It's MBR All the Way Down: Modern Generation Techniques Through the Lens of Minimum Bayes Risk

Amanda Bertsch* and **Alex Xie*** and **Graham Neubig** and **Matthew R. Gormley**
Carnegie Mellon University
[abertsch, alexx]@cs.cmu.edu

Abstract

Minimum Bayes Risk (MBR) decoding is a method for choosing the outputs of a machine learning system based not on the output with the highest probability, but the output with the lowest risk (expected error) among multiple candidates. It is a simple but powerful method: for an additional cost at inference time, MBR provides reliable several-point improvements across metrics for a wide variety of tasks without any additional data or training. Despite this, MBR is not frequently applied in NLP works, and knowledge of the method itself is limited. We first provide an introduction to the method and the recent literature. We show that several recent methods that do not reference MBR can be written as special cases of MBR; this reformulation provides additional theoretical justification for the performance of these methods, explaining some results that were previously only empirical. We provide theoretical and empirical results about the effectiveness of various MBR variants and make concrete recommendations for the application of MBR in NLP models, including future directions in this area.

1 Introduction

> "Sometimes innovation is only old ideas reappearing in new guises ... [b]ut the new costumes are better made, of better materials, as well as more becoming: so research is not so much going round in circles as ascending a spiral."
>
> (Jones, 1994)

Minimum Bayes Risk (MBR) decoding (Bickel and Doksum (1977); §2) is a decoding method following a simple intuition: when choosing a best output from a set of candidates, the desirable output should be both 1) high probability and 2) relatively consistent with the rest of the outputs (i.e., outputs that are not consistent with the other outputs are high *risk*– they may be dramatically better or worse than the consensus). MBR thus provides an alternative to the more standard maximum-likelihood decoding; when a sample of sufficient size is taken, MBR almost uniformly outperforms beam search and single-output sampling across tasks, metrics, and datasets (see §6). It is also notable in its flexibility; in §3 we organize and discuss several different design decisions that go into the use of MBR and how they affect the efficacy of the method.

While MBR is rarely applied by name in modern NLP, a number of methods with similar intuitions have gained popularity. In §4, we demonstrate that a number of generation techniques widely used with modern language models can be viewed as special instances of MBR: **self-consistency** (Wang et al., 2023) and its extensions, **range voting** (Borgeaud and Emerson, 2020), **output ensembling** (DeNero et al., 2010; Martínez Lorenzo et al., 2023), and some types of **density estimation** (Kobayashi, 2018). This view exposes connections between seemingly disparate methods and presents theoretical justifications for existing empirical results using these methods. We also discuss how insights from the MBR literature can inform the use of these other MBR-like methods.

With the framing of MBR, the theoretical justification for the empirical performance of several methods becomes clear; the extension of self-consistency to open-ended generations becomes trivial; and several promising modifications to self-consistency and output ensembling are exposed. In particular, modern MBR-like methods often do not apply the insights from research on MBR, suggesting that these methods could be further improved. In §5, we show that some design choices, though seemingly intuitive to a practitioner accustomed to search-based decoding methods, should be avoided when applying MBR.

2 Formalization

We begin with the basics of decoding and MBR.

*Denotes equal contribution.

2.1 Standard decoding

Decoding from an autoregressive model (such as a transformer decoder) is performed tokenwise. The distribution at each decoding step is conditioned on the prior tokens and the input text:

$$p(y_i|y_{<i}, x) \tag{1}$$

The model is *locally normalized*; the probabilities of next tokens sum to 1. The probability of a sequence under this global model distribution is

$$p(y|x) = \prod_{i=1}^{T} p(y_i|y_{<i}, x) \tag{2}$$

Given this distribution, there are several ways of extracting an output: by sampling at each decoding step from the distribution over next tokens (often with some modification to the distribution, e.g. temperature, nucleus, or epsilon sampling; Holtzman et al. (2019)); by always choosing the most probable next token (i.e. greedy decoding); or by performing a search over some subset of the output space, guided by the distribution (e.g. beam search, best-first search). These methods generally return a single output; if multiple output candidates are present, the one with the *maximum likelihood* under the model distribution is returned.

2.2 Minimum Bayes Risk decoding

The traditional formulation of MBR is as a minimization objective. Given a *output space* $\mathscr{Y}$ and a probability distribution over this space $p(y|x)$, we compute the risk $R(y')$ of a candidate decoding y' as the expected error (also called *loss*) under this distribution (Bickel and Doksum, 1977; Kumar and Byrne, 2004; Tromble et al., 2008). The MBR decoding is then the y' within $\mathscr{Y}$ that minimizes risk:

$$\hat{y} = \operatorname*{argmin}_{y' \in \mathscr{Y}} R(y') \tag{3}$$

$$= \operatorname*{argmin}_{y' \in \mathscr{Y}} \mathbb{E}_{y|x}[L(y, y')] \tag{4}$$

$$= \operatorname*{argmin}_{y' \in \mathscr{Y}} \sum_{y \in \mathscr{Y}} L(y, y')p(y|x) \tag{5}$$

We can trivially rewrite the risk as a maximization of gain (also called *utility*) rather than a minimization of error, where $G(y, y') = -L(y, y')$. Gain or loss functions are any function (e.g. a metric) that compares two sequences $G : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$.

