Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine Translation

Mathias Müller[1] and **Rico Sennrich**[1,2]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh

Abstract

Neural Machine Translation (NMT) currently exhibits biases such as producing translations that are too short and overgenerating frequent words, and shows poor robustness to copy noise in training data or domain shift. Recent work has tied these shortcomings to beam search – the de facto standard inference algorithm in NMT – and Eikema and Aziz (2020) propose to use Minimum Bayes Risk (MBR) decoding on unbiased samples instead.

In this paper, we empirically investigate the properties of MBR decoding on a number of previously reported biases and failure cases of beam search. We find that MBR still exhibits a length and token frequency bias, owing to the MT metrics used as utility functions, but that MBR also increases robustness against copy noise in the training data and domain shift.[1]

1 Introduction

Neural Machine Translation (NMT) currently suffers from a number of issues such as underestimating the true length of translations (Koehn and Knowles, 2017; Stahlberg and Byrne, 2019; Kumar and Sarawagi, 2019), underestimating the probability of rare words and over-generating very frequent words (Ott et al., 2018), or being susceptible to copy noise in the training data (Khayrallah and Koehn, 2018). In out-of-domain translation, *hallucinations* (translations that are fluent but unrelated to the source) are common (Koehn and Knowles, 2017; Lee et al., 2018; Müller et al., 2020).

Previous work has addressed these problems with decoding heuristics such as length normalization (Wu et al., 2016), data cleaning (Junczys-Dowmunt, 2018; Bañón et al., 2020) or model regularization (Bengio et al., 2015; Shen et al., 2016; Wiseman and Rush, 2016; Zhang et al., 2019; Ng et al., 2020).

Recently, Eikema and Aziz (2020) have highlighted the role of the decision rule, namely searching for the highest-scoring translation, and have argued that it is at least partially to blame for some of these biases and shortcomings. They found that sampling from an NMT model is faithful to the training data statistics, while beam search is not. They recommend the field look into alternative inference algorithms based on unbiased samples, such as Minimum Bayes Risk (MBR) decoding.

We believe MBR has potential to overcome several known biases of NMT. More precisely, if a bias can be understood as being caused by the mode-seeking nature of beam search then we hypothesize that MBR could exhibit less bias. We view short translations, copies of the source text and hallucinations as hypotheses that are probable, but quite different to other probable hypotheses. If such pathological hypotheses are in a pool of samples, it is unlikely that MBR would select them as the final translation.

While Eikema and Aziz (2020) compare the statistical properties of samples and beam search outputs, and show that MBR can perform favourably compared to beam search according to automatic metrics, our paper aims to perform a targeted study of MBR and its properties, specifically its effects on the biases and shortcomings discussed previously. In our experiments we find that

- If used with a utility function that favours short translations, MBR inherits this bias;
- MBR still exhibits a token probability bias in that it underestimates the probability of rare tokens and overestimates very common tokens;
- Compared to beam search, MBR decoding is more robust to copy noise in the training data;

[1] Code and documentation available at `https://github.com/ZurichNLP/understanding-mbr`

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 259–272
August 1–6, 2021. ©2021 Association for Computational Linguistics

- MBR exhibits higher domain robustness than beam search. We demonstrate that MBR reduces the amount of hallucinated content in translations.

2 Background

2.1 Maximum-a-posteriori (MAP) decoding

The de facto standard decoding algorithm in NMT is beam search (Graves, 2012; Boulanger-Lewandowski et al., 2013; Sutskever et al., 2014). Beam search belongs to a broader class of inference procedures called maximum-a-posteriori (MAP) algorithms. What MAP algorithms have in common is that they attempt to find the most probable translation under a given model. Essentially, they try to recover the *mode* of the output distribution over sequences.

An exact solution to this search problem is usually intractable. Beam search is an approximation that is tractable, but it also frequently fails to find the true mode of the distribution (Stahlberg and Byrne, 2019).

2.2 Known deficiencies of NMT systems

NMT systems are known to be deficient in a number of ways. We describe here only the ones relevant to our discussion and experiments.

Length bias: Systems underestimate the true length of translations. On average, their translations are shorter than references (Koehn and Knowles, 2017; Stahlberg and Byrne, 2019; Kumar and Sarawagi, 2019).

Skewed word frequencies: In translations, tokens that occur frequently in the training data are overrepresented. On the other hand, rare tokens occur fewer times than their probability in the training data would suggest (Ott et al., 2018).

Beam search curse: Increasing the beam size leads to finding translations that are more probable under the model. In theory, this should improve translation quality. Paradoxically, empirical results show that large beam sizes decrease quality (Koehn and Knowles, 2017; Ott et al., 2018).

