Sampling-Based Approximations to Minimum Bayes Risk Decoding for Neural Machine Translation

Anonymous ACL submission

Abstract

In NMT we search for the mode of the model distribution to form predictions. The mode and other high-probability translations found by beam search have been shown to often be inadequate in a number of ways. This prevents improving translation quality through better search, as these idiosyncratic translations end up selected by the decoding algorithm, a problem known as the beam search curse. Recently, an approximation to minimum Bayes risk (MBR) decoding has been proposed as an alternative decision rule that would likely not suffer from the same problems. We analyse this approximation and establish that it has no equivalent to the beam search curse. We then design approximations that decouple the cost of exploration from the cost of robust estimation of expected utility. This allows for much larger hypothesis spaces, which we show to be beneficial. We also show that mode-seeking strategies can aid in constructing compact sets of promising hypotheses and that MBR is effective in identifying good translations in them. We conduct experiments on three language pairs varying in amounts of resources available: English into and from German, Romanian, and Nepali.[1]

1 Introduction

NMT systems (Sutskever et al., 2014; Bahdanau et al., 2015) are trained to predict a conditional probability distribution over translation candidates of any given source sentence. After training, choosing a translation for a given input requires a decision rule: a criterion to elect a 'preferred' translation. MAP decoding, the most common decision rule in NMT, seeks the most probable translation under the model (*i.e.*, the mode of the distribution).

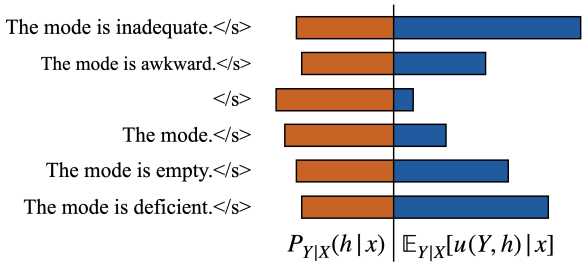

Figure 1: NMT spreads probability roughly uniformly over a large set of promising hypotheses (left). MBR (right) assigns hypotheses an expected utility, revealing clear preferences against those that are too idiosyncratic.

MAP decoding and its approximations such as beam search (Graves, 2012) have been under scrutiny. Stahlberg and Byrne (2019) show that the true mode is oftentimes inadequately short or empty. Better approximate search is known to hurt quality (Koehn and Knowles, 2017; Murray and Chiang, 2018; Kumar and Sarawagi, 2019), a problem known as the *beam search curse*. The success of beam search depends on search biases introduced by hyperparameters such as beam size and length normalisation, which are tuned not to correlate with the objective of MAP decoding, but rather to strike a compromise between mode-seeking search and properties of reasonable translations. Despite its success, a number of problems have been observed: length bias (Cho et al., 2014; Sountsov and Sarawagi, 2016), word frequency bias (Ott et al., 2018), susceptibility to copy noise (Khayrallah and Koehn, 2018; Ott et al., 2018), and hallucination under domain shift (Lee et al., 2019; Müller et al., 2020; Wang and Sennrich, 2020).

Eikema and Aziz (2020) argue that the inadequacy of the mode in NMT is a reasonable consequence of the translation space being combinatorial and unbounded. They show that, while distri-

[1] Code is available at `github.com/ANONYMISED`.

butions predicted by NMT do reproduce various statistics of observed data, they tend to spread probability mass almost uniformly over a large space of translation candidates. This makes their precise ranking in terms of probability mass a fragile criterion for prediction. While some of these candidates are possibly inadequate (*e.g.*, the empty sequence), most of them are similar to one another and exhibit appreciable structural similarity to reference translations. To make better use of the statistics predicted by NMT models, Eikema and Aziz (2020) recommend MBR decoding (Kumar and Byrne, 2004), a decision rule that seeks the translation candidate which maximises an external notion of utility (*e.g.*, an MT evaluation metric) in expectation under the model distribution. While MBR decoding promises robustness to idiosyncratic translations, it remains intractable, much like MAP decoding. Eikema and Aziz (2020) propose an approximation based on Monte Carlo (MC) sampling, which although tractable in principle, requires a prohibitive number of assessments of the utility function.

In this work, we first analyse the procedure by Eikema and Aziz (2020) and establish that it does not suffer from a counterpart to the beam search curse. That is, better search does not hurt translation quality. Their approximation is, however, computationally expensive, requiring a number of assessments of the utility function that is quadratic in sample size. We propose algorithms that scale linearly, allowing us to explore large hypothesis spaces, and considerably improve upon existing approximations to MBR with less computation. Finally, we find that mode-seeking strategies such nucleus sampling and beam search can still aid MBR decoding by constructing compact sets of high expected utility hypotheses, relying on MBR to filter idiosyncratic translations that may be present.

2 NMT and Decision Rules

NMT employs neural networks (NNs) to predict a conditional probability distribution $Y|\theta, x$ over translation candidates of any given source sentence x. The sample space $\mathcal{Y}$ is the set of all sequences of known target-language symbols (*e.g.*, sub-word units). NMT factorises the distribution as a chain of random draws from Categorical distributions

$$Y_j|\theta, x, y_{<j} \sim \mathrm{Cat}(f(x, y_{<j}; \theta)) \qquad (1)$$

parameterised in context. The prefix translation $y_{<j}$ starts empty and grows one symbol at a time until a special end-of-sequence symbol is drawn. At each step j, f maps from varying inputs $(x, y_{<j})$ to a probability distribution over the vocabulary. Common choices for f include recurrent networks (Sutskever et al., 2014; Bahdanau et al., 2015) and Transformers (Vaswani et al., 2017). The NN parameters θ are estimated to attain a local optimum of the regularised log-likelihood function.

After training, and for a given input, choosing a translation requires a *decision rule* to map from a distribution over translation candidates to a single 'preferred' translation. The most common decision rule in NMT is MAP decoding, which outputs the mode of the conditional distribution. Despite the widespread intuition that MAP decoding is an obvious choice, maximum likelihood estimation (MLE) is oblivious to our desire to form predictions.

