DC-MBR: Distributional Cooling for Minimum Bayesian Risk Decoding

Jianhao Yan[1,2], **Jin Xu**[3] , **Fandong Meng**[4] , **Jie Zhou**[4], **Yue Zhang**[2,5]
[1]Zhejiang University, China
[2]School of Engineering, Westlake University
[3]Institute for Interdisciplinary Information Sciences, Tsinghua University
[4]Pattern Recognition Center, WeChat AI, Tencent, China
[5]Institute of Advanced Technology, Westlake Institute for Advanced Study
elliottyan37@gmail.com

Abstract

Minimum Bayesian Risk Decoding (MBR) emerges as a promising decoding algorithm in Neural Machine Translation. However, MBR performs poorly with label smoothing, which is surprising as label smoothing provides decent improvement with beam search and improves generality in various tasks. In this work, we show that the issue arises from the inconsistency of label smoothing on the token-level and sequence-level distributions. We demonstrate that even though label smoothing only causes a slight change in the token level, the sequence-level distribution is highly skewed. We coin the issue *autoregressive over-smoothness*. To address this issue, we propose a simple and effective method, Distributional Cooling MBR (DC-MBR), which manipulates the entropy of output distributions by tuning down the Softmax temperature. We theoretically prove the equivalence between the pre-tuning label smoothing factor and distributional cooling. Extensive experiments on NMT benchmarks validate that distributional cooling improves MBR in various settings.

Keywords: Minimum Bayesian Risk, Distributional Cooling, Machine Translation

1. Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017; Yan et al., 2020a) has witnessed significant progress in recent years. It models the conditional probability distribution of target language candidates given a source sentence by a using neural architecture model. Given a well-trained NMT model, the task of decoding is to select high-quality candidates according to the model distribution. The most commonly used decoding is Maximum-a-Posteriori decoding (MAP), which aims to find the most probable candidate (i.e., mode of the distribution). However, as revealed by recent studies (Stahlberg and Byrne, 2019; Yan et al., 2022), MAP decoding can be degenerate, suffering from hallucination or being even empty.

Minimum Bayesian Risk Decoding (MBR) (Kumar and Byrne, 2002; Eikema and Aziz, 2020) emerges as a promising alternative to MAP decoding, which seeks the candidate with the largest utility instead of the largest probability. Several advantages have been observed for MBR, such as being robust against domain shift (Müller and Sennrich, 2021) and avoiding beam search curse (Eikema and Aziz, 2022). With the help of neural metrics (Freitag et al., 2022) such as BLEURT (Sellam et al., 2020), MBR exceeds the *de facto* MAP decoding algorithm – beam search, achieving the state-of-the-art on several benchmarks.

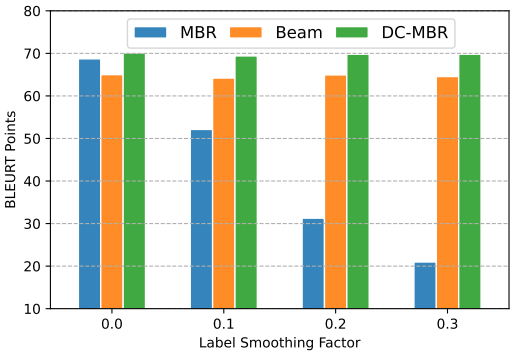

Figure 1: Translation quality against label smoothing factors. As the factor of label smoothing increases, beam search retains its performance while that of MBR drops drastically.

Despite the above promises, one crucial issue is identified for MBR but not yet solved in the literature: *MBR performs poorly with models trained with label smoothing (Eikema and Aziz, 2020).* We further find that the performance drops monotonically when increasing the label smoothing factor (see Figure 1, under the experiment settings in Section 4.1). It is counter-intuitive since label smoothing increases the generality of various tasks (Szegedy et al., 2016; Chorowski and Jaitly, 2017) and pro-

The early part of this work was done when Jianhao Yan was working at Pattern Recognition Center, Wechat AI, Tencent Inc, China.

LREC-COLING 2024, pages 4423–4437
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

vides steady improvements under the MAP setting in various NMT benchmarks (Vaswani et al., 2017; Chen et al., 2018).

We aim to investigate the root cause and address the issue that label smoothing benefits beam search but hurts MBR. As beam search makes use of token-level distribution and MBR relies on sequence-level distribution, we analyze the effect of label smoothing on the token-level and sequence-level distributions, finding that while label smoothing only slightly softens the token-level distribution, this effect makes the sequence-level distribution highly skewed with lots of low-quality candidates. We call the issue *autoregressive over-smoothness*, and quantify *autoregressive over-smoothness* using token-level entropy, finding that it correlates well with MBR's performance.

According to the above observations, we propose a conceptually simple and empirically effective approach, Distributional Cooling MBR (DC-MBR), which sharpens the model distributions by cooling down the Softmax temperature. It corrects the skewed sequence-level distribution and avoids sampling from candidates that the model is not confident with. We theoretically prove the equivalence between distributional cooling and label (un-)smoothing, validating that distributional cooling is a reverse process of label smoothing and can safely recover from the *autoregressive over-smoothness* without extra training.

We conduct experiments with two settings, bilingual NMT, under which we train Transformers (Vaswani et al., 2017) from scratch and evaluate them on three NMT benchmarks; and multilingual NMT, under which we evaluate with mBART-50 (Tang et al., 2020) on ten NMT benchmarks. Results show that DC-MBR mitigates the autoregressive over-smoothness and significantly outperforms the de facto standard unbiased setting of MBR. For instance, compared with naive MBR, DC-MBR improves up to 51.2 BLEURT points for the model trained with label smoothing.

