consider using neural metrics for MBR, which has not been done before. These metrics are neural networks that consider a pair of sentences (a hypothesis, and a reference)

or a triplet of sentences (a source, a hypothesis and a reference) and output a real-valued score estimating the quality of the hypothesis. They rely on pre-trained monolingual or multilingual neural language models. The first generation of neural utility metrics uses neural models to extract pre-trained sentence and word representations to compute distances indicative of semantic proximity, for example, BERTSCORE and YISI (Zhang et al., 2020; Lo, 2019). Later, a second generation of neural utilities proposed to fine-tune neural models on human judgments, either through regression or ranking tasks. These approaches, such as BLEURT and COMET (Sellam et al., 2020a; Rei et al., 2020), have shown better correlation with human judgments (Mathur et al., 2020).

3 Method

3.1 Minimum Bayes Risk Decoding

MBR relies on two essential components: a machine translation model and a utility metric. The translation model $P_{\text{model}}(y|x)$ estimates the probability of any target segment y given a source segment x. The utility metric $u(h, r)$ estimates quality of a candidate translation h given a reference translation r.

Given a set of hypotheses $\mathcal{H}$, we would like to select the best hypothesis according to its expected utility with respect to the distribution over human references in the space of all sequences Ω, namely,

$$
\begin{aligned}
h^{\text{best}} &= \underset{h \in \mathcal{H}}{\arg\max} \, \underset{r \sim P_{\text{human}}(\cdot|x)}{\mathbb{E}} \{u(h, r)\} \quad (1) \\
&= \underset{h \in \mathcal{H}}{\arg\max} \sum_{r \in \Omega} u(h, r) P_{\text{human}}(r|x).
\end{aligned}
$$

Because $P_{\text{human}}(r|x)$ is unknown, we need to rely on the model estimate instead, that is,

$$
h^{\text{model}} = \underset{h \in \mathcal{H}}{\arg\max} \sum_{y \in \Omega} u(h, y) P_{\text{model}}(y|x). \quad (2)
$$

This substitution assumes that the model provides a good approximation for the true underlying (human translation) distribution. As integrating over Ω, the space of all sequences, is intractable, MBR relies on a finite sample estimate by sampling a set of pseudo references $\mathcal{H}_{\text{model}}$ from $P_{\text{model}}(\cdot|x)$. This yields

$$
h^{\text{MBR}} = \underset{h \in \mathcal{H}}{\arg\max} \frac{1}{|\mathcal{H}_{\text{model}}|} \sum_{y \in \mathcal{H}_{\text{model}}} u(h, y). \quad (3)
$$

Commonly, one relies on the same set of model hypotheses for $\mathcal{H}$ (candidate pool) and $\mathcal{H}_{\text{model}}$ (pseudo-references), that is, $\mathcal{H} = \mathcal{H}_{\text{model}}$. In that case, growing $\mathcal{H}_{\text{model}}$ has two beneficial effects: A larger set provides a better approximation of the expected utility (reducing finite sample variance) while the maximum over a finite candidate pool obviously increases as the candidate pool grows.

Growing $\mathcal{H}_{\text{model}}$ is, however, computationally costly, both to obtain hypotheses and to evaluate their cross-utility. In all our experiments, we adopt the sampling-based approximation to MBR decoding (Eikema and Aziz, 2020) to generate a finite set of samples from a neural machine translation model. Eikema and Aziz (2020) showed that unbiased sampling provides a good approximation for the underlying model distribution. The cost of sampling is linear in the size of the set. Cross-utility can involve evaluating a large neural network as well and the cost of utility computation is generally quadratic in the size of the set. It is important to add that we generate independent samples, which implies that sentences with higher model probabilities have a higher chance to be drawn several times. By doing so and not deduping the candidate lists, we do not need to incorporate (again) the model probabilities during MBR decoding.

3.2 Utility Metrics

The automatic evaluation of machine translation is an active area of research (Mathur et al., 2020; Freitag et al., 2021b). MBR decoding centrally relies on a reference-based utility metric: Its goal is to identify a hypothesis with a high estimated utility (expectation under model distribution) with the hope that a high estimated utility translates into a high actual utility (with respect to a human reference), which itself should translate to a high human quality judgment. We experiment with utilities from different families of metrics:

Lexical Overlap: BLEU BLEU (Papineni et al., 2002) measures lexical overlap as the geometric mean of the precision of n-gram matches with $n \leq 4$ on the corpus level and adds a brevity penalty to penalize low recall hypotheses. As MBR decoding requires segment-level scores, we use add-one smoothed sentence-level BLEU (sBLEU) (Lin and Och, 2004). during MBR decoding as an approximation. We use Sacre-BLEU (Post, 2018) for reporting corpus-level BLEU scores.[2]

Lexical Overlap: CHRF We use CHRF (Popović, 2015) as an additional lexical overlap metric. CHRF uses character n-grams instead of word n-grams to compare the MT output with the reference. For CHRF we use the Sacre-BLEU `sentence_chrf` function (with default arguments[3]).