2.1 MAP Decoding

Maximum-a-posteriori (MAP) decoding outputs the most probable translation under the model. As this is intractable, beam search (Graves, 2012) is used. Beam search is a pruned version of breadth-first search which maintains an active set of k partial translations. For large beam size k, translation quality degrades (Koehn and Knowles, 2017) and the exact y^{MAP} is often the empty sequence (Stahlberg and Byrne, 2019). Therefore, in practice, the beam size is kept small and the objective is length normalised to up-rank longer hypotheses (Wu et al., 2016; Murray and Chiang, 2018).

2.2 MBR Decoding

Minimum Bayes risk (MBR) decoding stems from the principle of maximisation of expected utility (Berger, 1985). A utility function $u(y, h)$ measures the benefit in choosing $h \in \mathcal{Y}$ when $y \in \mathcal{Y}$ is the ideal decision. When forming predictions, we lack knowledge about ideal translations and must decide under uncertainty. MBR lets the model fill in 'ideal decisions' probabilistically as we search through the space of candidates for the one which is assigned highest utility *in expectation*:

$$y^{\text{MBR}} = \arg\max_{h \in \mathcal{Y}} \underbrace{\mathbb{E}[u(Y, h) \mid \theta, x]}_{=:\mu_u(h; x, \theta)} . \qquad (2)$$

MBR has a long history in parsing (Goodman, 1996; Sima'an, 2003), speech recognition (Stolcke et al., 1997; Goel and Byrne, 2000), and MT (Kumar and Byrne, 2002, 2004).

In MT, u can be a sentence-level evaluation metric (*e.g.*, METEOR (Denkowski and Lavie, 2011)

or Sentence BLEU (Chen and Cherry, 2014)). Intuitively, whereas the MAP prediction is the translation to which the model assigns highest probability, no matter how idiosyncratic, the MBR prediction is the translation that is closest (under the chosen u) to all other probable translations. See Figure 1 for an illustration of this concept.

Like in MAP decoding, exhaustive enumeration of the hypotheses is impossible, we must resort to a finite subset $\bar{\mathcal{H}}(x)$ of candidates. Unlike MAP decoding, the objective function $\mu_u(h; x, \theta)$ *cannot* be evaluated exactly. Most approximations to MBR decoding, from Kumar and Byrne (2004) to recent instances (Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017), use k-best lists from beam search for $\bar{\mathcal{H}}(x)$ and to form a biased estimate of expected utility. Eikema and Aziz (2020) use unbiased samples from the model for both approximations: *i)* they follow the generative story in Equation (1) to obtain N independent samples $y^{(n)}$, a procedure known as ancestral sampling (Robert and Casella, 2010); then, *ii)* for a hypothesis h, they compute an MC estimate of $\mu_u(h; x, \theta)$:

$$\hat{\mu}_u(h; x, N) \overset{\text{MC}}{\coloneqq} \frac{1}{N} \sum_{n=1}^{N} u(y^{(n)}, h) \,, \qquad (3)$$

which is unbiased for any sample size N. Eikema and Aziz (2020) use the same N samples as candidates and approximate Equation (2) by

$$y^{\text{N-by-N}} \coloneqq \underset{h \in \{y^{(1)}, \ldots, y^{(N)}\}}{\arg\max} \hat{\mu}_u(h; x, N) \,. \qquad (4)$$

We note that the candidates do not need to be obtained using ancestral sampling. We investigate alternative strategies in Section 5.4. It is important, however, to use ancestral samples to obtain an unbiased estimate of expected utility as we show in Section 5.1. We call this class of MBR algorithms using unbiased MC estimation instances of *sampling-based MBR decoding*.

3 Coarse-to-Fine MBR Decoding

A big disadvantage of $\text{MBR}_{\text{N-by-N}}$ is that it requires N^2 assessments of the utility function. If U is an upperbound on the time necessary to assess the utility function once, then $\text{MBR}_{\text{N-by-N}}$ runs in time $\mathcal{O}(N^2 \times U)$. For a complex utility function, this can grow expensive even for a modest hypothesis space. As NMT distributions have been shown to be high entropy (Ott et al., 2018; Eikema and Aziz, 2020), the quadratic cost prevents us from sufficiently exploring the space of translations. Therefore, we investigate and propose more flexible algorithms.

An important property of sampling-based MBR decoding is that MC estimation of expected utility, Equation (3), and approximation of the hypothesis space in Equation (4) really are two independent approximations. Tying the two is no more than a design choice that must be reconsidered. We start by obtaining N translation candidates from the model, which will form the hypothesis space $\bar{\mathcal{H}}(x)$. Then, we use any fixed number $S < N$ ancestral samples for approximating expected utility in Equation (3). We call this version $\text{MBR}_{\text{N-by-S}}$, which takes time $\mathcal{O}(N \times S \times U)$. Compared to $\text{MBR}_{\text{N-by-N}}$, this variant is able to scale to much larger hypothesis spaces $\bar{\mathcal{H}}(x)$. In practice, however, robust MC estimation for the utility of interest may still require S that is too large for the N we are interested in.

An idea that we explore in this work is to make use of a proxy utility that correlates with the target utility but is cheaper to compute. Even when those do not correlate perfectly, we can make use of the proxy utility to filter the hypothesis space to a manageable size T on which we can perform robust MC estimation of expected utility. We coin this approach coarse-to-fine MBR decoding (or MBR_{C2F}), which filters the hypothesis space to a manageable size in the coarse step, and performs robust MC estimation of expected utility in the fine step:

$$y^{\text{C2F}} \coloneqq \underset{h \in \bar{\mathcal{H}}_T(x)}{\arg\max} \hat{\mu}_{u_{\text{target}}}(h; x, L) \qquad (5a)$$

$$\bar{\mathcal{H}}_T(x) \coloneqq \underset{h \in \bar{\mathcal{H}}(x)}{\text{top-T}} \hat{\mu}_{u_{\text{proxy}}}(h; x, S) \,. \qquad (5b)$$

Upper-bounding the complexity of the proxy utility by U_{proxy}, the target utility by U_{target}, using S samples for MC estimation in the coarse step (5b) and L in the fine step (5a), the complexity of this algorithm is $\mathcal{O}(N \times S \times U_{\text{proxy}} + T \times L \times U_{\text{target}})$. MBR_{C2F} decouples robust MC estimation (large L) from exploration (large N) and the cost of exploration from the cost of the target utility.