In this paper, we take the label smoothing's incompatibility with MBR as a clue, dig into the hypothesis space, and propose a principled solution. Rather than simply avoiding label smoothing, MBR should be compatible with all kinds of models, especially in the recent tendency of LLMs (Touvron et al., 2023) where modifying training procedures is costly. Our proposed DC-MBR approach not only addresses the critical issue of label smoothing hurting MBR performance but also opens up new possibilities for improving the robustness and generalization of NMT models. By making MBR compatible with a wider range of training techniques, our work contributes to the development of more flexible and adaptable NMT systems that can better handle the challenges of real-world translation tasks. Furthermore, our findings shed light on the complex interplay between training methods and decoding strategies, paving the way for future research on optimizing NMT performance.[1]

2. Related Work

MT Decoding The dominant decoding method in NMT is Maximum-a-Posteriori (MAP) decoding, which seeks the hypothesis with the highest conditional probability. Among all MAP decoding methods, beam search is the *de facto* method in modern NMT systems. Many variants of beam search (Bahdanau et al., 2015; Wu et al., 2016; He et al., 2016; Yang et al., 2018; Murray and Chiang, 2018; Freitag and Al-Onaizan, 2017; Shu and Nakayama, 2018) are proposed to improve its performance. Other than beam search, exact decoding algorithms (Stahlberg and Byrne, 2019; Yan et al., 2022) use depth-first search to find the mode or top candidates of the whole candidate space. However, the computational cost of exact search hinders its applications.

Minimum Bayesian Risk Decoding (MBR), originated from SMT (Kumar and Byrne, 2002; Smith and Eisner, 2006; Tromble et al., 2008), recently emerges as the new alternative to MAP decoding algorithm in NMT. MBR selects the candidates with the highest utility, e.g., an evaluation metric, instead of the highest probability, which may avoid degenerate problems with MAP. Early attempts to incorporate MBR into NMT mainly use the k-best list obtained via beam search (Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017). Recently, Eikema and Aziz (2020) show that the model's sequence distribution provides a good approximation for human translation and proposes to approximate the hypothesis space and reference space of MBR by *unbiased* ancestral sampling. This *unbiased* sampling setting becomes the common practice of MBR and shows promising results. Müller and Sennrich (2021) show that MBR increases robustness against copy noise and domain shift. Eikema and Aziz (2022) demonstrate that MBR does not suffer from beam search curse (Koehn and Knowles, 2017a), i.e., better search always leads to better translations, and explores approximations for the expected utility. Freitag et al. (2022) propose to combine neural reference-based metric (i.e., BLEURT) as the utility function and demonstrate significant improvements. In this work, we take the inferior performance of MBR with label smoothing as a clue and propose a novel approach called distributional cooling. We demonstrate that, in contrast to the common *unbiased* setting, MBR can be further improved by

[1]The code can be found at https://github.com/ElliottYan/DC-MBR/tree/main.

cooling down the model distribution, effectively addressing the limitations of label smoothing in the context of MBR.

There is also work proposing the idea of desmoothing in decoding (Hewitt et al., 2022; Freitag et al., 2023), investigating the effectiveness of truncation methods. Compared with this line of work, distributional cooling is the more principled approach, clearly motivated by the inverse relation to label smoothing, while previous work does not explicitly discuss label smoothing.

Label smoothing First introduced by Szegedy et al. (2016), label smoothing is designed to improve the generality of neural models by replacing the one-hot targets with smoothed targets. It has shown to be effective in various NLP tasks (Szegedy et al., 2016; Chorowski and Jaitly, 2017; Pereyra et al., 2017), and provides a steady performance gain in machine translation (Vaswani et al., 2017; Chen et al., 2018). Müller et al. (2019) first study the effectiveness of label smoothing with beam search and attribute its effectiveness to better calibrating model predictions (Guo et al., 2017). However, label smoothing is not always helpful. Meister et al. (2020b) observe the case where higher entropy is detrimental to the performance of random sampling. Here we focus on label smoothing's negative effect on MBR and attribute the issue to the different behavior of label smoothing in sequence-level and token-level distribution. Further, we introduce distributional cooling that effectively resolves this issue.

3. Background

We take the standard Transformer (Vaswani et al., 2017) as the baseline, investigating label smoothing under the MAP and MBR decoding algorithms.

3.1. NMT and Label Smoothing

Given a model $f(\theta)$, Neural Machine Translation (NMT) predicts the conditional probability $P(y|x)$ of a target sentence y given a source sentence x, which can be factorised with an auto-regressive process:

$$P(y|x;\theta) = \prod_t P(y_t|y_{<t}, x; \theta). \tag{1}$$

We refer to $P(y_t|y_{<t}, x; \theta)$ and $P(y|x;\theta)$ as token-level distribution and sequence-level distribution, respectively. The token-level distribution $P(y_t|y_{<t}, x; \theta)$ is derived with a Softmax function,

$$o_t^i = f(y^i|y_{<t}, x; \theta), \tag{2}$$

$$P(y_t^i|y_{<t}, x; \theta) = \frac{\exp o^i}{\sum_j^{|V|} \exp o^j}, \tag{3}$$

where y^i is i-th token in the vocabulary V, and o represents the output logits. The widely used objective for training an NMT model is the label-smoothed cross-entropy loss, defined as,

$$\mathbf{L}_{ls} = -\sum_i Q_\lambda^i \cdot \log P(y_t^i). \tag{4}$$

Q_λ is the λ-smoothed target distribution. Its probability of the i-th token can be expressed as,

$$Q_\lambda^i = \begin{cases} 1-\lambda & \text{if } y^i \text{ is golden token} \\ \frac{\lambda}{|V|-1} & \text{otherwise} \end{cases}. \tag{5}$$

3.2. Decoding Algorithms

Given a model and an input, decoding algorithms select high-quality candidates from $P(y|x)$.