Embedding-based Overlap: YISI We also evaluate MBR decoding with neural utilities which has not been done before. We rely on Yisi-1-BERT (Lo, 2020) to represent first generation neural metrics, namely, metrics focusing on embedding-based overlap and not fine-tuned on human judgments. This metric relies on BERT (Devlin et al., 2019) to compute in-context word embeddings and then perform bi-directional alignments of n-gram matches in the embedding space to compute an F-score. For our experiments, we rely on base-cased BERT for English language evaluation and the multilingual model MBERT for other languages. We use our in-house reimplementation of YiSi.

Neural, Fine-tuned: BLEURT We rely on BLEURT to represent second generation neural metrics, that is, metrics not focusing on overlap but fine-tuned on human judgments instead. BLEURT is a regression model and relies on a learned embedding of the concatenation of the hypothesis and the reference translation. One of the strengths of BLEURT is that it can evaluate translations of different sentence structure, wording, and length in an unbiased fashion, as it is not focusing on any kind of overlap. This was one of our main motivations to revisit MBR decoding with neural metrics. We conducted experiments on two versions of BLEURT.

- **BLEURT v0.1**

  BLEURT v0.1 is a cased version of BLEURT (Sellam et al., 2020b) based on Rem-BERT (Chung et al., 2020). The model was pre-trained on more than 110 languages, and jointly fine-tuned on 13 target languages

---

[2] BLEU+case.mixed+lang.LANGPAIR-+numrefs.1 +smooth.exp+tok.13a-+version.1.5.0.

[3] chrF2+lang.LANGPAIR-+numchars.6+space.false-+version.1.5.0.

using the z-normalized WMT human evaluation data from 2015–2018.

- **BLEURT v0.2**

  BLEURT v0.2 is a joint model for all language pairs and is based on RemBERT. In addition to the fine-tuning data used for BLEURT v0.1, it also uses the WMT human evaluation data from 2019 and synthetic examples which consist of identities, alternative references, and random sentence pairs. Motivation for the latter was improved performance on very bad translations, a scenario frequently observed when scoring a candidate list during MBR decoding. Furthermore, instead of training BLEURT on the unbounded z-normalized scores, we manually scale them to a 0–1 range and clip the outliers.

4 Experimental Setup

4.1 Data and Model

We run experiments on two language pairs: English→German (En→De) and the reverse direction German→English (De→En) with models trained on WMT training data (Barrault et al., 2019). We use news-commentary-v15, paracrawl-v5.1, europarl-v10, and commoncrawl as training corpora with ~57 million training examples after filtering out noisy data with contrastive data selection, as proposed by Wang et al. (2018). We also remove sentences longer than 250 tokens and sentence pairs with a source/target ratio exceeding 1.5. We use newstest2019 as our dev set to pick checkpoints and newstest2021 (Barrault et al., 2021) as our test set. For newstest2021, we have two reference translations (Ref-C and Ref-D for En→De and Ref-A and Ref-B for De→En).

4.2 Model

We use the transformer implementation in *lingvo* (Shen et al., 2019), using a model similar to the transformer-big setting (Vaswani et al., 2017). The model has 6 encoder and 6 decoder layers, model dimension size of 1,024, hidden dimension size of 8,192, and the number of multi-attention heads is 16. Our models use a vocabulary of 32k subword units (Kudo and Richardson, 2018). We train the models until convergences for around 300,000 updates with a batch size of 43,000. We follow the suggestion of Eikema and Aziz (2020) and train our models without label smooth-

ing. This slightly drops accuracy by 0.5 BLEU points on both language pairs when compared with a model using label smoothing. We run beam search with beam size of 4 and length penalty as described in Equation 10 in Wu et al. (2016) using $\alpha = 0.5$. We do not use coverage penalty as this does not improve the results. For MBR decoding, we generate 1,000 unbiased samples for each source sentence.