As illustrated in Figure 2, we can find proxy utilities that correlate reasonably well with our target utility and are able to give us a rough—but useful—ordering of the hypothesis space. Rather than using a proxy utility, we could use the target utility itself in the coarse-step provided we pick a small S. This, however, most likely leads to too high variability in the ranking, as shown in Figure 2 (left).

src	Convercent erhielt $10 Millionen bei der Finanzierung im Februar von Firmen wie Sapphire Ventures und Tola Capital, womit das gesamte Kapital auf $47 Millionen angehoben wurde.
ref	Convercent raised $10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to $47 million.

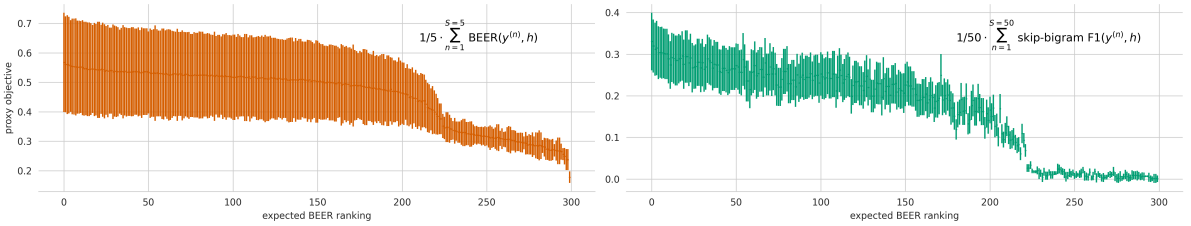

Figure 2: Motivation for coarse-to-fine MBR. We sort 300 candidates sampled from the model along the x-axis from best to worst according to a robust MC estimate (using 1,000 samples) of expected BEER under the model. Left: feasible MC estimates (5 samples) of each candidate's expected BEER. Right: robust and inexpensive MC estimates (100 samples) of expected utility w.r.t. a simpler metric (skip-bigram F1). As estimates are stochastic, we perform 100 repetitions and plot mean $\pm$ two deviations. We can see that the robust estimates (right) correlate fairly well with the expensive ranking we intend to approximate (x-axis), despite of the simpler utility. As we can afford more evaluations of the proxy utility, we obtain estimates of reduced variance, which leads to safer pruning.

4 Data, Systems and Utilities

We perform experiments on three language pairs with varying amount of resources for training: English into and from German, Romanian and Nepali. For German-English (de-en) we use all available WMT'18 (Bojar et al., 2018) news data except for Paracrawl, resulting in 5.9 million sentence pairs. We train a Transformer base model (Vaswani et al., 2017) until convergence and average the last 10 epoch checkpoints to obtain our final model. We test our models on `newstest2018`. For Romanian-English (ro-en) we use all available WMT'16 (Bojar et al., 2016a) news data amounting to 565k sentence pairs. We train a Transformer base model until convergence and pick the best epoch checkpoint according to the validation loss. We test our models on `newstest2016`. Finally, for Nepali-English (ne-en) we use the data setup by Guzmán et al. (2019). We apply the pre-processing step of removing duplicates as in Eikema and Aziz (2020). This results in 235k sentence pairs. We test our models on the FLORES test set, which is of a widely different domain than the training data. We mimick the training setup and models used in Guzmán et al. (2019). In all models we disable label smoothing, as this has been found to negatively impact model fit, which would compromise the performance of MBR (Eikema and Aziz, 2020).

For computational efficiency, we opt for non-neural evaluation metrics for use as utility function in MBR. BEER (Stanojević and Sima'an, 2014) is a non-neural trained metric that has shown good correlation with human judgements in previous WMT metrics shared tasks (Macháček and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016b). In experiments shown in Table 2 in Appendix B we found that using BEER as utility function performed well at pushing translation performance higher across a range of automatic evaluation metrics. We therefore use BEER as the utility of choice in our experiments and as a consequence will consistently report corpus-level BEER scores of MBR translations as well. We also report SacreBLEU (Papineni et al., 2002; Post, 2018a) scores where relevant to be able to detect overfitting to the utility and for comparison with other works.

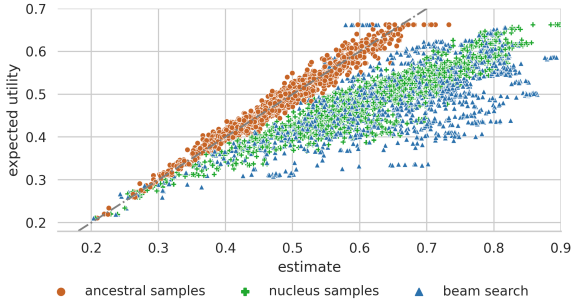

Figure 3: Estimates of expected utility for various hypotheses. We plot practical estimates of expected utility (x-axis) using either ancestral, nucleus or 'beam' samples against an accurate MC estimate using 1,000 ancestral samples. The gray line depicts a perfect estimator.

5 Experiments

5.1 Estimation of Expected Utility

We start by motivating the importance of unbiased estimates of expected utility using ancestral sam-

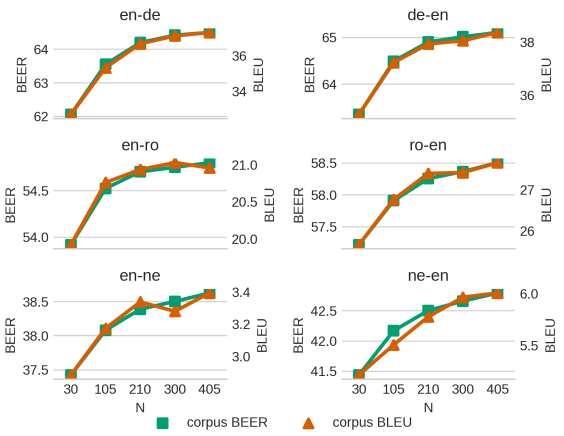

Figure 4: $\text{MBR}_{\text{N-by-N}}$ for various sizes of N using BEER as target utility. We report both BEER and BLEU scores.