Susceptibility to copy noise: Copied content in the training data disproportionately affects translation quality. More specifically, the most detrimental kind are copies of the source sentence on the target side of the training data (Khayrallah and Koehn, 2018). If such copies are present in the training data, copy hypotheses will be overrepresented in beam search (Ott et al., 2018).

Low domain robustness: Systems are not robust under distribution shifts such as domain shift. Having a system translate in an unknown test domain often does not gradually degrade translation quality, but leads to complete failure cases called hallucinations (Lee et al., 2018; Koehn and Knowles, 2017; Müller et al., 2020).

Much past research has attributed those deficiencies to model architectures or training algorithms, while treating beam search as a fixed constant in experiments. In contrast, Eikema and Aziz (2020) argue that the fit of the model is reasonable, which means that neither the model itself nor its training can be at fault. Rather, they argue that the underlying problem is beam search.

Inadequacy of the mode: Stahlberg and Byrne (2019) and Eikema and Aziz (2020) suggest that the mode of the distribution over output sequences is in fact not the best translation. On the contrary, it seems that in many cases the mode is the empty sequence (Stahlberg and Byrne, 2019). In addition, it appears that the probability of the mode is not much different from very many other sequences, as the output distribution is quite flat in an extensive region of output space (Eikema and Aziz, 2020).

Intuitively, it makes sense that such a situation could arise in NMT training: maximum likelihood estimation training does not constrain a model to be characterized well by its mode only. If the mode is inadequate, then obviously that is problematic for a mode-seeking procedure such as beam search, and MAP inference in general. In fact, MAP decoding should be used only if the mode of the output distribution can be trusted (Smith, 2011).

An alternative is a decision rule that considers how different a translation is from other likely translations.

2.3 Minimum Bayes Risk Decoding

MBR decoding was used in speech recognition (Goel and Byrne, 2000) and statistical machine translation (Kumar and Byrne, 2004; Tromble et al., 2008). More recently, MBR was also used to improve beam search decoding in NMT (Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017). Eikema and Aziz (2020) are the first to test a variant of MBR that operates on samples instead of an nbest list generated by beam search.

We give here a simplified, accessible definition of MBR in the context of NMT. Essentially, the goal of MBR is to find not the most probable translation, but the one that minimizes the expected risk for a given loss function and the true posterior distribution. In practice, the set of all possible candidate translations can be approximated by drawing from the model a pool of samples $\mathcal{S}$ of size n:

$$\mathcal{S} = (s_1, ..., s_n) \sim p(y|x, \theta). \quad (1)$$

The same set of samples can also be used to approximate the true posterior distribution. Then for each sample s_i in $\mathcal{S}$, its expected utility (the inverse risk) is computed by comparing it to all other samples in the pool. The sample with the highest expected utility is selected as the final translation:

$$y^{\star} = \operatorname*{argmax}_{s_i \in \mathcal{S}} \frac{1}{n} \sum_{s_j=1}^{n} u(s_i, s_j) \quad (2)$$

The size of the pool n and the utility function u are hyperparameters of the algorithm. A particular utility function typically computes the *similarity* between a hypothesis and a reference translation. Therefore, MBR "can be thought of as selecting a *consensus* translation [...] that is closest on average to all likely translations" (Kumar and Byrne, 2004).

3 Motivation for experiments

We hypothesize that MBR decoding is useful for a certain class of failure cases encountered with beam search. Namely, if an incorrect translation from beam search can be characterized as a hypothesis that is likely but fairly different from other hypotheses with similar probability, then MBR is expected to improve over beam search.

Several known deficiencies of NMT systems outlined in Section 2.2 belong to this class of beam search failures. For instance, length bias occurs when a beam search translation is shorter than other hypotheses with comparable probability. Likewise, translations that are copies of the input sentence or hallucinations (translations that are fluent, but unrelated to the input) can be avoided with MBR if they are not common in a pool of samples.

Finally, we study the skewedness of token frequencies in translations. Eikema and Aziz (2020) study lexical biases in NMT models, showing that model samples have higher agreement with the training distribution than MAP output. We investigate whether this is also true for MBR decoding, focusing on the well-known bias towards frequent tokens.

4 Experimental Setup

4.1 Data

We use data for a number of language pairs from the Tatoeba Challenge (Tiedemann, 2020). Individual language pairs are fairly different in terms of language families, scripts and training set sizes. See Appendix A for details about our data sets.

For one additional experiment on out-of-domain robustness we use data from Müller et al. (2020). This data set is German-English and defines 5 different domains of text (medical, it, koran, law and subtitles). Following Müller et al. (2020) we train our model on the medical domain, and use data in other domains to test domain robustness.

We hold out a random sample of the training data for testing purposes. The size of this sample varies between 1k and 5k sentences, depending on the overall size of the training data.