4.3 Human Evaluation

We run two different human evaluations in this paper. For our main results, we run a human evaluation based on the Multidimensional Quality Metrics (MQM) methodology (Uszkoreit and Lommel, 2013) with professional translators. Freitag et al. (2021a) showed that this human evaluation is more reliable than typical scalar-value evaluation using crowd-workers. For ablation studies, we use a scalar-value human evaluation with professional translators similar to what is typically implemented in WMT as this human evaluation setup is cheaper and less time-consuming.

4.3.1 MQM

We hired 9 professional translators (4 for En→De and 5 for De→En) and measure translation quality with a document context version of MQM (Lommel et al., 2014) which mimics the setup proposed in Freitag et al. (2021a). This includes using the same error categories, severity levels, and error weighting schema. As suggested in the study, we weight each major error with 5 and each minor error with 1, except for minor punctuation errors, which get a score of 0.1. The final segment-level score is an average over scores from all annotators. We refer the reader to Freitag et al. (2021a) for the details on error categories and annotator instructions.

4.3.2 pSQM

In some of our ablation experiments, we conduct a human evaluation via **p**rofesional **S**calar **Q**uality **M**etric (Freitag et al., 2021a). This evaluation presents each source and translated segment from a document in a table row, asking professional translators to pick a rating from 0 through 6. The rater can scroll up or down to see all the other source/translation segments from the document. The final score for each of the systems is an average over their segment-level scores. We

Method		Automatic Evaluation						Model	Human Eval
		BLEU	sBLEU	CHRF	YISI	BL.1	BL.2	logP	MQM ↓
Human Transl.	Ref-D	31.5	31.6	60.9	84.7	37.1	75.6	−38.0	**0.388**[†]
Beam 4		34.3	34.2	62.5	85.3	26.8	71.6	−11.5	2.030
MBR	sBLEU	34.7	**<u>34.8</u>**	62.5	85.4	23.4	70.5	−11.2	1.855
	CHRF	34.2	34.3	**<u>64.1</u>**	85.7	25.8	71.4	−13.2	2.139
	YISI	34.2	34.2	62.8	**<u>86.0</u>**	26.4	71.6	−11.4	2.445
	BLEURT v0.1	29.2	29.4	60.0	84.3	**<u>50.0</u>**	77.1	−18.7	**1.571**[†]
	BLEURT v0.2	25.4	26.0	57.7	83.1	43.9	**<u>79.0</u>**	−24.4	**1.661**[†]

Table 2: Actual utility, log-likelihood (logP) and MQM score for different MBR methods and beam search on newstest2021 En→De computed with human reference *Ref-C*. All MQM results labeled with † are significantly better than beam search based on PERM-BOTH significance testing (Deutsch et al., 2021) with p = 0.001.

Method		Automatic Evaluation						Model	Human Eval
		BLEU	sBLEU	CHRF	YISI	BL.1	BL.2	logP	MQM ↓
Human Transl.	Ref-B	29.5	30.4	57.7.	82.8	38.3	75.4	−23.0	0.447
Beam 4		33.1	34.2	61.2	84.1	41.1	75.2	−6.1	0.345
MBR	sBLEU	33.3	**<u>34.7</u>**	61.1	84.1	40.1	75.0	−7.1	**0.323**
	CHRF	32.5	34.1	**<u>62.2</u>**	84.2	41.7	75.3	−8.0	0.380
	YISI	32.6	33.8	60.8	**<u>84.4</u>**	41.5	75.1	−7.7	**0.307**
	BLEURT v0.1	28.2	29.7	58.5	82.9	**<u>41.9</u>**	77.3	−11.8	**0.302**
	BLEURT v0.2	28.4	30.0	58.2	82.9	41.2	**<u>78.2</u>**	−12.2	**0.272**

Table 3: Actual utility of different MBR methods on newstest2021 De→En. Actual utility is computed with respect to reference A. This table is the equivalent of Table 2 for En→De.

run pSQM evaluations in our ablation studies for En→De with 3 professional translators.

5 Experimental Results

In this section, we discuss the main results of our study. First, we look into the automatic scores to investigate if MBR results in higher *actual* utility scores when *estimating* the expectation of the same utility. Second, we look into the human evaluation results to investigate how well the improvements in utility scores can transfer to human judgments.

5.1 Automatic Evaluation

MBR decoding chooses the translations with the highest estimated utility in a candidate list with the hope that this translation also gets a high actual utility score with respect to a human reference. We run MBR decoding with the utilities sBLEU, CHRF, YISI, BLEURT v0.1, and BLEURT v0.2. We verify

whether our NMT model is accurate enough for its candidate list to serve as a proxy for the human distribution. Experimental results with a 1,000 candidate list generated by unbiased sampling are summarized in Tables 2 and 3. For all utilities, the hypotheses with the highest *estimated* utility can generate a higher *actual* utility (bold, underlined numbers) when compared to the beam search output. This shows that the expectation of utility under the model distribution is a good proxy for the actual utility with respect to a human translation.