Maximum a Posteriori (MAP) The standard decoding algorithm in NMT is MAP decoding, which finds the candidate with the highest sequence probability (mode of the sequence distribution).

$$y^{\text{MAP}} = \text{argmax}_y P(y|x) \tag{6}$$

$$= \text{argmax}_{y_1,\cdots,y_T} \prod_t^T P(y_t|y_{<t}, x; \theta). \tag{7}$$

The exact solution of MAP is computationally costly due to NMT's exponentially large search space. Hence, practitioners turn to beam search, a decoding algorithm relies on greedy token selections.

Maximum Bayesian Risk (MBR) Recently, it has been shown that the mode of the model's sequence distribution (i.e., MAP's optimal solution) may be degenerate or even empty (Stahlberg and Byrne, 2019; Yan et al., 2022), which makes the mode a bad target. In contrast, MBR (Kumar and Byrne, 2002) chooses the candidate with the highest expected utilities:

$$y^{\text{MBR}} = \text{argmax}_{h \in \mathcal{Y}_h} \underbrace{\mathbb{E}_{r \in \mathcal{Y}_r}[u(h, r)|x, \theta]}_{=:\mu_u(h;x,\theta)}, \tag{8}$$

where the utility function u can be a certain evaluation metric measuring the similarity between a hypothesis h and a reference r. The hypothesis space $\mathcal{Y}_h$ and reference space $\mathcal{Y}_r$ are sets of all possible translations. Clearly, the above formulation is also intractable as both spaces are prohibitively large. Recently, Eikema and Aziz (2020) propose a sampling-based approach that approximates both spaces with the help of the model distribution. The authors argue that the model distribution is a good approximation for human translations. Specifically, their approach relies on finite candidates sampled from the model's distribution,

$$\mathcal{Y}_{\text{model}} \sim \prod_t P(y_t|y_{<t}, x; \theta), \tag{9}$$

and uses these candidates as both the pseudo references and hypotheses:

$$\hat{\mu}_u(h; x, \theta) := \frac{1}{N} \sum_{r}^{\mathcal{Y}_{\text{model}}} u(h, r), \quad (10)$$

$$\hat{y}^{\text{MBR}} = \text{argmax}_{h \in \mathcal{Y}_{\text{model}}} \hat{\mu}_u(h; x, \theta), \quad (11)$$

where $N = |\mathcal{Y}_{\text{model}}|$ is the number of candidates sampled. In practice, the choice of the utility function can be NMT n-gram matching metrics such as BLEU (Papineni et al., 2002), ChrF (Popović, 2015) or neural metrics such as BLEURT (Sellam et al., 2020).

4. Analyses

In this section, we will examine the cause of label smoothing negatively impacting the performance of Minimum Bayes Risk (MBR). We will also compare the effects of label smoothing on beam search, which has been shown to work well with it in previous studies (Szegedy et al., 2016; Chorowski and Jaitly, 2017; Pereyra et al., 2017; Vaswani et al., 2017). We will begin by outlining the details of our experimental setup.

4.1. Setup

For the bilingual setting, we conduct experiments on three benchmarks: WMT 2020 English-German (En-De), WMT 2020 German-English (De-En), and WMT 2016 English-Romanian (En-Ro). We train Transformers from scratch using the training set and evaluate on dev/test sets. All models are trained for 300k steps. The batch size is 32k for En-De/De-En tokens and 16k for En-Ro. Hyperparameters settings except label smoothing are the same as Vaswani et al. (2017). We train models from scratch and preprocess datasets following previous work. For En-De and De-En, We apply the same filtering process described in Zeng et al. (2021) and get about 37M parallel sentence pairs, which are tokenized with Moses [2] and segmented by byte pair encoding BPE (Sennrich et al., 2016) with 32000 merge operations. For En-Ro, we have 608k parallel sentences tokenized and segmented using the same tool as En-De and De-En.

For the multilingual setting, we use the pretrained mBART-50 (Tang et al., 2020), which is designed specifically for NMT and trained to support over 50 languages. We choose 10 directions of WMT16 to evaluate our method, including En↔De, En↔Cs, En↔Fi, En↔Ro, En↔Ru. For the multilingual NMT setting, we use the large version of the released mBART-50[3]. The benchmarks we used are the ten tasks of WMT16 supported by mBART-50. The dataset statistics of both bilingual and multilingual settings can be found in Table 1.

	Dataset	Train	Valid	Test
Bilingual	WMT20 En-De	37M	1997	1418
	WMT20 De-En	37M	2000	785
	WMT16 En-RO	608K	1999	1999
Multilingual	WMT16 En-De	-	2169	2999
	WMT16 De-En	-	2169	2999
	WMT16 En-Cs	-	2656	2999
	WMT16 Cs-En	-	2656	2999
	WMT16 En-Ro	-	1999	1999
	WMT16 Ro-En	-	1999	1999
	WMT16 En-Ru	-	2818	2998
	WMT16 Ru-En	-	3003	2998
	WMT16 En-Fi	-	1500	3000
	WMT16 Fi-En	-	1500	3000

Table 1: Dataset statistics.