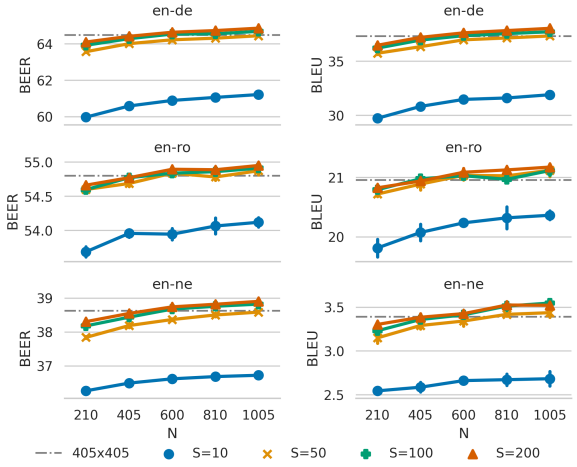

Figure 5: $\text{MBR}_{\text{N-by-S}}$: we estimate the expected utility of N hypotheses using S samples. We show average performance over 3 runs with 1 standard deviation. The dashed line shows $\text{MBR}_{\text{N-by-N}}$ performance at $N = 405$.

ples (*i.e.* sampling-based MBR). In Figure 3 we verify the biasedness of alternatives to ancestral sampling for this computation: nucleus sampling (Holtzman et al., 2020) and 'beam sampling' (*i.e.*, using k-best outputs from beam search for estimating expected utility; Blain et al. (2017)). We can see, rather clearly, that estimates using nucleus samples or beam search bias away from expected utility under the model, while ancestral sampling is unbiased by design and hence should be preferred when approximating the objective function in search. Therefore, in all experiments that follow, we shall use ancestral samples for making unbiased estimates of expected utility, even when different methods are used to construct the hypothesis space.

5.2 N-by-N MBR

Now, we look into scaling $\text{MBR}_{\text{N-by-N}}$. Eikema and Aziz (2020) only explored 30 by 30 approximations to the MBR objective. Our aim is to investigate whether MBR decoding is indeed able to scale to better translation performance with more computation. In Figure 4, we explore N from 30 to 405.[2] As MBR optimises a specific utility (we use BEER), we report translation quality along both BEER and BLEU to detect overfitting to the metric.

We find that MBR steadily improves across language pairs as N grows larger. BLEU scores improve at a similar rate to that of BEER, showing no signs of overfitting to the utility. This is strong empirical evidence that *sampling-based* MBR has no equivalent to the beam search curse. We see this as an important property of a decoding objective.

[2]A batch size of 15 is convenient on our hardware, which is why we work with multiples of 15 in most experiments.

5.3 N-by-S MBR

$\text{MBR}_{\text{N-by-N}}$ couples two approximations, namely, tractable exploration and unbiased estimation of expected utility are based on the same N ancestral samples. Our aim is to learn more about the impact of these two approximations, for which we look into $\text{MBR}_{\text{N-by-S}}$. Moreover, with less than N^2 assessments of utilities per decoding, we can also investigate larger $\bar{\mathcal{H}}(x)$. We explore N ranging from 210 to 1005, while keeping the number of samples used for approximating expected utility of each hypothesis smaller, with S ranging from 10 to 200. We argue that S does not need to grow at the same pace as N, as MC estimates should stabilize after a certain point.[3] See our results in Figure 5.

We find that growing N beyong 405 improves translation quality further, even when the estimates of expected utility are less accurate. Increasing S also steadily improves translation quality, with diminishing returns in the magnitude of improvement. On the other hand, smaller values of S lead to notable deterioration of translation quality and we note higher variance in results. For all language pairs it is possible to improve upon the best $\text{MBR}_{\text{N-by-N}}$ results by considering a larger hypothesis spaces and smaller S. This experiment shows that the two approximations can be controlled independently and better results are within reach if we explore more. On top of that, the best setting of $\text{MBR}_{\text{N-by-N}}$ takes 164,025 utility assessments per

[3]The standard error of the mean scales with the inverse square root of the sample size.

decoding, $\text{MBR}_{\text{N-by-S}}$ with $S = 100$ brings this number down to 100,500 for the largest N considered, while improving BEER scores on all language pairs. We note that again increasing either N or S generally improves translation quality in our experiments. This further strengthens our previous finding that sampling-based MBR does not seem to have an equivalent of the beam search curse.

5.4 Choice of Hypothesis Space

While our focus thus far has been on reducing the number of target utility calls, allowing the exploration of larger $\bar{\mathcal{H}}(x)$, one should also take sampling time in consideration. For example, we found that in $\text{MBR}_{\text{N-by-N}}$ with $N = 100$, sampling time made up about 60% of the total translation time for our setup. Therefore, it is computationally attractive to construct compact $\bar{\mathcal{H}}(x)$ with promising translation candidates. Ideally, for better search in MBR, we enumerate a set of high expected utility hypotheses. Up until now we have constructed $\bar{\mathcal{H}}(x)$ using ancestral samples, following Eikema and Aziz (2020). Strategies like nucleus sampling and beam search are known empirically to produce higher quality translations than ancestral sampling and might therefore also enumerate outcomes that have high expected utility. We explore ancestral sampling, nucleus sampling and beam search. In a hyperparameter search we found $p = 0.7$ for nucleus sampling to work best. For beam search we use a length penalty of 1.2 (ne) or 0.6 (de, ro). We compare each strategy by the expected BEER values of the translations generated, using accurate estimates of expected BEER (using 1,000 samples for MC estimation). We show results in Figure 6.

We find ancestral sampling to produce hypotheses across the entire range of expected BEER scores. Nucleus sampling and beam search generally produce translations at the higher end of expected BEER. Therefore, these seem more suitable for generating effective $\bar{\mathcal{H}}(x)$ at smaller N. Nucleus sampling seems to lead to the largest proportion of high expected utility translations across language pairs. Beam search has a noticeably high proportion of poor translations for English-Nepali, a low-resource language pair where mode-seeking search has been observed to be less reliable. Results in the opposite direction were similar. We explore both nucleus sampling and beam search for constructing $\bar{\mathcal{H}}(x)$ in the next experiment, as well as combining all three strategies together.