Interestingly, MBR with overlap-based metrics (sBLEU, CHRF, YISI) prefers high log likelihood hypotheses, with *logP* similar to MAP decodes. Rewarding reference overlap—even with an embedding distance in the case of YISI—favors the most common wording with the highest chance to match the surface form or embedding of a phrase in the reference translation. The BLEURT metrics, on the other hand, do not rely on overlap

evaluation and can reward less frequent translations. BLEURT selects alternative translations, which are not scored highly by overlap metrics like BLEU and which are not among the highest likelihood ($logP$) sentences according to the underlying NMT model.

5.2 Human Evaluation

Automatic metric results are encouraging but need to be confirmed with human assessments. We ran MQM-based human evaluations with professional translators for all MBR decoding outputs, beam search, and one human translation. MQM generates an interpretable error score (lower is better) and a score of 1 is equivalent to an average of one minor error per sentence, while a score of 5 is equivalent to an average of 1 major error. The MQM results in Tables 2 and 3 show that MBR decoding with BLEURT clearly outperforms (significantly, in the case of En→De) beam search decoding and MBR decoding with sBLEU, CHRF and YISI, demonstrating that when comparing different decoding strategies, model probability and actual human assessment poorly correlate. Interestingly, MBR using BLEU as the utility function is also better than beam search decoding, while CHRF and YISI are ranked below beam search for at least one language pair.

We have to mention that the human translation for En→De outperforms all machine generated translations. For De→En, the human translation is ranked behind all machine generated translations. We looked into the ratings and confirm that the human translation contains critical errors (this is in line with the official WMT21 human evaluation [Barrault et al., 2021]), showcasing how important it is to generate a good human translation when comparing MT with humans.

6 Ablation

We run ablation experiments to better understand the properties of MBR. We will mostly focus on experiments for English→German due to space and cost constraints.

6.1 Smaller Model

The candidate lists used by MBR in the main results section (Section 5) were generated by an NMT model using 375 million parameters similar to the transformer-big architecture. We raise the question if MBR using BLEURT v0.2 still avoids

Model		BLEU	BL.2	pSQM ↑
Transformer-big	Beam	34.3	71.6	4.47
	MBR-BL.2	25.4	79.0	4.67
Transformer-base	Beam	32.2	69.7	4.31
	MBR-BL.2	21.8	70.5	3.55
$\mathbb{E}$=base; max=big	MBR-BL.2	23.5	76.2	n/a
$\mathbb{E}$=big; max=base	MBR-BL.2	23.5	73.0	n/a

Table 4: Candidate list generation with either transformer-big or transformer-base model. The last column shows pSQM human evaluations results (higher is better). The results demonstrate that MBR needs a good model to outperform beam search.

Pitfall 1 and outperforms beam search when using a candidate list that is generated by a weaker model that is trained with 93 million parameters (model dimension size of 512, hidden dimension size of 2,048, and 8 transformer heads) similar to the transformer-base architecture. Experimental results can be seen in Table 4. We can see that the performance drops by 2 BLEU and 2 BLEURT points when comparing the beam hypotheses of the two different NMT models, indicating that the smaller model is indeed of lower quality.

Even though MBR outperforms beam decoding by 0.8 BLEURT points on the transformer-base model, the gap is much smaller than what we observe with the bigger model (7.4 BLEURT points). This already indicates that MBR is less effective on the smaller model, and the candidate list might not be good enough as a proxy for human references. We run a human evaluation comparing the two decoding algorithms on the small model and find that translation quality actually drops for the small setup when using MBR decoding. This shows that MBR requires a good quality candidate list to outperform beam search.