Approximating risk Computing this sum over the space of all possible outputs $\mathscr{Y}$ is intractable for most models.[1] In these cases, we approximate the risk $R(y')$ by using a subset of the full space $\mathcal{Y} \subset \mathscr{Y}$; that is, instead of exact computation of the expectation, we approximate it with a sum over independent samples from $p(y|x)$. Generally, this is performed by sampling repeatedly from a model (or several models) and estimating the probability of each individual output as proportional to the relative frequency that the output occurs.[2] For an unbiased sampling method[3] (e.g. ancestral sampling), as the number of outputs drawn goes to infinity, this recovers the model's true distribution of probability over sequences. Thus, we approximate risk using this sample:

$$R(y') \approx \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} L(y, y') \tag{6}$$

$$= -\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} G(y, y') \tag{7}$$

Thus, given a sample (which may include duplicates) $\mathcal{Y}$ and a gain function, we approximate the true MBR decoding rule as:

$$\hat{y} = \operatorname*{argmax}_{y' \in \mathcal{Y}} \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}_e} G(y, y') \tag{8}$$

Separation of evidence and hypothesis sets In many cases, the same subset of the output space is used for both the risk estimate and the candidate outputs. However, when the sample is substantially smaller than the full output space, it is often beneficial to use separate sets (Eikema and Aziz, 2022; Yan et al., 2023). Following prior work (§2.2), we refer to these as the *evidence set* ($\mathcal{Y}_e$) and *hypothesis set* ($\mathcal{Y}_h$).

This separation is beneficial because there are distinct and potentially contradictory desiderata for the two sets. We wish for our evidence set to cover a large, representative portion of the search space to obtain a more accurate estimate of risk. However, we want our hypothesis set to only cover the narrower, high-quality region of the space, as we do not want to consider candidate hypotheses that are low-quality. Applying the separation of evidence and hypothesis sets yields the equation for MBR over two subsets of the output space:

[1] This is the case for many deep generative models, such as a transformer language model and other autoregressive models without conditional independence assumptions.

[2] This is called a Monte Carlo approximation.

[3] We discuss the use of biased samplers in §3.2 and §3.1.

$$\hat{y} = \underset{y' \in \mathcal{Y}_h}{\operatorname{argmax}} \sum_{y \in \mathcal{Y}_e} G(y, y') \tag{9}$$

Note that this implicitly encodes the distribution of the evidence set samples in the sum. That is, by averaging over the gain on evidence set examples, we are estimating the expected gain *under this evidence set's distribution over sequences*.

3 Taxonomy of MBR

Equation 9 demonstrates four major axes along which an MBR method may vary:

1. Choice of hypothesis set $\mathcal{Y}_h$
2. Choice of evidence set $\mathcal{Y}_e$
3. Choice of gain (or error) function $G(y, y')$
4. Choice of evidence distribution $p(y|x)$

In this section, we examine how these four factors affect the efficacy of MBR and give recommendations for each; in Section 4, we discuss how these apply to other MBR-like methods.

3.1 Sampling a hypothesis set

Several recent works show benefits from improving the quality of the hypothesis space. Fernandes et al. (2022) apply a two-stage approach where they first apply an N-best (referenceless) reranker and then do MBR over only the most highly ranked hypotheses, which they also use as the evidence set. Eikema and Aziz (2022) introduce a method, Coarse-to-Fine MBR, that first uses MBR with a cheap-to-compute metric to filter a large hypothesis space to a smaller set, then uses MBR with a better but more expensive to compute metric over the smaller set; they separate evidence and hypothesis sets. Freitag et al. (2023) further investigates sampling strategies for MBR, finding that epsilon sampling (Hewitt et al., 2022) outperforms other strategies in automated and human evaluations.

Another earlier line of work has considered growing *post hoc* the hypothesis set in order to obtain hypotheses with higher expected gain (González-Rubio et al., 2011; González-Rubio and Casacuberta, 2013; Hoang et al., 2021).

3.2 Sampling an evidence set

Comparatively less work has studied strategies for sampling the evidence set. Most recent work has adopted the unbiased sampling strategy of Eikema and Aziz (2020), i.e. drawing i.i.d. samples from the model distribution $p(y|x)$ (equation 2). This strategy is motivated by their observation that unbiased sampling is reasonably reflective of the data distribution, much more so than beam search. However, their approach is incompatible with models trained via label smoothing (Szegedy et al., 2016). Yan et al. (2023) attempt to remedy this by sampling the evidence set with temperature $\tau < 1$, sharpening the model distribution.

3.3 What metric do we want to maximize?

The gain G (alternatively, error L) may be an arbitrary function $\mathcal{Y}_e \times \mathcal{Y}_h \to \mathbb{R}$. Early work focused on simple, token-level metrics like word error rate and BLEU (Kumar and Byrne, 2004; Ehling et al., 2007), but more recent work has explored the use of neural metrics (Amrhein and Sennrich, 2022; Freitag et al., 2022), as well as executing outputs in code generation (Shi et al., 2022; Li et al., 2022).

Generally, for both neural and non-neural metrics, MBR with metric G as a gain function will yield the largest downstream improvements on G (Müller and Sennrich, 2021; Freitag et al., 2022; Fernandes et al., 2022). In other words, if one aims to optimize system performance on metric M, one should perform MBR with M as gain. Although MBR uses pseudoreferences, using a metric M to score candidates against these pseudoreferences generally produces a candidate that also scores quite highly on M against the gold reference.

However, MBR also inherits the weaknesses and biases of the gain metric used. MBR has been shown to suffer from length and token frequency biases brought on by the metric, i.e. MBR with BLEU prefers shorter sentences (Nakov et al., 2012; Müller and Sennrich, 2021). Similarly, Amrhein and Sennrich (2022) find that MBR using the metric COMET (Rei et al., 2020) causes higher rates of errors for named entities and numbers due to a lack of sensitivity in the metric. Moreover, MBR is susceptible to overfitting to the metric; Freitag et al. (2023) show that the MBR setting that maximizes the metric is not the one that humans prefer. Thus, if the same metric is used for both MBR and evaluation of the output, *not all of the improvement in that metric can be attributed to higher quality*: it is possible that some of the improvement comes from gaming the metric. This provides an additional reason to evaluate across multiple, diverse metrics.