For MBR settings, our results rely on a candidate list of two sizes: low-cost (N=10) and high-cost (N=50). In the analysis part, we use the high-cost setting. Following Freitag et al. (2022), we use BLEURT v0.2 (Sellam et al., 2020) as our utility function to achieve state-of-the-art performance. [4] For evaluation, we mainly report BLEURT points for both the bilingual and multilingual settings, except for bilingual experiments on WMT16 En-Ro, where we report sacreBLEU as its training corpus is case insensitive and the BLEURT metric is trained on case-sensitive data.

For the significance test, we first tokenize the generated sequence and reference with the tokenizer[5], and then use the script provided by the Moses toolkit[6]. It is worth noting that the significance test is conducted with NIST (Doddington, 2002) and BLEU (Papineni et al., 2002) scores.

4.2. Inconsistency between Token- and Sequence-level Distributions

One key distinction between MBR and beam search is that MBR relies on sequence-level distribution $P(y|x)$ and re-ranks among sequence samples, while beam search generates tokens greedily from the token-level distribution $P(y_t|y_{<t}, x)$. Therefore, we explore the effect of label smoothing

[2] http://www.statmt.org/moses/

[3] https://huggingface.co/facebook/mbart-large-50

[4] Due to the space limitation, we will provide results using other utility functions on the additional page upon acceptance.

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[6] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/analysis/bootstrap-hypothesis-difference-significance.pl

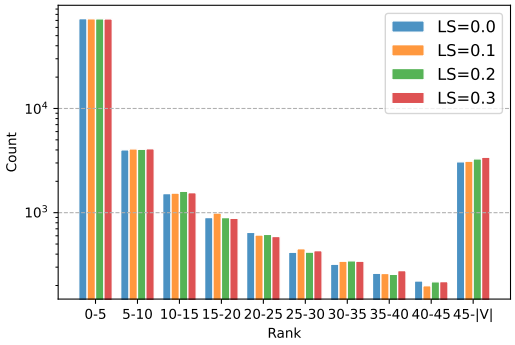

Figure 2: Ranking statistics for tokens in the ground-truth sentence within the token-level distribution $P(y_t|y_{<t}, x)$.

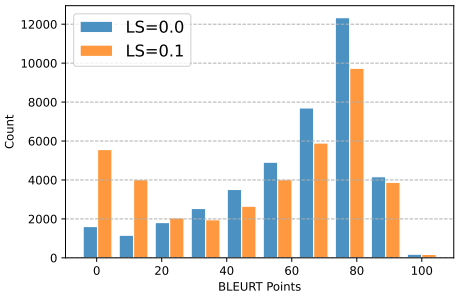

Figure 3: Translation quality statistics for sequences within the top-20 candidates of the sequence-level distribution $P(y|x)$.

over the model's sequence-level and token-level distributions.

Figure 2 demonstrates rankings of ground-truth tokens on the token-level distributions. We predict the teacher-forcing probabilities of reference tokens and investigate how the ranks of these tokens within the distribution $P(y_t|y_{<t}, x)$ change with label smoothing factors. We can see that most of the reference tokens are ranked within the top 0-5, indicating that our models are well-trained. When we increase the label smoothing factor, the rankings only change slightly. Specifically, the count of reference tokens in rank 0-5 slightly drops, and that in rank 45-$|V|$ (tail of the token-level distribution) slightly increases. It implies that label smoothing makes the model mildly less confident at the token-level as intended. It improves the models' generality and accords with our experiments in Figure 1 that label smoothing provides minor improvements with beam search.

The minor impact on the token-level distribution can lead to a huge disparity regarding the sequence-level distribution. With the exact top-N (Yan et al., 2022) decoding algorithm, which is a DFS-based search algorithm equipped with a min-heap, we decode the topmost (i.e., top-20) sequences of the sequence-level model distribution. Figure 3 plots the translation qualities of these topmost sequences for models trained with and without label smoothing. As we can see, compared to the model without label smoothing, the model trained with label smoothing has more low-quality sequences in its top region of sequence-level distribution. It suggests that label smoothing skews the model distribution and gives poor sequences higher ranks/probabilities. This may relate to the well-known label bias problem (Lafferty et al., 2001), wherein the sampling process of the model's short-sighted decisions on certain steps lead to poor translations. This leads to low-quality hypotheses and reference spaces and explains MBR's deteriorated performance in Figure 1.

To further understand why a small distortion in the model's token-level distribution results in a much skewed sequence-level distribution, we examine the auto-regressive nature of machine translation models. As a concrete example, suppose we have a reference sequence of 30 tokens. Given a model trained without label smoothing and *perfectly* fit the data set, it should receive 100% probability for each reference token and the whole sequence. In contrast, with a model with label smoothing $\lambda = 0.1$, each reference token receives a 90% probability, whereas the reference sequence as a whole receives only $90\%^{30} = 4\%$. This effect further enlarges. When $\lambda = 0.2$, the reference sequence only receives about 0.1% probability. Consequently, as shown in Figure 1, the model re-distributes the probability mass to many low-quality sentence candidates. We adopt the term *autoregressive over-smoothness* for this issue. In the next section, we discuss how to quantify *autoregressive over-smoothness*.

5. Method

We propose DC-MBR to address the above issue. The idea is to sharpen the sequence-level distribution, thus allowing the model to benefit from its own confidence. To this end, we first propose a measure of *autoregressive over-smoothness*, and then introduce our DC-MBR.