MBR uses the candidate list in two ways: (i) as a candidate pool from which it picks the hypothesis with the maximum estimated Bayes risk (*max step*) and (ii) as a list of pseudo-references to calculate the expected risk for each entry in the candidate pool ($\mathbb{E}$ *step*). It is not required that both operations use the same list. We run MBR decode using the candidate list of the small model on the $\mathbb{E}$ side and the candidate list of the larger model on the max side and vice versa. The BLEURT v0.2 results in the last two rows of Table 4 show that the candidate list generated by the smaller model

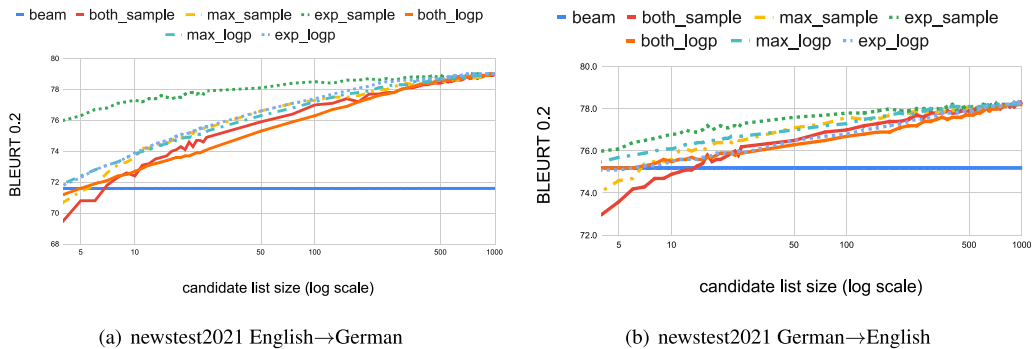(a) newstest2021 English→German

(b) newstest2021 German→English

Figure 1: Effect of different candidate list sizes on MBR decode with utility BLEURT v0.2 by either randomly sampling or choosing the candidates with the highest logP. We can reduce the number of candidates either only on the maximization or the expectation step alone or tight the two lists together. The graph shows that randomly subsampling the candidate list outperforms choosing candidates based on logP. Another evidence that we want the translations to steer away from the most probable translations. Further, pruning via sampling on the expectation side is more effective than reducing the candidate pool on the maximization side.

has a larger negative effect when used on the max operation compared to using it on the $\mathbb{E}$ side only. Overall, the results show that it is not sufficient to use the candidate list of the smaller model on either the $\mathbb{E}$ or the max operation.

6.2 Candidate List Size

All our MBR decoding results in the main results section (Section 5) rely on a candidate list size of 1,000. Generating 1,000 candidates and computing $1,000{\times}1,000{=}1M$ BLEURT scores for each source sentence is computationally costly and would not be practical at scale. We explore two different strategies to prune the candidate list via either (i) random sampling, or (ii) based on the model probabilities (logP). Similar to Section 6.1, we can apply the pruning strategies to either the $\mathbb{E}$ list, the max list or both lists. Experimental results can be seen in Figure 1.

There are three major insights: (i) if we prune both operations in MBR, randomly downsampling the candidate list size to a size of 8 (En→De) or 13 (De→En) already outperforms beam decoding based on BLEURT. (ii) We can aggressively sub-sample the candidate list used for the expectation ($\mathbb{E}$). For En→De, we observe major improvements over beam search decoding, shrinking the candidate list to 5 on the $\mathbb{E}$ side, resulting in only $5{\times}1,000{=}5,000$ BLEURT computations for a single source sentence. This confirms the findings of Section 6.1 that we rely more on the quality and size of the candidate pool on the maximization step than on the expectation. (iii) The results in Figure 1 suggest that the MBR

output most likely further improves when increasing the candidate list size beyond 1,000. This is different from beam search where accuracy gains are typically not achieved by growing beam size beyond a small number (< 10).

6.3 Oracle Experiments

We conduct oracle experiments to evaluate how the MBR hypotheses compare with selecting the best hypothesis with respect to a human reference. Given a human translation $\text{ref}_{\text{human}}$, we select the best hypothesis according to $\max_{h \in \mathcal{H}_{\text{model}}}$ BLEURT(h, $\text{ref}_{\text{human}}$) and report its BLEURT score. This assesses the gap between our decoding strategy and an oracle decision.

We consider two scenarios: selecting and evaluating a hypothesis with the *same* human reference, or selecting a hypothesis with a first reference before evaluating it with a second, *different* reference. The second method considers the selection reference and the evaluation reference as two independent samples of the human translation space. This avoids biasing selection to translation choices specific to the evaluation conditions.