Note that in the most trivial case, where the met-

Method	Evidence Gen.	Hypothesis Gen.	Metric	$p(y\|x)$
Lattice MBR (Tromble et al., 2008)	N-best list	N-best list	BLEU	translation lattice
Coarse-to-fine MBR (Eikema and Aziz, 2022)	ancestral sampling	`filter`(sample)	BEER	single model
Wiher et al. (2022)	ancestral sampling	evidence + more decodings	BEER	single model
MBR-DC (Yan et al., 2023)	temperature sampling[1]	temperature sampling[1]	BLEURT	single model
Ours (§ 3.3)	ancestral sampling	temperature sampling	BERTScore	single model
Ours (§ 3.4)	ancestral sampling	temperature sampling	BERTScore	length-corrected scores
Freitag et al. (2023)	epsilon sampling		BLEURT	single model
Crowd sampling[2] (Suzgun et al., 2023)	temperature sampling		neural score metric	single model
MBR-Exec (Shi et al., 2022)	temperature sampling		execution match	single model
Self-consistency (SC) (Wang et al., 2023)	temperature sampling		exact answer match	single model
Complex SC (Fu et al., 2022)	`filter`(temperature sample)		exact answer match	single model
SC for open-ended gen (Jain et al., 2023)	temperature sampling		n-gram overlap	single model
Range voting (Borgeaud and Emerson, 2020)	beam search		n-gram overlap	single model
Post-Ensemble (Kobayashi, 2018)	beam search for each model in ensemble		cosine similarity	model set
AMRs Assemble! (Martínez Lorenzo et al., 2023)	model set	beam search	perplexity	model set

Table 1: Recent work under our taxonomy. The line separates methods that are explicitly MBR (above) from those that we identify as MBR-like (below).
[1] Different temperatures used for evidence and hypothesis.
[2] While Suzgun et al. (2023) coin the new term *crowd sampling*, they also explicitly refer to their method as MBR.

ric is $G(y, y') = \mathbb{1}[y = y']$, MBR recovers mode-seeking methods like beam search– i.e. MBR under this metric, in expectation, yields the maximum likelihood decoding. This is because, as the size of the sampled evidence set grows to infinity, the most frequent evidence set sequence (and thus the sequence with the highest gain) becomes the one with the highest probability under the sampling distribution.

3.4 What probability distribution should we use to estimate risk?

Most MBR decoding methods use the model's score distribution over outputs, s, as the (unnormalized) evidence distribution. Alternately, this distribution may be normalized by a temperature (during minimum risk training (Smith and Eisner, 2006) or decoding (Yan et al., 2023)). Some work (e.g Suzgun et al. (2023)) interprets this as a weak proxy for the human or true distribution, arguing that the true objective is to minimize error under the human distribution:

$$\operatorname*{argmin}_{y' \in \mathcal{Y}_h} \mathbb{E}_{y \sim p_{\text{human}}}[L(y, y')]$$

Note that this is not the only reasonable choice of $p(y|x)$; other possible distributions include a distribution over outputs from multiple models (§4.2) or the length-penalized distribution over a single model's outputs $p_l(y|x)$ (§5.3).

4 MBR as a frame for other methods

Self-consistency, output ensembling, density estimation, and range voting can all be viewed through the framing of MBR. This exposes unstated connections between the methods and provides some theoretical backing to the empirical success of these methods. We discuss each in turn.

4.1 Self-consistency as MBR

Self-consistency (Wang et al., 2023) is a method for choosing outputs from language models. In self-consistency, the model is prompted to generate an explanation and then an answer. Multiple outputs $\mathcal{O} = \{y_1, \ldots, y_m\}$ are sampled from the model, the answers $\mathcal{A} = \{a_1, \ldots, a_m\}$ are extracted $a_i = \text{ans}(y_i)$, and the most frequent answer is returned:

$$\operatorname*{argmax}_a \sum_{i=1}^{m} \mathbb{1}(a_i = a) \tag{10}$$

Self-consistency only computes exact match over the *answer*, not the reasoning chain. It is possible to recover MBR from this method by either taking the hypothesis/evidence sets to be the set of resulting answers $\mathcal{Y}_h = \mathcal{Y}_e = \mathcal{A}$ discarding the reasoning chain, or by defining a gain function $G(y, y') = \mathbb{1}(\text{ans}(y) = \text{ans}(y'))$ over full outputs $\mathcal{O}$; though notationally different, they are mathematically equivalent.

Thus, self-consistency is a type of MBR decoding in which we approximate the risk with a Monte Carlo estimate (cf. Eq. 6), the answers are sampled from the model (conditioned on the prompt), and the metric is exact match of the "final answer."

This framing additionally explains some results from the self-consistency paper. Wang et al. (2023) compare the performance of self-consistency across sampling strategies, finding that

the best of the strategies they tried are those that are closest to ancestral sampling (nucleus sampling with $p = 0.95$ and $\tau = 0.7$ without top-k sampling). They also find that self-consistency works better with a sampled output rather than outputs from beam search (their Table 6). Through the lens of MBR, this empirical result has a clear theoretical justification: ancestral sampling of evidence sets generally yields the best performance for MBR because this provides an unbiased estimator of the probabilities of the sampled sequences. This also presents an opportunity for improvement: while Wang et al. (2023) do not evaluate on ancestral sampling, it is possible that this would outperform their best results.

Self-consistency is a special case of MBR. Proposed extensions to self-consistency have recovered aspects of generalized MBR decoding, including filtering to smaller hypothesis/evidence sets (Fu et al., 2022) and the use of alternative gain metrics (Jain et al., 2023). As a result, the term *self-consistency* has widened in definition from a specific type of MBR to a catch-all for MBR-based decoding methods on large language models.

4.2 Output Ensembling as MBR

Model ensembling techniques that operate on *completed outputs* of models may also be cast in MBR terms. Note that this does not include methods that operate on model weights or partial outputs. Common ensembling methods such as averaging model weights (Izmailov et al., 2018) or averaging token-level probabilities (Sennrich et al., 2016; Manakul et al., 2023) cannot be explicitly formulated as MBR.