5.1. Measuring Autoregressive Over-Smoothness Using Entropy

Measuring to what extent the model suffers from *autoregressive over-smoothness* is non-trivial, as the search space of sequence-level distribution is prohibitively large. We turn to a token-level measure that performs well empirically, the token distribution entropy,

$$H = \sum_{i}^{|V|} P(y_t^i) \log P(y_t^i). \tag{12}$$

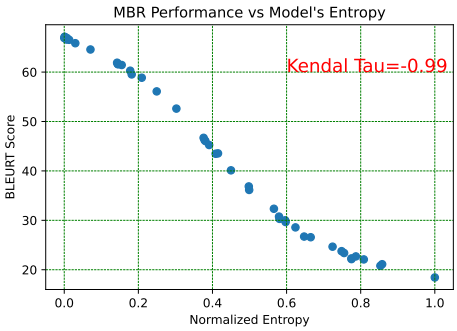

Figure 4: The performance of MBR against the smoothness of the model distribution. Each data point is a Transformer-based model. The performance of MBR is negatively correlated with entropy values. ($\tau = -0.99$).

The connection is straightforward. The lower the token level entropy is, the less probability mass is distributed to the golden reference.

To validate our measure, we conduct experiments on the WMT20 En-De task and investigate the relationship between entropy and MBR performance. To control the entropy values, we train 45 Transformer-base models with various hyper-parameters (i.e., $\alpha \in [0.1, 0.9]$ and $\beta \in [0.1, 0.5]$) of generalized entropy regularizations (GER, Meister et al. 2020b), where label smoothing is one of the special cases. We use a Transformer-base model (Vaswani et al., 2017) and fine-tune the model for 10k more steps. Figure 4 plots the BLEURT points against corresponding model entropy values. Our measure of *autoregressive over-smoothness* is a good proxy as it correlates strongly (-0.99) with the MBR performance. The model with smaller token distribution entropy suffers less from *autoregressive over-smoothness* and achieves a better performance with MBR.

5.2. Distributional Cooling for MBR

In order to mitigate the *autoregressive over-smoothness* and further improve MBR, it is essential to sharpen the model distribution. This is simply the reverse of label smoothing. While it can be achieved by training with negative entropy regularizations, it would require extra computational overhead and it would sacrifice the model's performance with beam search.

Instead, we manipulate the generation process of both hypothesis space and reference space by distributional cooling. It turns down the Softmax temperature and thus reduces the token-level entropy. Formally, we divide the logits o by temperature T before normalization,

$$P(y_t|y_{<t}, x; \theta, T) = \frac{\exp \frac{o_t}{T}}{\sum_j^{|V|} \exp \frac{o_j}{T}}, \quad (13)$$

and generate candidates in hypothesis space and reference space with

$$\mathcal{Y}^T_{\text{model}} \sim \prod_t P(y_t|y_{<t}, x; \theta, T). \quad (14)$$

The computation of MBR becomes

$$\hat{\mu}_u(h; x, \theta) := \frac{1}{N} \sum_r^{\mathcal{Y}^{T_r}_{\text{model}}} u(h, r), \quad (15)$$

$$\hat{y}^{\text{DC-MBR}} = \text{argmax}_{h \in \mathcal{Y}^{T_h}_{\text{model}}} \hat{\mu}_u(h; x, \theta). \quad (16)$$

We use separate temperatures T_h and T_r for the hypothesis and reference spaces due to their distinct roles in MBR. See more discussion in Section 5.6

With distributional cooling, we model a label (un-)smoothing process, as it forces the model to focus on its most confident candidates and avoid distributing probability mass on unconfident ones. It is simple and easy to implement. It only modifies the decoding phase, which can be easily applied to off-the-shelf MT models and does not affect the model's performance with beam search.

5.3. Proof of Equivalence

Distributional cooling also theoretically connects with the label (un-)smoothing. In this section, we provide the proof.

Proposition 1 *The optimal solution of the model trained with label smoothing λ is $\hat{P}_\lambda$ whose probability of i-th token is:*

$$\hat{P}^i_\lambda = \begin{cases} 1-\lambda & y^i \text{ is golden token} \\ \frac{\lambda}{|V|-1} & \text{otherwise} \end{cases}. \quad (17)$$

Intuitively, this solution is straightforward since minimizing cross-entropy loss equals minimizing the Kullback–Leibler divergence between target distribution Q and model distribution P. The loss achieves zero if and only if two distributions are the same.

With the assistance of Proposition 1, we can further derive the following lemma. [7]

Lemma 1 *Given two models that achieve the optimal solutions with different label smoothing factors $\lambda_1, \lambda_2 < 1$, there exists a Softmax temperature factor $T = (\log \frac{1-\lambda_1}{\lambda_1})/(\log \frac{1-\lambda_2}{\lambda_2})$ that can transform $\hat{P}_{\lambda_1}$ to $\hat{P}_{\lambda_2}$.*

[7] The detailed proof for Proposition 1 and Lemma 1 can be found in Appendix upon acceptance.

Models	Models	BS	MBR	Ours	Δ
$N = 10$	Transformer	29.1	28.9	28.9	+0.0
	+ LS 0.1	31.1	25.1	30.9	+5.4
	+ LS 0.2	31.4	19.7	31.2	+12.5
	+ LS 0.3	31.5	12.5	30.9	+28.4
$N = 50$	Transformer	29.1	29.4	29.1	-0.3
	+ LS 0.1	31.1	27.7	31.2	+3.5
	+ LS 0.2	31.4	22.5	31.5	+9.0
	+ LS 0.3	31.5	15.8	31.3	+15.5

(a) Utility: Sacrebleu, Score: SacreBLEU, on En-Ro task.