Table 5 reports these results. In the *different* reference scenario, MBR performs better than the cross human selection, for example, selecting the best hypotheses with Ref-C yields a BLEURT score of 0.774 with Ref-D which is lower than 0.789, the BLEURT score of MBR with Ref-D. It is remarkable that the inter-translator variability in single reference automated evaluation causes more damage in oracle selection than the drop

818

| | | Actual | | | Model |
		Ref-C	Ref-D	mean	Est.
Human	Ref-C	0.963	0.757	0.860	0.680
	Ref-D	0.756	0.963	0.860	0.677
Oracle	Ref-C	0.827	0.774	0.801	0.709
	Ref-D	0.779	0.828	0.805	0.711
	Ref-C+D	0.810	0.815	0.813	0.719
MBR	BL.2	0.790	0.789	0.790	0.739

Table 5: Actual versus estimated BLEURT v0.2 of human references, oracle selection and MBR on Newstest2021 En→De. This table shows that BLEURT estimates that the oracle method is biased toward a specific human reference.

| | | Rank wrt BLEURT v0.2 Ref-C | | | | |
		p5	p25	p50	p75	p95
MAP		13	78	181	355	717
Oracle	Ref-D	1	4	18	78	327
MBR	BL.2	1	3	8	26	105

Table 6: Ranking (lower is better) of the top candidate selected by each decoding method, as ranked among the 1,000 candidates using BLEURT v0.2 (BL.2). The percentiles are calculated on the 1,002 test queries of Newstest2021 En→De. A smaller value indicates that the chosen candidate is also preferred by the actual Ref-C BL.2 metric. This table shows that MBR provides more stable quality estimates than single references.

due to swapping human references for model estimates.

Table 6 shows percentiles of the rankings of the selected translations among the candidate list as ranked by BLEURT v0.2 with respect to Ref-C. The median ranking (p50) of the MBR output is 8 out of 1,000, while the median raning of the MAP hypothesis is only 181. Interestingly, the MBR output even achieved higher ranking than the oracle candidate selected by Ref-D BLEURT v0.2 score, confirming the observation in Table 5 that model-estimated MBR provides more reliable quality estimates than selecting hypothesis with a single human reference translation.

6.4 Comparison to QE Metrics

Similar to reference-based metrics, reference-free –Quality Estimation (QE)–metrics have made huge improvements in the last years and show promising performance for some language pairs and test sets (Mathur et al., 2020). We pose the question whether a QE metric alone is sufficient to rerank the candidate list that we usually use for MBR decoding. The obvious advantage is that we only need N (N being the size of the candidate list), instead of $N \times N$ metric calculations. We present results with two different QE metrics: *COMET-QE-20* (Rei et al., 2020) and *COMET-QE-21* (Rei et al., 2021). These two metrics were the best QE metrics based on the two most recent WMT metric tasks (Mathur et al., 2020; Freitag et al., 2021b). Experimental results for En→De and De→En can be seen in Table 7.

Both reranking experiments show similar patterns: The QE-based reranking outputs outperform beam search and MBR with BLEURT v0.2 on both QE-metrics. Nevertheless, we can see that most reference-based metrics set the QE-based reranked output below both the beam search and the MBR output. When looking into the translations, we observed that some sentences in the QE-based reranking approach contain translations with crucial errors or the translation is unrelated to the source sentence. The human evaluation results in Table 7 confirm our impression that the reranked translations are of lower quality when compared to our MBR output or the beam search hypothesis. One potential reason of the underperforming reranking experiments can be the quality of the candidate list. As a reminder, the candidate list consists of unbiased samples drawn from the NMT model. Some of the samples are of bad quality and partially or entirely unrelated to the source sentence. While MBR compares the different samples with each other and penalized samples that are different to the other ones, the reranking approach solely relies on the QE metrics and does not have this safety mechanism.

7 How Different are Beam and MBR Hypotheses?

In Section 5, we observed that the model probabilities of the MBR output using BLEURT v0.2 is lower when compared to the beam search output. We want to further characterize the differences between these two decoding algorithms.

7.1 Cross BLEU

BLEU measures the lexical overlap between a hypothesis and a reference translation. It can also be

Method		Reference-based Evaluation					COMET-QE		Model	
		Bleu	Chrf	Yisi	BL.1	BL.2	2020	2021	logP	pSQM ↑
Human Transl.	Ref-D	31.5	60.9	84.7	37.1	75.6	39.7	11.4	−38.0	n/a
Beam 4		34.3	62.5	85.3	26.8	71.6	36.0	10.9	−11.5	4.47
MBR	Bleurt v0.2	25.4	57.7	83.1	43.9	**79.0**	43.4	10.8	−24.4	4.67
Reranking	COMET-QE-20	20.1	52.2	80.7	10.2	39.8	**60.6**	11.9	−31.7	4.05
	COMET-QE-21	15.2	44.3	76.9	−12.4	63.1	43.5	**12.8**	−32.8	3.44

Table 7: Reranking results with COMET-QE on Newstest2021 En→De. Actual utility is computed with respect to reference C. pSQM are human evaluation results on the same sentences (higher is better).