The connection to MBR is most straightforward in methods that perform MBR decoding over the outputs of multiple models (DeNero et al., 2010; Duh et al., 2011; Barzdins and Gosko, 2016; Lee et al., 2022, *inter alia*). Representative of this family of methods is Post-Ensemble (Kobayashi, 2018), which ensembles multiple text generation models $\theta_1, \theta_2, \ldots, \theta_n$ by separately decoding from each model, computing pairwise sentence embedding similarity between all pairs of outputs, and yielding the output with greatest average similarity. Observe that this may be framed as MBR minimizing the expected risk over the mixture distribution

$$p_{\text{ensemble}}(y|x) = \begin{cases} p_{\theta_1}(y|x) & \text{with probability } \pi_1 \\ \cdots & \\ p_{\theta_n}(y|x) & \text{with probability } \pi_n \end{cases}$$

where $\sum_{i=1}^{n} \pi_i = 1$. While π_i is usually taken to be uniform over the ensemble, this need not always be the case (Duan et al., 2010).

Other methods may be viewed as relaxations of MBR decoding. Assemble! (Martínez Lorenzo et al., 2023) ensembles Abstract Meaning Representation (AMR) graph parsers by computing the pairwise perplexities of each output under *each parser*. While this is not precisely MBR, it may be viewed as a variation where the evidence set is *a set of models*, not a set of model outputs.

$$\hat{y} = \underset{y' \in \mathcal{Y}_h}{\operatorname{argmin}} \, \mathbb{E}_{\theta \sim \pi(\cdot)}[L(\theta, y')]$$

In this case, the error $L(\theta, y')$ is the perplexity of y' under model θ, i.e. $\exp(-\log p_\theta(y')) = \frac{1}{p_\theta(y')}$, and $\pi(\cdot)$ is the distribution over models.

4.3 MBR as Density Estimation

Interestingly, Post-Ensemble (Kobayashi, 2018) (§4.2) was not formulated as MBR (and in fact never referred to by name as MBR), but rather as kernel density estimation. Kernel density estimation is a non-parametric method for estimating the probability density function p of an unknown distribution, given samples $(x_1, x_2, \cdots, x_n)$ from that distribution (Rosenblatt, 1956; Parzen, 1962).

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) \tag{11}$$

Indeed, Equation 11 very closely resembles the Monte Carlo estimator of expected loss in Equation 6. This connection allowed (Kobayashi, 2018) to propose approximation error bounds on MBR, drawing from the density estimation literature.[4]

Note that the kernel function $K(x, x_i)$ is more commonly written as $K(x - x_i)$, or $K(x^T x_i)$ for directional statistics. While this may seem limiting, we can rewrite commonly used MBR metrics in this form; we show this for ROUGE-n as an example. For a sequence y, define $T_n(y)$ to be a vector of size $|V|^n$, where $|V|$ is the size of the vocabulary, containing the number of times every possible n-gram appears in y. Then we can rewrite ROUGE-n as the following:

$$\begin{aligned} & K_{\text{R}}(T_n(y) - T_n(y')) \\ &= 1 - \frac{|T_n(y) - T_n(y')|_1}{|T_n(y)|_1 + |T_n(y')|_1} \end{aligned} \tag{12}$$

[4] We do not reproduce their bounds here; we direct interested readers to the original paper.

where $|\cdot|_1$ is the $L1$ norm.

The similarity between density estimation and MBR yields an alternative interpretation of MBR as a mode-seeking search. However, we are not seeking the mode of the model's distribution over outputs, $p(y|x)$, but rather that of a distribution over some features $\phi(y)$ of our output, $p'(\phi(y)|x)$. For instance, in the case of ROUGE-n MBR,

$$\hat{y} = \underset{y' \in \mathcal{Y}_h}{\text{argmax}} \sum_{y \in \mathcal{Y}_e} K_{\text{R}}(T_n(y') - T_n(y)) \quad (13)$$

$$\approx \underset{y' \in \mathcal{Y}_h}{\text{argmax}}\, p'(T_n(y')|x) \quad (14)$$

We posit that this alternative distribution $p'(T_n(y')|x)$ may be better correlated with performance on specific downstream metrics than the original model distribution, potentially adding an additional justification for MBR's effectiveness. We hope this may inspire future work investigating the theoretical underpinnings of MBR.

4.4 Range Voting as MBR

Methods that take inspiration from outside of NLP may also be MBR-like; in particular, some MBR-like algorithms in the literature are formulated from a voting theory perspective where candidate hypotheses are assigned votes based on similarity to some set of voters (Wang et al., 2023; Jain et al., 2023; Suzgun et al., 2023; Hoang et al., 2021). We show here that range voting (Borgeaud and Emerson, 2020), which broadly encapsulates these proposed voting methods, reduces to MBR.

Range voting describes a family of voting systems in which each voter assigns each candidate a score and the candidate with the greatest total or average score is elected. Observe that the set of candidates C corresponds to the hypothesis set $\mathcal{Y}_h$ and the set of voters V corresponds to the evidence set $\mathcal{Y}_e$. Then, if voter v's score for candidate c is taken to be a gain $G(v, c)$ and each voter is assigned uniform weight, range voting is equivalent to the MBR decision rule in Equation 8:

$$c_{\text{elected}} = \underset{c \in C}{\text{argmax}} \frac{1}{|V|} \sum_{v \in V} G(v, c) \quad (15)$$

Other range-voting methods can similarly be cast as MBR variants.

5 Design Decisions Impact MBR Performance

Although all the methods in Section 4 are MBR-like, they make very different decisions about the four design choices in our MBR taxonomy. To demonstrate the importance of the method design, we consider empirically two cases where changing design impacts the performance of the method.

5.1 Experimental Details

We run MBR experiments for abstractive summarization on CNN/DM (Nallapati et al., 2016) with a fine-tuned BART-Large[5] released by the BART authors (Lewis et al., 2020) as our base model. In §5.3, we additionally report results for translation on WMT'16 Romanian-English (Ro-En) (Bojar et al., 2016) using mBART-50 (Liu et al., 2020).[6]

We draw n_e ancestral samples for our evidence set and n_t temperature samples ($\tau = 0.5$ for CNN/DM, $\tau = 0.3$ for WMT'16 Ro-En) for our hypothesis set. We set $n_e = n_t = 30$ in §5.2 and $n_e = n_t = 50$ in §5.3. Unless otherwise specified, we take ROUGE-1 (Lin, 2004) as our gain metric for summarization and BLEU-4 (Papineni et al., 2002)[7] as our gain metric for translation.