5.5 Coarse-to-Fine MBR

We now turn to the coarse-to-fine procedure (MBR_{C2F}) described in Section 3.

5.5.1 Choice of Proxy Utility

We compare various proxy utilities by their effectiveness as filtering strategies in obtaining high expected utility sets, where we again use accurate estimates of expected utility using 1,000 samples for MC estimation. We filter the top-20 hypotheses from an initial 100 hypotheses obtained using ancestral sampling. This ensures a high variety of expected utilities in the initial set. We also compare each proxy utility on their runtime performance. We compare both cheap estimates of expected BEER using either 1 or 5 samples for MC estimation (BEER-1 and BEER-5 respectively) as well as cheap-to-compute proxy metrics: unigram F1 using 50 samples for MC estimation (UF-50) and skip-bigram F1 [4] using 50 samples for MC estimation (SBF-50). We use expected BEER using 100 samples for MC estimation (BEER-100) as a reference point. See our results on the English-German system in Figure 2.

We surprisingly find nearly all strategies to lead to equally good filtered sets as BEER-100 in terms of expected BEER of the filtered set. The only strategy that performs slightly worse than the others is BEER-1, which is likely too noisy to be a reliable filtering strategy. We observed very similar results for the other five language pairs. In terms of runtime performance we find BEER-1 to be fastest followed by UF-50 at a 22.2x performance increase over BEER-100.[5] In follow-up experiments, we will use UF-50 as a proxy utility, providing high quality filtered sets at good runtime performance.

5.5.2 Coarse-to-Fine MBR Results

In Table 1 we compare MBR_{C2F} with $\text{MBR}_{\text{N-by-S}}$ using $N = 405$ nucleus samples ($p = 0.7$) to construct the hypothesis space. We filter the top-$T = 50$ hypotheses using UF-50 as proxy utility and use $L = 100$ samples for MC estimation of the top-set, following our findings in Sections 5.5.1 and 5.3 respectively. For $\text{MBR}_{\text{N-by-S}}$ we set $S = 13$ to roughly match the amount of computation available to MBR_{C2F}, based on a 22.2x speed-up of UF-50 relative to BEER-100 observed in Figure 7.

[4] Skip-bigrams are bigrams that do not enforce adjacency.

[5] Our Python implementations of unigram and skip-bigram F1 are not optimized and we deem it likely that a greater speed-up is possible with a more efficient implementation.

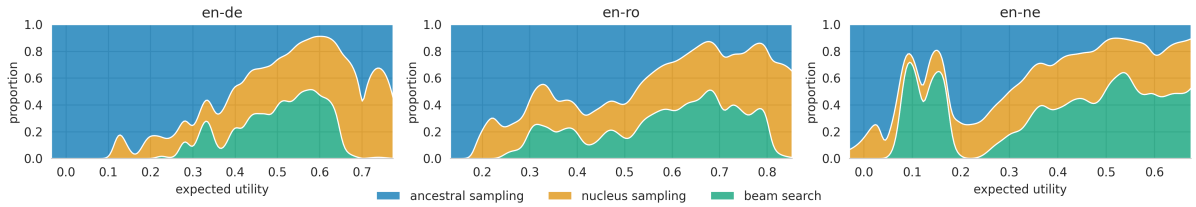

Figure 6: Proportion plots of expected utility for 3 strategies for constructing $\bar{\mathcal{H}}(x)$, using 100 translation candidates per strategy. We estimate expected utility using 1,000 samples. Results are aggregated over 100 source sentences.

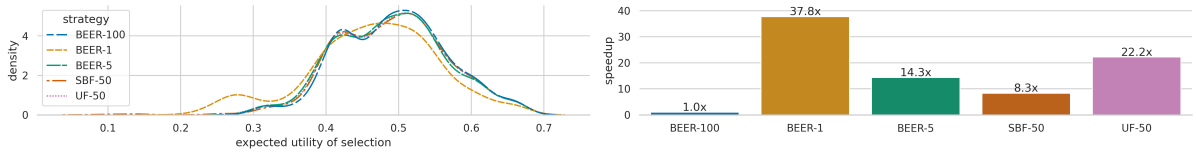

Figure 7: Comparison of proxy utilities on English to German: BEER using 1, 5 or 100 samples for MC estimation, and unigram F1 (UF) and skip-bigram F1 (SBF) each using 50 samples for MC estimation. We use each proxy utility to filter a top-20 from 100 ancestral samples. We show the resulting expected target utilties (BEER, an accurate estimate) (left), as well as a runtime comparison (right). Results are aggregated over 100 source sequences.

We find that across language pairs MBR_{C2F} consistently outperforms $\text{MBR}_{\text{N-by-S}}$ showing improvements between +0.4 and +1.1 BEER and +0.2 to +1.9 BLEU. MBR_{C2F} thus is effective at obtaining higher translation quality than $\text{MBR}_{\text{N-by-S}}$ at the same amount of computation available for MBR.

We also explore the effects on translation quality of changing and combining strategies for constructing $\bar{\mathcal{H}}(x)$. We find that using a beam of $N = 405$ (using the same length penalty as in Section 5.4) to construct $\bar{\mathcal{H}}(x)$ produces better results than nucleus sampling for most language pairs. Notably, reordering a large beam considerably improves over standard beam search decoding (using the usual beam size of 5 (ro, ne) or 4 (de)) for all language pairs in terms of BEER and for most language pairs in terms of BLEU scores. Combining all strategies for creating hypothesis spaces: ancestral sampling, nucleus sampling and beam search leads to the best results overall. For all language pairs both BEER and BLEU scores either improve or remain similar. This is more empricial evidence that expected utility is a robust and reliable criterion for picking translations: enlarging the hypothesis space or improving MC estimation under reasonable choices of hyperparameters seemingly never unreasonably hurts translation quality, but generally improves it.