		Beam				MBR					Human	
		FB	O-W	UEdin	Ours	Bleu	Chrf	Yisi	BL.1	BL.2	Ref-C	Ref-D
Beam	Facebook		59.5	67.6	56.9	55.6	54.0	54.1	43.3	35.0	42.0	38.4
	Online-W	59.4		56.4	53.9	52.9	52.8	51.8	42.6	34.7	41.3	40.4
	UEdin	67.6	56.5		62.1	59.5	57.4	57.8	43.7	35.4	38.0	35.7
	Ours	57.0	54.0	62.2		77.0	69.8	71.9	50.6	39.8	34.3	33.9
MBR	Bleu	55.6	53.0	59.6	77.0		73.5	76.8	50.7	40.0	34.7	33.9
	Chrf	53.9	52.8	57.4	69.7	73.4		72.1	50.6	40.0	34.2	33.1
	Yisi	54.2	51.9	57.9	71.8	76.7	72.2		50.4	39.5	34.2	33.7
	BL.1	43.3	42.6	43.7	50.5	50.6	50.6	50.3		50.7	29.2	28.7
	BL.2	35.0	34.7	35.3	39.8	39.9	40.0	39.5	50.7		25.4	24.6
Human	Ref-C	42.0	41.4	38.0	34.3	34.6	34.3	34.1	29.2	25.5		31.4
	Ref-D	38.5	40.4	35.7	33.9	33.9	33.2	33.7	28.7	24.6	31.5	

Table 8: Overlap (cross-Bleu) between beam search output from different systems, our MBR hypotheses and human references on newstest2021 En→De. Lower cross-Bleu means lower word overlap between 2 translations. Facebook (Tran et al., 2021), Online-W, and UEdin (Chen et al., 2021) are submissions of the WMT21 evaluation campaign. Bleurt v0.1 and v0.2 are shortened BL.1, BL.2. We observe that the beam search output and MBR with Bleu, Chrf, and Yisi form a cluster of similar translations, while human references and the MBR output with Bleurt (in particular Bleurt v0.2) are different. Cross-Bleus lower than 50 are highlighted in green.

used to measure the lexical similarity of two alternative machine translations. In that case, Bleu does not assess translation quality but surface proximity between sentences.

Cross Bleu scores of our MBR outputs with our MAP decode and the best submissions in WMT21 can be seen in Table 8. Bleu scores lower than 50 are highlighted in the table. Our MAP hypothesis, the WMT21 submissions, and our MBR hypotheses using Bleu, Chrf, or Yisi have high crossBleu, which shows that they yield similar translations. The MBR output using Bleurt and the human translations have low cross-BLEU with all MAP hypotheses, which means that they use different

words and sentence structures. It is worth highlighting that the two human translations are as different from each other as they are to our MBR output using Bleurt.

7.2 MQM Error Categories

In addition to an overall quality score, MQM provides individual error labels with category and severity information. Table 9 reports major error counts for the most frequent categories, excluding categories with similar counts from beam and MBR. This table shows a clear advantage for the MBR output for four categories. Specifically, the number of errors in the category

	En→De		De→En	
	beam	MBR BL.2	beam	MBR BL.2
Terminology/Inappropriate for context	151	98	7	6
Accuracy/Mistranslation	70	58	33	23
Style/Awkward	66	46	10	5
Accuracy/Omission	18	7	0	0

Table 9: Number of major errors for selected categories for the MQM human evaluation.

Terminology/Inappropriate for context which is problematic for En→De shows a reduction of one third with MBR.

8 Conclusion

We explored an alternative to the commonly used beam search decoding algorithm typically used in NMT. We run the sampling-based approximation of Minimum Bayes Risk (MBR) decoding to optimize BLEU, CHRF, YISI, and BLEURT. Our experimental results showed that MBR decoding using BLEURT as utility function results in translations that significantly outperform beam search decoding based on expert-based human evaluation. We showed that the resulting translations are significantly different from both the beam search decode and MBR decoding output using one of the other overlap-based metrics as utility function, and have a lower model probability.