Our code is available at `https://github.com/abertsch72/minimum-bayes-risk`.

5.2 The MBR metric matters – but perhaps not as much as the hypothesis set

We find that using MBR with the summarization n-gram metric ROUGE-1 (Lin, 2004) improves abstractive summarization performance over beam search on CNN/DM, even when evaluating performance with neural metrics; using the general-purpose neural metric BERTScore (Zhang et al., 2020) as the MBR metric yields highest BERTScore but smaller gains on non-neural metrics, a finding consistent with past work; and even BEER (Stanojević and Sima'an, 2014), a translation metric, works as an MBR metric for this task.

However, prior work using the same dataset and model (Wiher et al., 2022) found that BEER (Stanojević and Sima'an, 2014) underperforms beam search. This divergence in results is likely due to our different choices in hypothesis set – Wiher et al. (2022) use the evidence set plus additional

[5] `facebook/bart-large-cnn` on HuggingFace (Wolf et al., 2020)

[6] `facebook/mbart-large-50-many-to-many-mmt`

[7] We use the implementation from `sacrebleu` (Post, 2018) with signature `nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.3.1`

Method	R1	R2	RL	BS
Greedy	43.98	20.88	30.88	88.04
BS (k = 5)	43.16	20.63	30.53	87.82
BS (k = 10)	42.62	20.23	30.02	87.71
DBS ($k = g = 5$)	43.77	20.85	30.77	87.97
MBR ROUGE-1	**46.89**	22.29	32.01	88.41
MBR BEER	46.31	**22.36**	32.02	88.38
MBR BERTSCORE	46.04	22.09	**32.09**	**88.68**

Table 2: MBR results on CNN/DM for various gain functions. We additionally test the same non-MBR, (approximate) mode-seeking baselines as Wiher et al. (2022). All MBR methods outperform all non-MBR methods tested.

outputs from other decoding methods as hypotheses, while we use temperature samples at $\tau = 0.5$. While reusing the evidence set is more efficient than sampling a separate set of hypotheses, it leads to performance degregation in this case; this further emphasizes the importance of choosing the hypothesis set in MBR.

5.3 Varying the risk distribution: lessons from beam search don't translate to MBR

By nature, autoregressive text generation models suffer from length bias: sequence probability monotonically decreases with increasing length, causing shorter, potentially less informative sequences to be favored by the model distribution (Koehn and Knowles, 2017; Stahlberg and Byrne, 2019). For non-sampling methods such as beam search, the sequence probabilities are generally modified with a length-dependent term when comparing sequences (Murray and Chiang, 2018; Cho et al., 2014). Hence, it stands to reason that a length-corrected distribution with these biases alleviated may provide a better estimate of the risk $R(y')$.

Vanilla Monte Carlo MBR (as depicted in Equation 6) yields an estimate of the expected risk under the distribution that our evidence samples are drawn from. To modify the distribution used in our estimate, we turn to **importance sampling**, a method for estimating the expected value of a quantity under target distribution p, given samples from proposal distribution q (Kloek and van Dijk, 1978). For a brief tutorial on importance sampling and description of our estimator, see Appendix A.

We take the *score* of a sequence to be the log probability: We then experiment with two of the strategies described in Murray and Chiang (2018) for constructing the length corrected score $s_l(y|x)$:

(a) **Length normalization**: The model distribution is smoothed with temperature T^β, where T is the sequence length and β is the length penalty, a hyperparameter. A larger β more heavily prioritizes longer sequences.

$$s_l(y|x) = s(y|x)/T^\beta \tag{16}$$

(b) **Length reward** (He et al., 2016): A fixed reward γ is added to the score per token generated.

$$s_l(y|x) = s(y|x) + \gamma T \tag{17}$$

The length-corrected distribution is then $p_l(y|x) \propto \exp s_l(y|x)$. We apply **normalized importance sampling** (Rubinstein and Kroese, 2016) to estimate the risk under the length corrected distribution, i.e. $R(y') = \mathbb{E}_{y\sim p_l}[L(y, y')]$, given samples drawn from the model distribution $p(y|x)$.

We compare our MBR results against beam search both with and without length normalization. We use the models' default values for length penalty ($\beta = 2$ for BART, $\beta = 1$ for mBART).

Our results are Tables 3 and 4. In line with past work, we find that beam search generally benefits from incorporating a length penalty. However, we find that length-corrected MBR underperforms vanilla MBR. This may be due to a gap between the sampling and length-correction distibutions, leading to a high-variance estimator of risk.

Method	R1	R2	RL	BS	LR
Beam search, no correction	43.88	20.96	30.77	87.79	**108.00**
Beam search	43.95	21.00	30.84	87.81	114.39
MBR, No correction	**47.70**	**23.00**	**32.54**	**88.50**	111.64
MBR, Length norm, $\beta = 0.5$	44.29	19.95	29.99	88.03	110.75
MBR, Length norm, $\beta = 1.0$	44.29	19.98	30.0	88.03	110.77
MBR, Length reward, $\gamma = 0.5$	47.60	22.93	32.48	88.48	112.52
MBR, Length reward, $\gamma = 1.0$	47.41	22.72	32.25	88.43	112.50

Table 3: MBR results for various length correction schemes on CNN/DM. We report ROUGE-1, ROUGE-2, ROUGE-L, BERTSCORE, and length ratio, respectively.