6 Related Work

In recent NMT literature MBR has started being explored either in combination with MAP decoding or replacing it altogether. Stahlberg et al. (2017) adapt lattice minimum Bayes risk decoding (Tromble et al., 2008) on SMT translation lattices to be incorporated in left-to-right beam search decoding in NMT, thereby proposing a hybrid decoding scheme. They adapt lattice MBR to work on partial hypotheses and perform beam search to find translations that are both high probability under the NMT model and have high expected utility under the SMT model. Shu and Nakayama (2017) also combine beam search with MBR decoding to find low risk hypotheses, after which they re-rank all hypotheses with MBR again. They report having to restrict the number of hypotheses as not to degrade the effectiveness of MBR re-ranking, a finding that is likely due to biased estimation of expected utility, as in our work we find that increasing the number of hypotheses always improves translation quality. Blain et al. (2017) explore the quality of k-best lists obtained from beam search in NMT models and find that while MAP is not a good criterion for ranking the resulting hypotheses, re-ranking using MBR with BEER as a utility leads to improvements on top of standard beam search decoding (with a small beam size), in terms of both BLEU scores as well as human evaluation scores. Borgeaud and Emerson (2020) approach decoding from a voting theory perspective and derive a decoding strategy similar to MBR. They explore a range of utility functions, achieving similar BLEU scores to beam search, but showing improvements in terms of length, diversity and human judgement.

All of the above works make use of beam search to provide both the hypothesis space as well as to make a biased estimate of expected utility. Eikema and Aziz (2020) are the first work in NMT that propose to use sampling from the model to both make unbiased estimates of expected utility, the importance of which we confirm in experiments, and to form the hypothesis space. The authors only explore $\text{MBR}_{\text{N-by-N}}$, however, and never explore hypothesis spaces larger than $N = 30$ samples. We show that it is beneficial to scale MBR to much larger hypothesis spaces and that it can be beneficial to construct them using mode-seeking strategies. Müller and Sennrich (2021) study the properties of the sampling-based algorithm proposed in Eikema and Aziz (2020) and explore hypothesis spaces up to a size of $N = 100$ as well as multiple utility functions. They find that MBR decoding outputs exhibit a similar but smaller bias towards short translations and frequent tokens compared to beam search, but do observe that this is dependent on the choice of utility function. They further find that MBR decoding mitigates spurious copying and hallucinations under domain shift. Similar to our work, they find that MBR decoding scales well with larger hypothesis spaces and better estimation of expected utility. Freitag et al. (2021) explore the use of large hypothesis spaces and a range of utilities, including neural utilities, on the $\text{MBR}_{\text{N-by-N}}$ approximation. They find that using BLEURT as utility leads to significantly better translations in a human evaluation, while producing considerably lower probability translations.

We provide a more extensive overview of historical approximations to the MBR objective as well as an overview of alternatives for tackling the inadequacy of the mode in Appendix A.

7 Conclusion

We have shown MBR to be a robust decision rule for NMT that can find high quality translations. In particular, we have found that MBR, under reasonable hyperparameter choices, generally leads to improved translation quality with more computation (*i.e.*, searching a larger search space and/or using more samples for more accurate MC estimation). Big challenges in decoding with MBR are constructing the hypothesis space and keeping computational cost of estimating expected utility tractable. We have proposed effective strategies for both, by exploring more efficient ways of forming the hypothesis space and proposing an approximation to MBR that is linear in the size of this hypothesis space. Our coarse-to-fine MBR procedure is able to considerably reduce the number of calls to the utility function without compromising translation quality. We have shown that sampling-based MBR in general can outperform beam search on all the language pairs we explored and can continue to improve with better and more accurate search. We believe sampling-based MBR to be a promising, albeit still more expensive, alternative to beam search decoding. Unlike beam search, where it is not obvious how to further improve translation quality, sampling-based MBR is likely to benefit from improvements of different aspects of the algorithm. We believe fruitful avenues of research to be among *i)* clever algorithms for constructing hypothesis spaces, *ii)* more robust estimates of expected utility using fewer samples, *iii)* use of modern neural utilities and *iv)* improving the modelling capacity of NMT systems. We hope that this work motivates researchers and practitioners to make more conscious considerations of the choice of decision rule and that it paves the way for use of tractable sampling-based MBR decoding in NMT.

		en-de		en-ro		en-ne	
MBR	$\bar{\mathcal{H}}$	BEER	BLEU	BEER	BLEU	BEER	BLEU
NxS	N	64.3	38.0	54.9	21.4	38.9	3.6
C2F	N	+1.1	+1.9	+0.4	+0.2	+0.4	+0.2
	B	+0.9	+1.5	**+0.5**	**+0.5**	+0.5	**+0.5**
	all	**+1.3**	+2.4	**+0.5**	+0.4	**+0.6**	**+0.5**
BS	-	+0.9	**+2.8**	-0.1	+0.1	-0.8	+0.2

		de-en		ro-en		ne-en	
MBR	$\bar{\mathcal{H}}$	BEER	BLEU	BEER	BLEU	BEER	BLEU
NxS	N	64.8	38.7	58.5	28.0	43.1	6.3
C2F	N	+0.9	+1.1	+0.5	+0.7	+0.5	+0.2
	B	**+1.0**	**+1.5**	**+0.7**	**+1.2**	+0.5	**+0.9**
	all	**+1.0**	+1.4	+0.6	+1.1	**+0.8**	+0.8
BS	-	+0.5	+1.2	-0.0	+0.8	-1.0	+0.4

Table 1: Comparing $\text{MBR}_{\text{N-by-S}}$, MBR_{C2F} and beam search (BS) in terms of BEER and BLEU performance. We use BEER as utility, UF-50 as proxy utility, set top-$T = 50$ and use $L = 100$ samples for MC estimation. We use various strategies for constructing $\bar{\mathcal{H}}(x)$: 405 nucleus samples (N), the 405-best list from beam search (B) and combining both of these along with 1,005 ancestral samples (all). We use $S = 13$ in $\text{MBR}_{\text{N-by-S}}$ to mimic the computational cost of MBR_{C2F} at $N = 405$. The last row shows standard beam search performance using a typical beam size of 4 or 5 depending on the language. MBR results are averaged over 3 runs.

References

Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez, and Philipp Koehn. 2009. Monte Carlo inference and maximization for phrase-based translation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 102–110, Boulder, Colorado. Association for Computational Linguistics.

Wilker Aziz, Marc Dymetman, and Sriram Venkatapathy. 2013. Investigations in exact inference for hierarchical translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 472–483, Sofia, Bulgaria. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR, 2015*, San Diego, USA.