Models	Models	BS	MBR	Ours	Δ
$N = 10$	Transformer	65.0	63.8	68.8	+5.0
	+ LS 0.1	64.2	41.0	67.9	+26.9
	+ LS 0.2	64.9	24.0	68.3	+44.3
	+ LS 0.3	64.6	17.1	68.3	+51.2
$N = 50$	Transformer	65.0	68.7	70.1	+1.4
	+ LS 0.1	64.2	52.1	69.4	+17.3
	+ LS 0.2	64.9	31.3	69.8	+38.5
	+ LS 0.3	64.6	21.0	69.8	+48.8

(b) Utility: BLEURT, Score: BLEURT, on En-De task.

Table 2: Gray: Models perform poorly with original MBR. We investigate two settings: Low cost, N=10, 100 utility function calls per sentence; High cost, N=50, 2500 utility calls per sentence. Our results are significantly better than "MBR"($p < 0.01$).

The above Lemma proves the equivalence between distributional cooling and label smoothing training. Thus, we can exactly manipulate the Softmax temperature to recover the over-smoothness brought by label smoothing, and, furthermore, improve the performance of MBR. This justifies our approach in that distributional cooling with a temperature $T < 1.0$ does not just make the model's output distribution sharp in any direction. It transforms the distribution towards the optimal solution of a model trained by a smaller label smoothing.

5.4. Main Results

Table 2a and 2b show our results on the bilingual NMT setting. Table 3 provides our results under the multilingual NMT setting. The default value of temperature is set to 0.5.

Mitgating Autoregressive Over-smoothness. As shown in the gray rows of the two tables, we confirm that models trained by label smoothing ('*+LS xx*') generally improve with beam search (*BS*), but label smoothing performs poorly with naive MBR (*MBR*). The performance drops drastically no matter the choice of the number of candidates or tasks. On the other hand, DC-MBR (column *Ours*) achieves strong and consistent performance across different choices of label smoothing, where the performance gap between ours and naive MBR can even reach about 28 BLEU scores and 50 BLEURT points. This consistency indicates that our methods address the *autoregressive over-smoothness*.

Improving Sub-optimal Settings. We compare our performance with naive MBR under an unbiased setting. In bilingual NMT (Table 2b), we observe significant gaps (+5.0/+1.4 BLEURT point) in both the low-cost scenario and the high-cost scenario. In multilingual NMT (Table 3), the conclusions are similar. Our method significantly outperforms MBR with +2.5 BLEURT points when $N = 10$ and with +0.5 BLEURT points when $N = 50$. Our results suggest that the widely used *unbiased* setting is sub-optimal, and the construction of hypothesis and reference space needs exploration.

DC-MBR vs Beam Search. Further, we compare beam search ('BS') with our approach, as beam search is the widely applied decoding algorithm in NMT applications. As shown in Table 2b and 3, our methods strongly outperform beam search with +4.8 to +5.1 BLEURT points in the bilingual setting and with +1.2 and +2.2 BLEURT points in the multilingual setting. Compared with naive MBR, which performs weaker than beam search when $N = 10$, our methods perform much better in the low-cost scenario. The results indicate that our approach makes MBR more applicable in place of beam search in NMT applications, with lower costs and higher translation quality.

Computational Cost Reduction. A by-product of our approach is our method can achieve the same performance with much less computational cost, i.e., the number of candidates, and thus enable a much faster decoding process. For instance, our low-cost result (68.8, 'Transformer, Ours, N=10' in Table 2b) is comparable to that of the original MBR's high-cost result (68.7, 'Transformer, MBR, N=50'). The computational cost is reduced from 2500 to 100 BLEURT calls (25x speedup), due to the quadratic nature of MBR. Compared with other acceleration methods in MBR (Eikema and Aziz, 2022; Freitag et al., 2022), which mainly focus on truncating the hypothesis space or the reference space and accelerating the MBR computation process solely, our methods additionally reduce the cost of candidate generation.

5.5. The Number of Candidates

In our approach, we decrease the Softmax temperature to sharpen the token-level distribution. This may reduce the diversity of generated candidates. Thus, one possible concern is whether our method would limit the potential of MBR when using a large number of candidates. To this end, we study MBR's performance as the number of candidates increases. We plot different temperature choices and report the corresponding BLEURT points over

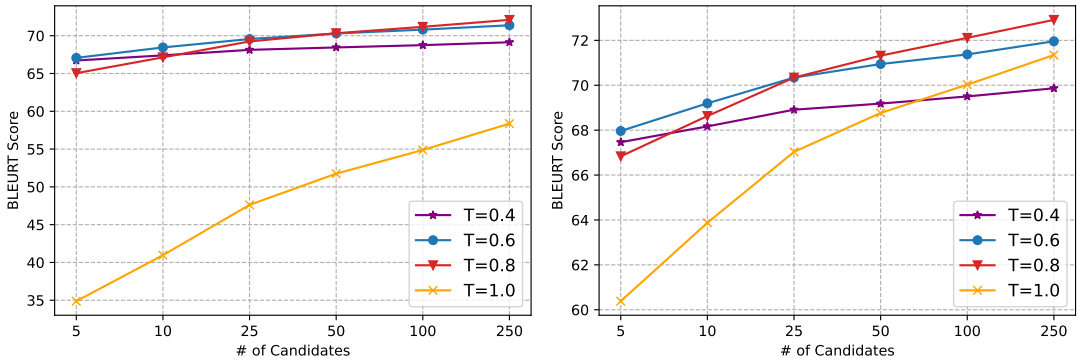

Figure 5: BLEURT points against the number of candidates used with different temperature values. All results are averaged over three random runs. **Left**: Transformer-base w/ LS=0.1; **Right**: Transformer-base w/o LS