Method	BLEU	chrF	BLEURT	BS	LR
Beam search, no correction	33.21	59.81	65.50	94.95	99.37
Beam search	33.06	**60.05**	**65.60**	**94.96**	101.58
MBR, No correction	**33.56**	**60.00**	65.53	**94.96**	**100.04**
MBR, Length norm, $\beta = 0.5$	31.14	58.53	64.70	94.71	102.82
MBR, Length norm, $\beta = 1.0$	31.09	58.51	64.68	94.71	102.60
MBR, Length reward, $\gamma = 0.5$	32.09	59.63	65.19	94.82	105.00
MBR, Length reward, $\gamma = 1.0$	31.29	59.17	64.91	94.73	105.63

Table 4: MBR results for various length correction schemes on WMT'16 Romanian-English. We report BLEU, chrF, BLEURT, BERTSCORE, and length ratio, respectively. We use the chrF (Popović, 2015) implementation from `sacrebleu`. We use the smaller `BLEURT-20-D6` checkpoint for efficiency (Sellam et al., 2020; Pu et al., 2021).

However, our results are also emblematic of a wider trend among minimum-risk techniques. Past work has found that models trained with Minimum Error Rate Training (Och, 2003; Shen et al., 2016), an error-aware training method, do not require length correction in beam search (Neubig, 2016). Similarly, we find that MBR without length correction generates outputs relatively close in length to the references, more so than length-normalized beam search. This suggests that MBR may be to some extent immune from length biases, when they are not introduced by the MBR metric (Müller and Sennrich, 2021).

6 MBR applications in NLP

The use of minimum Bayes risk decoding in NLP predates these MBR-like methods; MBR has been applied by name in NLP since the 1990s.

Historical context Minimum Bayes Risk decoding has roots in Bayesian decision theory, a field of study that dates as far back as the Age of Enlightenment (Bernoulli, 1738; Parmigiani, 2001). Central to Bayesian decision theory is the principle of risk minimization: in the face of uncertainty, an optimal decision maker should choose the option that minimizes the amount of error they can expect to suffer – or, in other terms, maximizes the amount of utility they can expect to enjoy (DeGroot, 1970; Bickel and Doksum, 1977). This is precisely the intuition encoded in MBR (i.e. Equation 3).

Adoption in NLP MBR was adopted by the speech and NLP communities in the 1990s and early 2000s, finding applications in syntactical parsing (Goodman, 1996; Sima'an, 2003), automatic speech recognition (Stolcke et al., 1997; Goel and Byrne, 2000), and statistical machine translation (Kumar and Byrne, 2004; Tromble et al., 2008; Kumar et al., 2009). Many NLP tasks during this time relied upon graph structures as inductive biases (i.e. parse trees or translation lattices/hypergraphs). As such, early MBR works often used these graphical models as hypothesis and evidence spaces. Work on lattice MBR (Tromble et al., 2008), for instance, treated the set of all hypotheses encoded in a word lattice, of which there are exponentially many, as both evidence and hypothesis sets. This is in contrast to most later MBR work, which operates on a relatively small list of text outputs obtained from a neural model. As a result, early work relied on rather involved dynamic programming algorithms

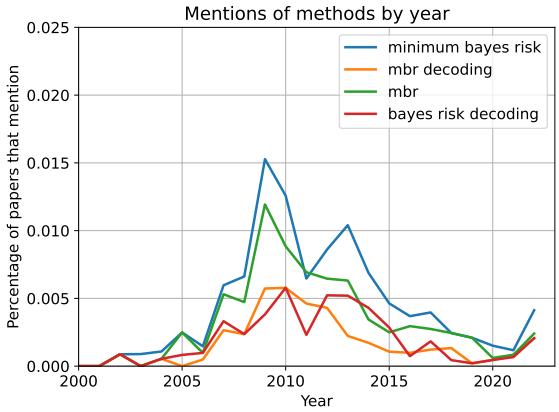

Figure 1: The use of MBR (by name) peaked in the mid-2010s. This graph shows the percentage of ACL Anthology papers that mention several MBR-related phrases by year, from 2000 to 2022.

for exact MBR decoding and were restricted to token-factorizable metrics such as BLEU and edit distance. Later work additionally demonstrated the efficacy of MBR for question answering (Duan, 2013) and for joining statistical and neural approaches to translation (Stahlberg et al., 2017).

Recent usage In an effort to move past beam search, which has well-known pathologies (Stahlberg and Byrne, 2019), MBR has in recent years resurfaced as a decision rule for text-generation models (Eikema and Aziz, 2020). As discussed earlier in §3, several lines of work have sprung up investigating the properties of MBR in modern neural text generation setups. Notably, however, most of these works have focused on applications of the method to neural machine translation, with only a few very recent works studying its applications in other text generation tasks (Shi et al., 2022; Wiher et al., 2022; Suzgun et al., 2023).

Outside of these areas, the method has largely been applied in shared task papers (e.g. Manakul et al. (2023); Yan et al. (2022); Barzdins and Gosko (2016)), as it provides a reliable boost in performance. The fraction of papers in the ACL Anthology that reference MBR (at least by this name) has declined from its peak around 2009 (Figure 1).

7 Conclusion

Minimum Bayes Risk decoding has declined in popularity, but the underlying concept of sampling a set from a distribution and choosing an output to minimize risk according to that set has remained. This concept now takes many surface forms– from self-consistency to range voting to

output ensembles– and current research in these areas rarely draws connections to MBR. While rediscovery is a key part of science, so is recontextualizing new methods within a broader research narrative. This can often reveal new insights or cast findings in a different light. For instance, the empirical benefits of self-consistency can be justified through an MBR framing; work on extensions to self-consistency has rediscovered other properties of MBR; and work on ensembling has raised questions about how to weight mixtures of models that can be reasoned about within the framework of noisy estimates of global probability distributions.

The adoption of newer terms for MBR-like methods may be a type of terminology drift. Related phenomena have been studied in the philosophy of science literature, including pressures to coin new terms (Dyke, 1992; Merton, 1957), potential negative consequences of divergent terminology (Calvert, 1956; Samigullina et al., 2020), and decreased citation of older methods in NLP (Singh et al., 2023). For a more involved discussion of the literature on term coining and possible connections, see Appendix B.

Language is not static, so some degree of terminology drift in scientific literature is unavoidable. However, recognizing the connections between modern techniques and older work is crucial to understanding why such methods are effective. We must not forget the lessons of the past as we search for the methods of the future.