James O Berger. 1985. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *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 4528–4537, Online. Association for Computational Linguistics.

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

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A Gibbs sampler for phrasal synchronous grammar induction. 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 782–790, Suntec, Singapore. Association for Computational Linguistics.

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. 2016a. 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.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. 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.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii. Association for Computational Linguistics.

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. 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 567–575, Suntec, Singapore. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of WMT, 2011*, pages 85–91, Edinburgh, Scotland.

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.

Martina Forster, Clara Meister, and Ryan Cotterell. 2021. Searching for search errors in neural morphological inflection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1388–1394, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021. Minimum bayes risk decoding with neural metrics of translation quality.

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

Joshua Goodman. 1996. Parsing algorithms and metrics. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 177–183, Santa Cruz, California, USA. Association for Computational Linguistics.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *ICML Workshop on Representation Learning*, volume abs/1211.3711.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

T. Jaeger and Roger Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. 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.

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. 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.

Rémi Leblond, Jean-Baptiste Alayrac, Laurent Sifre, Miruna Pislar, Jean-Baptiste Lespiau, Ioannis Antonoglou, Karen Simonyan, and Oriol Vinyals. 2021. Machine translation decoding beyond beam search. *arXiv preprint arXiv:2104.05336*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. 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. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

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.

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, Stockholmsmässan, Stockholm Sweden. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL, 2002*, pages 311–318, Philadelphia, USA.

Ben Peters and André F. T. Martins. 2021. Smoothing and shrinking the sparse Seq2Seq search space. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654, Online. Association for Computational Linguistics.

Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy. Association for Computational Linguistics.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018a. A call for clarity in reporting BLEU scores. In *Proceedings of WMT, 2018*, pages 186–191, Brussels, Belgium.

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, Brussels, Belgium. Association for Computational Linguistics.

Christian P. Robert and George Casella. 2010. *Monte Carlo Statistical Methods*. Springer Publishing Company, Incorporated.

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

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.

Pavel Sountsov and Sunita Sarawagi. 2016. Length bias in encoder decoder models and a case for global conditioning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1525, Austin, Texas. 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, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. 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.

Miloš Stanojević and Khalil Sima'an. 2015. Reordering grammar induction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 44–54, Lisbon, Portugal. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *NIPS, 2014*, pages 3104–3112. Montreal, Canada.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 6000–6010.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Yonghui Wu, M. Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Qin Gao, Klaus Macherey, J. Klingner, Apurva Shah, M. Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, George Kurian, Nishant Patil, W. Wang, C. Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. Corrado, Macduff Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Hao Zhang and Daniel Gildea. 2008. Efficient multi-pass decoding for synchronous context free grammars. In *Proceedings of ACL-08: HLT*, pages 209–217, Columbus, Ohio. Association for Computational Linguistics.

A Additional Related Work

A.1 Approximations to MBR

Most instances of MBR decoding in machine translation, from the original work of Kumar and Byrne (2004) to recent instances in NMT (Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017), approximate the objective function by computing expectations not w.r.t. the model distribution, but rather, w.r.t. a proxy distribution. This proxy is obtained by enumeration via beam-search of a subset of the sample space (*e.g.*, a k-best list), and renormalisation of the probabilities of the outcomes in this subset. This has the undesirable effect of exaggerating differences in probability due to underestimation of the normalisation constant, and, like MAP decoding, it over-represents pathologies around the mode. Similarly, most prior work uses mode-seeking search to explore a tractable subset of the hypothesis space. Mode-seeking approximations bias the decoder towards the mode making MBR decoding less robust to idiosyncratic outcomes in the hypothesis space (Eikema and Aziz, 2020).This is in stark contrast with our work, where we sample from the model to construct unbiased estimates of expected utility, as well as to enumerate a tractable hypothesis space.

There are cases in statistical machine translation (SMT) where the computation of expected utility can be factorised along a tractable directed acyclic graph (DAG) via dynamic programming (Tromble et al., 2008; Zhang and Gildea, 2008; DeNero et al., 2009; Kumar et al., 2009). In such cases, the DAG contains a much larger subset of the sample space than any practical k-best list, still some pruning is necessary to construct a compact DAG containing only the most probable outcomes. These strategies are only available for models and utility functions that make strong Markov assumptions. For example, Tromble et al. (2008) and DeNero et al. (2009) develop linearisation strategies for BLEU, and Zhang and Gildea (2008) maximise expected trigram counts as a proxy to BLEU proper. The idea of utilising a proxy utility is something we also explore in this paper, though only as an intermediate step to decoding with the target utility.

In some (rarer) cases, unbiased (or asymptotically unbiased) samples have been used to approximate the MBR objective and/or to reduce the search space. For example, Stanojević and Sima'an (2015) use ancestral sampling in MBR decoding for permutation-trees-based reordering models, and Arun et al. (2009) use Gibbs sampling for MBR decoding in phrase-based MT. Unbiased samples for estimation of expected utility or exploration of a tractable hypothesis space are simply not common in machine translation. In SMT, the reason is a technical one, most SMT models are not based on a left-to-right factorisation of the joint distribution, thus unbiased sampling requires MCMC (DeNero et al., 2008; Blunsom et al., 2009) or expensive adaptive rejection sampling (Aziz et al., 2013). This limitation does not extend to NMT models, but NMT most likely simply inherited from SMT the practice of using beam-search-based approximations, at least until the work of Eikema and Aziz (2020).

A.2 Tackling the Inadequacy of the Mode

Eikema and Aziz (2020) link the inadequacy of the mode in NMT to the entropy of the conditional distribution, or, more precisely, to the fact that NMT models tend to spread probability mass over large subsets of the sample space (Ott et al., 2018). It is plausible that strategies to concentrate probability mass (*e.g.*, reducing entropy or pruning the

support of the model) will do so by making inadequate translations less probable. For example, Forster et al. (2021) find that the inadequacy of the mode problem does not seem to affect sequence-to-sequence models of morphological inflection, an essentially deterministic task, whose combinatorial space is built upon a smaller vocabulary (*i.e.*, characters instead of sub-word units), and whose observations are typically very short (*i.e.*, words rather than sentences). Peters and Martins (2021) train sparse sequence-to-sequence models (Peters et al., 2019) which assign zero probability to many outcomes dramatically reducing the support of the conditional distribution over complete sequences. They show that sparsity leads to inadequate candidates such as the empty string being pruned out of the support. They also find that label smoothing increases the rate at which the empty string is more probable than the beam-search output.