mBART	BS	N=10			N=50		
		MBR	Ours	Δ	MBR	Ours	Δ
En-De	71.8	70.3	**73.7**	+3.4	74.2	**74.8**	+0.6
De-En	74.0	73.4	**74.3**	+0.9	**75.1**	74.7	-0.4
En-Cs	73.0	70.1	**75.3**	+5.2	75.0	**77.2**	+2.2
Cs-En	70.8	69.8	**71.2**	+1.4	71.7	**71.7**	+0.0
En-Ro	77.2	77.3	**78.6**	+1.3	**79.8**	79.5	-0.3
Ro-En	72.3	72.2	**72.5**	+0.4	**73.6**	72.8	-0.8
En-Ru	70.8	68.3	**72.9**	+4.6	72.5	**74.4**	+1.9
Ru-En	71.6	70.3	**72.2**	+1.9	72.4	**72.7**	+0.3
En-Fi	77.4	75.6	**79.9**	+4.2	80.1	**81.8**	+1.7
Fi-En	68.9	68.1	**69.6**	+1.5	70.1	**70.2**	+0.1
Average	72.8	71.5	**74.0**	+2.5	74.5	**75.0**	+0.5

Table 3: BLEURT points for the ten tasks on WMT16 with mBART. Our results are significantly better than "MBR" ($p < 0.01$).

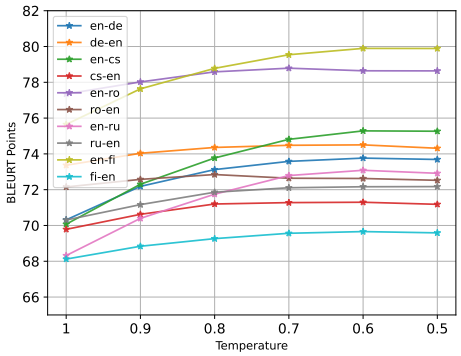

Figure 6: BLEURT points for varying DC-MBR's temperature values on multilingual tasks. We use $N = 10$. Best viewed in color.

WMT20 En-De, and present results with both models trained with and without label smoothing.

Figure 5 shows the results. Given the model with label smoothing, distributional cooling is necessary. Even with 250 candidates sampled per sentence, our methods ($T < 1.0$) strongly outperform the naive MBR ($T = 1.0$) by a considerable margin. Given the model without label smoothing, our methods still significantly outperform naive MBR in most cases, except in the scenario of high costs (e.g., N=250), where a proper choice of temperature (e.g., T=0.6/0.8) is required. For both models, our approach helps MBR achieve strong performance at a low cost (N=5,10). *The above results resolve the concern that sharpening model distribution would limit the gains with a large number of candidates.*

In addition, we find that the performance of our approach improves with an increasing number of candidates, indicating our approach retains the advantage of MBR of not suffering from the *beam search curse* problem (Koehn and Knowles, 2017b; Eikema and Aziz, 2022).

5.6. Tempearture of DC-MBR

Temperature is another key factor for DC-MBR.

Applicability of Distributional Cooling. Besides DC-MBR's effectiveness shown in previous experiments, we want to know whether DC-MBR is applicable to wider settings such as different translation directions. To this end, we plot the performance of each direction of our multilingual experiments in Figure 6. As shown, tuning down temperature almost monotonically improves translation performance in all directions, proving the general applicability of DC-MBR.

Distinct Roles for $\mathcal{Y}_h$ and $\mathcal{Y}_r$. In the above experiments, we use the same temperature to generate both the hypothesis and reference space. Since they have very different roles in MBR decoding, we study how the performances are affected by temperature.

Figure 7 shows the choices of T_h and T_r for the hypothesis and reference space, respectively. Experiments are conducted on the valid set of WMT20 En-De. The model we use is the Transformer-based trained with label smoothing 0.1. As shown, T_h has a significant effect on the performance of MBR. The BLEURT point gap

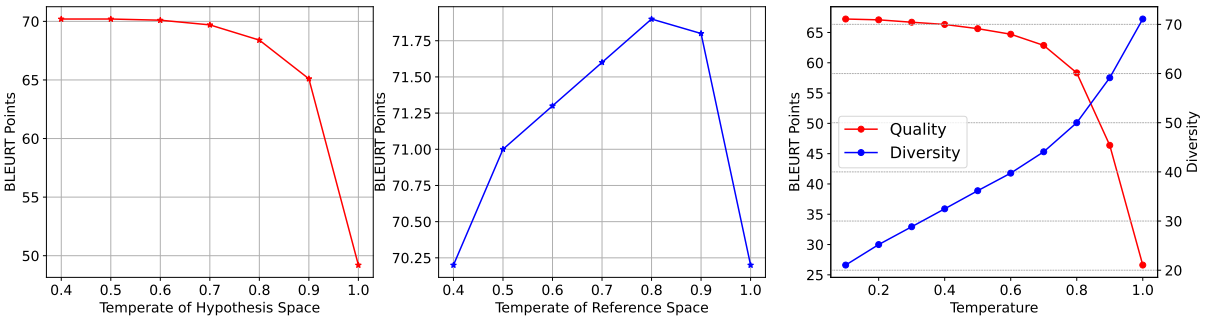

Figure 7: Temperature study of DC-MBR. The number of candidates is 10. **Left**: Fix $T_r = 1.0$ and tune T_h; **Middle**: Fix $T_h = 0.5$ and tune T_r. **Right**: Quality/Diversity scores of sampling candidates versus Temperature T. *Red: Quality; Blue: Diversity*

between the best ($T_h = 0.5$) and worst settings ($T_h = 1.0$) is about 20 points. A sharp hypothesis space is more favorable than a smooth one. On the other hand, a good T_r value also provides a considerable gain on the performance, about +1.5 BLEURT points. Different from T_h, a sharp reference space is not always the best choice. A T_r value that is too high or too low can result in a drop in BLEURT.