Acknowledgments

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

James O. Berger. 1985. *Statistical Decision Theory and Bayesian Analysis; 2nd ed.* Springer series in statistics, Springer, New York. https://doi.org/10.1007/978-1-4757-4286-2

Peter J. Bickel and Kjell A. Doksum. 1977. Mathematical statistics: Basic ideas and selected topics. *Holder-Day Series in Probability and Statistics, Holder-Day, San Francisco*.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German

and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *arXiv preprint arXiv:2104.00054*. https://doi.org/10.1162/tacl_a_00417

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. https://doi.org/10.18653/v1/N18-1033

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. https://doi.org/10.18653/v1/2020.coling-main.398

Bryan Eikema and Wilker Aziz. 2021. Sampling-based minimum Bayes risk decoding for neural machine translation.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. https://doi.org/10.1162/tacl_a_00437

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.5

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metric shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Vaibhava Goel and William J. Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135. https://doi.org/10.1006/csla.2000.0138

Joshua Goodman. 1996. Parsing algorithms and metrics. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 177–183, USA. Association for Computational Linguistics. https://doi.org/10.3115/981863.981887

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*. https://doi.org/10.18653/v1/D18-2012

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. https://doi.org/10.3115/1118693.1118712

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical

machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

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. https://doi.org/10.18653/v1/2021.emnlp-main.662

Chin-Yew Lin and Franz Josef Och. 2004. Orange: A method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507.

Chi-kiu Lo. 2019. YiSi—A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 455–463. https://doi.org/10.5565/rev/tradumatica.77

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*. https://doi.org/10.18653/v1/2021.acl-long.22

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR. https://doi.org/10.18653/v1/W18-6301

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. https://doi.org/10.3115/1073083.1073135

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

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*, Online. Association for Computational Linguistics,

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.704`

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur P. Parikh. 2020b. Learning to evaluate translation beyond english: Bleurt submissions to the WMT metrics 2020 shared task. *arXiv preprint arXiv: 2010.04297*.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X. Chen, Ye Jia, Anjuli Kannan, Tara N. Sainath, Yuan Cao, Chung-Cheng Chiu, Yanzhang He, Jan Chorowski, Smit Hinsu, Stella Laurenzo, James Qin, Orhan Firat, Wolfgang Macherey, Suyog Gupta, Ankur Bapna, Shuyuan Zhang, Ruoming Pang, Ron J. Weiss, Rohit Prabhavalkar, Qiao Liang, Benoit Jacob, Bowen Liang, HyoukJoong Lee, Ciprian Chelba, Sébastien Jean, Bo Li, Melvin Johnson, Rohan Anil, Rajat Tibrewal, Xiaobing Liu, Akiko Eriguchi, Navdeep Jaitly, Naveen Ari, Colin Cherry, Parisa Haghani, Otavio Good, Youlong Cheng, Raziel Alvarez, Isaac Caswell, Wei-Ning Hsu, Zongheng Yang, Kuan-Chieh Wang, Ekaterina Gonina, Katrin Tomanek, Ben Vanik, Zelin Wu, Llion Jones, Mike Schuster, Yanping Huang, Dehao Chen, Kazuki Irie, George Foster, John Richardson, Klaus Macherey, Antoine Bruguier, Heiga Zen, Colin Raffel, Shankar Kumar, Kanishka Rao, David Rybach, Matthew Murray, Vijayaditya Peddinti, Maxim Krikun, Michiel A. U. Bacchiani, Thomas B. Jablin, Rob Suderman, Ian Williams, Benjamin Lee, Deepti Bhatia, Justin Carlson, Semih Yavuz, Yu Zhang, Ian McGraw, Max Galkin, Qi Ge, Golan Pundak, Chad Whipkey, Todd Wang, Uri Alon, Dmitry Lepikhin, Ye Tian, Sara Sabour, William Chan, Shubham Toshniwal, Baohua Liao, Michael Nirschl, and Pat Rondon. 2019. Lingvo: A modular and scalable framework for sequence-to-sequence modeling. *CoRR*, abs/1902.08295.

Khalil Sima'an. 2003. On maximizing metrics for syntactic disambiguation. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 183–194, Nancy, France.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787–794. `https://doi.org/10.3115/1273073.1273174`

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. `https://doi.org/10.3115/v1/D14-1025`

Andreas Stolcke, Yochai Konig, and Mitchel Weintraub. 1997. Explicit word error minimization in n-best list rescoring. In *Fifth European Conference on Speech Communication and Technology*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisbon, Portugal. European Association for Machine Translation.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Roy Tromble, Shankar Kumar, Franz Josef 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. `https://doi.org/10.3115/1613715.1613792`

Hans Uszkoreit and Arle Lommel. 2013. Multidimensional quality metrics: A new unified paradigm for human and machine translation quality assessment. *Localization World, London*, pages 12–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics. `https://doi.org/10.18653/v1/W18-6314`

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.