Meister et al. (2020) interprets the algorithmic approximations of beam search as an inductive bias towards outputs with uniform information density (Jaeger and Levy, 2007). They develop variants of beam search where this preference is a tunable hyperparameter and show that deviating from the mode with this type of bias can lead to improved translation quality. Another way to deviate from the mode is to augment the decoding objective with an auxiliary model. Li and Jurafsky (2016) re-rank a k-best list using a combination of two model probabilities, namely, $p_{Y|X}(h|x, \theta_{\text{fwd}})$ and $p_{X|Y}(x|h, \theta_{\text{bwd}})$. They think of this as maximising the mutual information (MI) between source and translation. The motivation is that the target-to-source component will push against inadequate candidates, as those are unlikely to be mapped back to the source with high probability. Bhattacharyya et al. (2021) find that 100 samples from an NMT model contain better candidates (measured in terms of BLEU) than the output of beam search (an observation Eikema and Aziz (2020) also make based on 30 samples and METEOR, instead). They propose to rerank these samples using an energy-based model trained to order candidates as sentence-BLEU would. Like these works, sampling-based MBR decoding, can be seen as a form of *explore and rank* approach, however, the ranking function in MBR is derived from the NMT model itself, whereas both MI- and EBM-based re-ranking involve an auxiliary trained model. For the EBM, in particular, in the limit of a too large hypothesis space, the beliefs of the NMT model are completely overwritten by the EBM. MBR, instead, does not overwrite the model's beliefs, it re-expresses those beliefs in terms of utility.

Leblond et al. (2021) recast NMT as a reinforcement learning problem and learn both a policy (*i.e.*, a mechanism to explore the space of translations one word at a time from left-to-right) and a value function (*i.e.*, an estimate at the expected reward of finishing a given prefix translation). For reward they investigate what they call privileged metrics, which require access to references (*e.g.*, sentence-level BLEU), and unprivileged metrics, which do not use references but access the source (*e.g.*, a quality estimation score). Compared to sampling-based MBR, their work tightly integrates search and value estimation, thus going beyond ranking a fixed set of candidates. The objective function of MBR can be thought of as an 'unprivileged metric' in their terminology, one that is based on the NMT model itself (and a choice of utility). But, the policy in sampling-based MBR (*i.e.*, the NMT model) is not trained to be aware of the evaluation metric.

B Comparing Target Utilities

We compare a number of utility functions for use in MBR decoding. In principle any function that measures some notion of similarity across sequences and can be reliably assessed on the sentence-level is suitable as a utility function for MBR. As BLEU is the predominant automatic evaluation metric on which translation quality is assessed, we experiment with a smoothed version of BLEU (Papineni et al., 2002) that can work on the sentence-level: sentence-BLEU (Chen and Cherry, 2014) using the default parameters in Post (2018b). We further try METEOR (Denkowski and Lavie, 2011) as this was used in Eikema and Aziz (2020) and showed good results.[6] BEER (Stanojević and Sima'an, 2014) is a character-based metric that has shown to correlate well with human judgements in many WMT metrics tasks (Macháček and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016b). Finally, we also explore ChrF++ (Popović, 2017), another character based metric that is an improved version of ChrF (Popović, 2015).

We perform $\text{MBR}_{\text{N-by-S}}$ with $N = 405$ and $S = 100$ in order to perform the comparisons. We

[6] We use a slightly different version of METEOR than in Eikema and Aziz (2020). We use language-specific versions rather than a language-agnostic version used in that work.

Task	Utility	BEER	BLEU	METEOR	ChrF++
en-de	BEER	**64.3**	37.0	56.6	61.3
	sentence-BLEU	63.3	**37.5**	55.9	60.2
	METEOR	62.5	33.4	**57.8**	60.5
	ChrF++	63.2	34.9	56.9	**61.4**
de-en	BEER	**64.9**	38.0	39.3	61.0
	sentence-BLEU	64.3	**38.3**	38.9	60.3
	METEOR	63.5	36.1	**39.7**	59.8
	ChrF++	64.4	37.2	39.5	**61.5**
en-ro	BEER	**54.8**	21.0	33.9	47.8
	sentence-BLEU	54.4	**21.3**	40.4	47.4
	METEOR	54.5	20.9	**40.9**	47.7
	ChrF++	54.2	20.2	40.3	**48.0**
ro-en	BEER	**58.4**	27.5	32.4	52.0
	sentence-BLEU	57.8	**27.8**	32.2	51.4
	METEOR	57.5	26.6	**32.9**	51.5
	ChrF++	58.0	27.1	32.7	**52.6**
en-ne	BEER	**38.4**	**3.4**	11.0	26.1
	sentence-BLEU	34.9	3.0	10.9	22.7
	METEOR	37.3	**3.4**	**13.2**	25.3
	ChrF++	36.8	2.6	12.3	**26.6**
ne-en	BEER	**42.7**	**6.0**	17.0	31.2
	sentence-BLEU	39.9	5.7	15.1	28.4
	METEOR	40.4	4.6	**17.3**	30.8
	ChrF++	40.6	4.8	17.0	**32.0**

Table 2: Comparing BEER, sentence-BLEU, METEOR and ChrF++ as utility functions in $\text{MBR}_{\text{N-by-S}}$ using $N = 405$ and $S = 100$.

measure the performance of each utility on BEER, BLEU, METEOR and ChrF++. Our results are shown in Table 2. As expected, using a certain utility achieves the best performance under the lens of that metric as well. Sometimes we find a small deviation from this when BEER or METEOR outperforms sentence-BLEU in terms of BLEU score. This is likely due to sentence-BLEU only being an approximation to BLEU itself. We find that overall BEER seems to do best across metrics followed by ChrF++. Herefore, in the main paper, we have used BEER as the utility of choice. The finding that BEER works well as a utility function in MBR was also made before in the work of Blain et al. (2017).