Acknowledgments

We would like to thank Jason Eisner, Patrick Fernandes, and Sireesh Gururaja for useful early discussions about this work, and Saujas Vaduguru, Daniel Fried, and Shuyan Zhou for feedback on this draft.

This work was supported in part by grants from the Singapore Defence Science and Technology Agency, 3M — M*Modal, the Air Force Research Laboratory (AFRL), and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE2140739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Guntis Barzdins and Didzis Gosko. 2016. RIGA at SemEval-2016 task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1143–1147, San Diego, California. Association for Computational Linguistics.

Daniel Bernoulli. 1738. Specimen theoriae novae de mensura sortis. *Commentarii academiae scientiarum imperialis Petropolitanae*, 5:175–192.

Peter J. Bickel and Kjell A. Doksum. 1977. *Mathematical Statistic: Basic Ideas and Selected Topics*. Holden-Day Inc., Oakland, CA.

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

Marcel Bollmann and Desmond Elliott. 2020. On forgetting to cite older papers: An analysis of the ACL Anthology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7819–7827, Online. Association for Computational Linguistics.

Sebastian Borgeaud and Guy Emerson. 2020. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 97–109, Online. Association for Computational Linguistics.

E. S. Calvert. 1956. Technical terms in science and technology. *The American Journal of Psychology*, 69(3):476–479.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Morris H. DeGroot. 1970. *Optimal Statistical Decisions*. McGraw-Hill, Inc., New York.

John DeNero, Shankar Kumar, Ciprian Chelba, and Franz Och. 2010. Model combination for machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 975–983, Los Angeles, California. Association for Computational Linguistics.

Nan Duan. 2013. Minimum Bayes risk based answer re-ranking for question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 424–428, Sofia, Bulgaria. Association for Computational Linguistics.

Nan Duan, Mu Li, Dongdong Zhang, and Ming Zhou. 2010. Mixture model-based minimum Bayes risk decoding using multiple machine translation systems. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 313–321, Beijing, China. Coling 2010 Organizing Committee.

Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. 2011. Generalized minimum Bayes risk system combination. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1356–1360, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Carolynn Van Dyke. 1992. Old words for new worlds: Modern scientific and technological word-formation. *American Speech*, 67(4):383–405.

Nicola Ehling, Richard Zens, and Hermann Ney. 2007. Minimum Bayes risk decoding for BLEU. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 101–104, Prague, Czech Republic. Association for Computational Linguistics.

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

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

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

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation.

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.

Yao Fu, Hao-Chun Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *ArXiv*, abs/2210.00720.

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

Jesús González-Rubio and Francisco Casacuberta. 2013. Improving the minimum Bayes' risk combination of machine translation systems. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

Jesús González-Rubio, Alfons Juan, and Francisco Casacuberta. 2011. Minimum Bayes-risk system combination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1277, Portland, Oregon, USA. Association for Computational Linguistics.

Joshua Goodman. 1996. Efficient algorithms for parsing the DOP model. In *Conference on Empirical Methods in Natural Language Processing*.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with smt features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 151–157. AAAI Press.

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.

Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam Nguyen, Dzung Phan, Vanessa Lopez, and Ramon Fernandez Astudillo. 2021. Ensembling graph predictions for amr parsing. In *Advances in Neural Information Processing Systems*, volume 34, pages 8495–8505. Curran Associates, Inc.

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

Anna Kristina Hultgren. 2013. Lexical borrowing from english into danish in the sciences: An empirical investigation of 'domain loss'. *International Journal of Applied Linguistics*, 23(2):166–182.

Pavel Izmailov, Dmitrii Podoprikhin, T. Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*.

Siddhartha Jain, Xiaofei Ma, Anoop Deoras, and Bing Xiang. 2023. Self-consistency for open-ended generations.

Karen Sparck Jones. 1994. *Natural Language Processing: A Historical Review*, pages 3–16. Springer Netherlands, Dordrecht.

T. Kloek and H. K. van Dijk. 1978. Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica*, 46(1):1–19.

Hayato Kobayashi. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4165–4176, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum Bayes-risk decoding for translation hypergraphs and lattices. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 163–171, Suntec, Singapore. Association for Computational Linguistics.

Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. Maximum Bayes Smatch ensemble distillation for AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Potsawee Manakul, Yassir Fathullah, Adian Liusie, Vyas Raina, Vatsal Raina, and Mark Gales. 2023. CUED at ProbSum 2023: Hierarchical ensemble of summarization models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 516–523, Toronto, Canada. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Pere Lluís Huguet Cabot, and Roberto Navigli. 2023. AMRs assemble! learning to ensemble with autoregressive models for AMR parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1595–1605, Toronto, Canada. Association for Computational Linguistics.

Robert K. Merton. 1957. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659.

Saif M. Mohammad. 2020. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, 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*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 259–272, Online. Association for Computational Linguistics.

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

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of COLING 2012*, pages 1979–1994, Mumbai, India. The COLING 2012 Organizing Committee.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Graham Neubig. 2016. Lexicons and minimum risk training for neural machine translation: NAIST-CMU at WAT2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 119–125, Osaka, Japan. The COLING 2016 Organizing Committee.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

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.

G. Parmigiani. 2001. Decision theory: Bayesian. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 3327–3334. Pergamon, Oxford.

Emanuel Parzen. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.

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.

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.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

B. L. Raad. 1989. Modern trends in scientific terminology: Morphology and metaphor. *American Speech*, 64(2):128–136.

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.

Murray Rosenblatt. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837.

Reuven Y. Rubinstein and Dirk P. Kroese. 2016. *Simulation and the Monte Carlo Method*, 3rd edition. Wiley Publishing.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

L.Z. Samigullina, E.F. Samigullina, O.V. Danilova, and I.A. Latypova. 2020. Linguistic borrowing as a way to enrich oil and gas terminology. In *Proceedings of the International Session on Factors of Regional Extensive Development (FRED 2019)*, pages 58–61. Atlantis Press.