To further reveal the characteristics of both spaces, the right plot in Figure 7 plots the quality and diversity for sampled candidates over different choices of temperature. We directly use the BLEURT point for quality, and the diversity score is defined as

$$\text{Div} = \frac{1}{|\mathcal{Y}|^2} \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{Y}} \text{ChrF}(a, b), \qquad (18)$$

which is the average ChrF score (Popović, 2015) of each candidate against the others. We do not use neural metrics such as BLEURT because they are trained to be robust against surface changes. In conclusion, the hypothesis space provides the possible candidates for translation, which should be of high quality and insensitive to diversity. In contrast, the reference space is responsible for the comprehensive evaluation of utilities, which should balance both quality and diversity.

6. Conclusion

We investigated the negative effect of label smoothing on MRB, finding that MBR's performance decreases monotonically with the increase of label smoothing value, and showing that the above phenomenon is due to the *autoregressive oversmoothness* caused by the autoregressive factorization. We then presented a conceptually simple and theoretically well-motivated approach, DC-MBR, to address this issue. Extensive experiments on NMT benchmarks demonstrated the effectiveness of our approach.

Acknowledgement

This work is funded by the Ministry of Science and Technology of China (grant No.2022YFE0204900).

7. Bibliographical References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

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.

Frédéric Blain, Pranava Swaroop Madhyastha, and Lucia Specia. 2017. Exploring hypotheses spaces in neural machine translation. *Asia-Pacific Association for Machine Translation (AAMT), editor, Machine Translation Summit XVI. Nagoya, Japan*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*, 13(4).

Jan Chorowski and Navdeep Jaitly. 2017. Towards better decoding and language model integration in sequence to sequence models. *Proc. Interspeech 2017*, pages 523–527.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Courtney D Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 13–18.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

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.

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.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. 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, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv preprint arXiv:2202.11822*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2019. An empirical investigation of global and local normalization for recurrent neural sequence models using a continuous relaxation to beam search. 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 1724–1733.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with smt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Michael D Hendy and David Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 59(2):277–290.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017a. Six challenges for neural machine translation. *ACL 2017*, page 28.

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

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

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.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to

document distances. In *International conference on machine learning*, pages 957–966. PMLR.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Rémi Leblond, Jean-Baptiste Alayrac, Laurent Sifre, Miruna Pislar, Lespiau Jean-Baptiste, Ioannis Antonoglou, Karen Simonyan, and Oriol Vinyals. 2021. Machine translation decoding beyond beam search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8410–8434, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. Cder: Efficient mt evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Bei Li, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang, and Jingbo Zhu. 2020. Shallow-to-deep training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 995–1005.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2395–2405, Melbourne, Australia. Association for Computational Linguistics.

Chi-kiu Lo. 2017. Meant 2.0: Accurate semantic mt evaluation for any output language. In *Proceedings of the second conference on machine translation*, pages 589–597.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Alan Mackworth. 2013. Lecture notes in introduction to artificial intelligence.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. 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.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020a. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online.

Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020b. Generalized entropy regularization or: There's nothing special about label smoothing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6870–6886.

Fandong Meng and Jinchao Zhang. 2019. DTMT: A novel deep transition architecture for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 224–231.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2017. Analyzing neural MT search and model performance. In *Proceedings of the First*

Workshop on Neural Machine Translation, pages 11–17, Vancouver.

Mohammad Norouzi, Samy Bengio, Navdeep Jaitly, Mike Schuster, Yonghui Wu, Dale Schuurmans, et al. 2016. Reward augmented maximum likelihood for neural structured prediction. *Advances In Neural Information Processing Systems*, 29:1723–1731.

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.

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.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

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.

Matt Post. 2018a. 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.

Matt Post. 2018b. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

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. *arXiv preprint arXiv:1704.03169*.

Raphael Shu and Hideki Nakayama. 2018. Improving beam search by removing monotonic constraint for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Citeseer.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936. PMLR.

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.

Felix Stahlberg, Danielle Saunders, Gonzalo Iglesias, and Bill Byrne. 2018. Why not be versatile? applications of the SGNMT decoder for machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*,

pages 208–216, Boston, MA. Association for Machine Translation in the Americas.

Miloš Stanojević and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum bayes-risk decoding for statistical machine translation.

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*, pages 5998–6008.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079.

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

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*.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2020a. Multi-unit transformers for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1047–1059.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2020b. Multi-unit transformers for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1047–1059, Online.

Jianhao Yan, Chenming Wu, Fandong Meng, and Jie Zhou. 2022. Digging errors in NMT: Evaluating and understanding model errors from partial hypothesis space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12067–12085, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-) scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. Wechat neural machine translation systems for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 243–254.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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*.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019a. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019b. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.