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, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. 2022. Natural language to code translation with execution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3533–3546, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. Forgotten knowledge: Examining the citational amnesia in NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.

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, Sydney, Australia. Association for Computational Linguistics.

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

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

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

Andreas Stolcke, Yochai Konig, and Mitchel Weintraub. 1997. Explicit word error minimization in n-best list rescoring.

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 (CVPR)*.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. CMU's IWSLT 2022 dialect speech translation system. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 298–307, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Jianhao Yan, Jin Xu, Fandong Meng, Jie Zhou, and Yue Zhang. 2023. Dc-mbr: Distributional cooling for minimum bayesian risk decoding.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A More details on importance sampling for MBR

We present in this section the normalized importance sampling estimator of risk used in our experiments in §5.3.

The core insight of importance sampling is that we can rewrite the expected value of a random variable $f(x)$ under target distribution p as another expectation under some proposal distribution q:

$$\begin{aligned}\mathbb{E}_p[f(x)] &= \sum_x f(x)p(x) \\ &= \sum_x f(x)\frac{p(x)}{q(x)}q(x) \\ &= \mathbb{E}_q\left[f(x)\frac{p(x)}{q(x)}\right]\end{aligned}$$

Importance sampling can be particularly useful when sampling from the proposal distribution is easy, but sampling from the target distribution is costly or intractable; this is indeed the case for MBR, as sampling from the length-corrected distribution $p_l(y|x)$ requires computation of its partition function, which has exponential complexity.

Hence, for MBR, if we draw evidence samples $\mathcal{Y}_e$ according to model distribution $p(y|x)$ but wish to compute the risk under some length-corrected distribution $p_l(y|x)$, we may compute

$$\begin{aligned}R(y') &= \mathbb{E}_{y\sim p_l}[L(y,y')] \\ &= \mathbb{E}_{y\sim p}\left[L(y,y')\frac{p_l(y|x)}{p(y|x)}\right] \\ &= \sum_{y\in\mathcal{Y}_e} L(y,y')\frac{p_l(y|x)}{p(y|x)} \\ &= \sum_{y\in\mathcal{Y}_e} L(y,y')w(y)\end{aligned}$$

where we let $w(y) = p_l(y|x)/p(y|x)$, commonly referred to as the importance weight.

Note, however, that importance sampling requires us to be able to exactly compute the probabilities $p(y|x)$ and $p_l(y|x)$; while the former can be computed efficiently (Equation 2), the latter is intractable, again because it requires the partition function. What we can efficiently compute is the unnormalized probability $\tilde{p}_l(y|x) = \exp s_l(y|x)$, where s_l is the length-corrected score given by either Equation 16 or 17.

Fortunately, we can use **normalized importance sampling** to obtain a consistent estimator of the risk by adjusting importance weights (Rubinstein and Kroese, 2016):

$$\begin{aligned}R(y') &= \mathbb{E}_{y\sim p_l}[L(y,y')] && (18)\\ &= \frac{\mathbb{E}_{y\sim p}[L(y,y')\tilde{w}(y)]}{\mathbb{E}_{y\sim p}[\tilde{w}(y)]} && (19)\\ &= \sum_{y\in\mathcal{Y}_e} L(y,y')\cdot\frac{\tilde{w}(y)}{\sum_{y\in\mathcal{Y}_e}\tilde{w}(y)} && (20)\end{aligned}$$

where $\tilde{w}(y) = \tilde{p}_l(y|x)/p(y|x)$. As it is the ratio of two estimates, the normalized importance sampling estimator is *biased* for finite sample sizes.

B Contextualizing this work within philosophy of science

In this section, we contextualize our work in the broader framings of meta-analysis of scientific research.

Patterns of citation in NLP Several factors have been shown to correlate with citation rate in NLP, including author geographic location (Rungta et al., 2022), author gender (Mohammad, 2020), and publication date (Bollmann and Elliott, 2020; Singh et al., 2023). Bollmann and Elliott (2020) conduct a bibliometric anaylsis of the ACL Anthology, finding that the mean age of papers cited decreased significantly from 2010 to 2019. Singh et al. (2023) expand this analysis to the full anthology, finding that, while citations of older papers rose briefly in the mid-2010s, it has since declined, with 2021 marking a historic low for the percentage of citations that went to older papers[8]. They term this *citational amnesia* and discuss several possible reasons for the result, including the shift to neural methods and the rise of new areas of NLP.

Our work raises another potential explanation: some citational amnesia is due to *terminology drift* over time, as old methods begin to be referred to by newer names.

Term coining in science Work in science and technology studies has examined the broader phenomenon of term coining in science. Dyke (1992) argues that neologisms emerge more frequently in fields that prize novelty and see science as fundamentally about leaps of discovery, and fields that are perceived as synthesizing findings from multiple fields are most likely to recycle terms from other disciplines. She cites computer science as an example of a field where most new terms of art emerge from recycling common words, often those that draw a metaphor to some basic physical or human concept; this is reflected in the adoption of the humanizing "self-consistency" and the political-science-inspired "range voting" in decoding. Raad (1989) suggests that evocative, metaphor-laden names are more likely to emerge as a scientific field grows more public-facing and in times where many new terms are being coined; both of these descriptors apply to modern NLP. While several works in linguistics and STS have considered the coining of new terms for new phenomena, relatively little work has focused on the divergence of terminology for previously observed phenomena.

The consequences of divergent or distinct terminology have also been studied, with differences in terminology across fields blamed for slow adaptation of research to practical applications (e.g. in studying visual distortions during plane takeoff (Calvert, 1956)). Borrowing terminology from another language (often Latin or Greek) or from another field has been described as a method to build common ground between researchers (Samigullina et al., 2020) and as a possibly concerning pressure against developing language-specific scientific terminology in lower-resourced languages (Hultgren, 2013). However, most work on lexical divides in science has focused on divides across language or field rather than divides across time in the same field.

[8] They define an "older paper" as one that is more than 10 years older than the paper that is citing it.