4.2 Models

Our preprocessing and model settings are inspired by OPUS-MT (Tiedemann and Thottingal, 2020). We use Sentencepiece (Kudo, 2018) with subword regularization as the only preprocessing step, which takes care of both tokenization and subword segmentation. The desired number of pieces in the vocabulary varies with the size of the data set.

We train NMT models with Sockeye 2 (Domhan et al., 2020). The models are standard Transformer models (Vaswani et al., 2017), except that some settings (such as word batch size and dropout rate) vary with the size of the training set. Following Eikema and Aziz (2020) we disable label smoothing so as to get unbiased samples.

4.3 Decoding and evaluation

In all experiments, we compare beam search to MBR decoding and in most cases also to single samples. For beam search, we always use a beam size of 5. Single samples are drawn at least 100 times to show the resulting variance.

If not stated otherwise, all results presented are on a test set held out from the training data, i.e. are certainly in-domain, which avoids any unintended out-of-domain effects.

We evaluate automatic translation quality with BLEU (Papineni et al., 2002), CHRF (Popović, 2016) and METEOR (Denkowski and Lavie, 2014). We compute BLEU and CHRF with SacreBLEU (Post, 2018). See Appendix B for details.

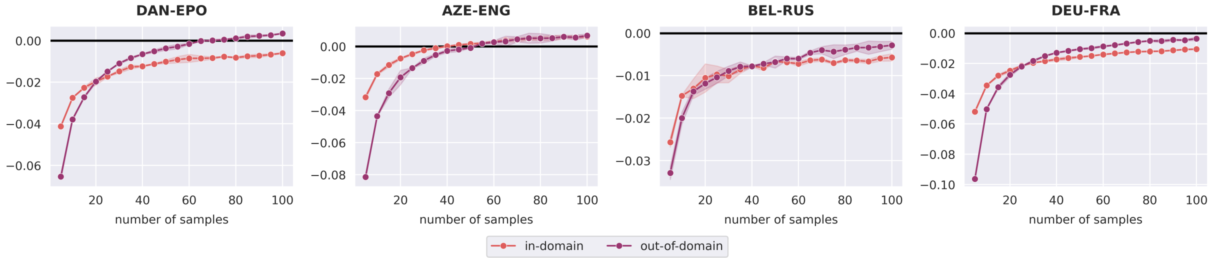

Figure 1: CHRF1 scores of MBR decoding on two test corpora: the standard Tatoeba test set (out-of-domain) and a test set of held-out training data (in-domain). Plots show the **difference** between MBR and beam search, as a function of the number of samples used for MBR.

	smoothed?	α	β	γ	δ
bleu	✗	-	-	-	-
bleu-floor	✓	-	-	-	-
bleu-add-k	✓	-	-	-	-
bleu-exp	✓	-	-	-	-
chrf-0.5	✗	-	0.5	-	-
chrf-1	✗	-	1.0	-	-
chrf-2	✗	-	2.0	-	-
chrf-3	✗	-	3.0	-	-
meteor	✗	0.85	0.2	0.6	0.75
meteor-0.5	✗	0.50	0.2	0.6	0.75

Table 1: Utility functions used with MBR. The smoothed variants of BLEU correspond to the ones implemented in SacreBLEU (Post, 2018) and are defined in Chen and Cherry (2014).

MBR also depends on samples, so we repeat each MBR experiment twice to show the resulting variance. We also vary the number of samples used with MBR, from 5 to 100 in increments of 5. Finally, we produce MBR translations with different utility functions. All of the utility functions are sentence-level variants of our evaluation metrics: BLEU, CHRF or METEOR. See Table 1 for an overview of utility functions. If not stated otherwise, MBR results are based on 100 samples and use `chrf-1` as the utility function.

5 Length bias

We evaluate MBR decoding with different utility functions. There is no single utility function which performs best on all evaluation metrics. Instead, any of our evaluation metrics can be optimized by choosing a closely related utility function (see Figure 2 and Appendix D). For instance, `chrf-2` as the utility function leads to the best CHRF2 evaluation scores.

Number of samples: We find that the translation quality of MBR increases steadily as the number of samples grows (see Figure 2). This means that MBR does not suffer from the beam search curse where single pathological hypotheses in a large beam can jeopardize translation quality.

We analyze the lengths of translations produced by different decoding methods in Table 2 (see Appendix E for additional statistics). We find that in terms of mean length of translations, beam search underestimates the true length of translations, even when hypotheses are normalized. Hypotheses generated by sampling better match the reference length. This is in line with the findings of Eikema and Aziz (2020).

For MBR decoding, it is clear that the choice of utility function has an impact on the mean length of the resulting translations. For instance, employing sentence-level BLEU as the utility function leads to translations that are too short. BLEU is a precision-based metric known to prefer shorter translations on the sentence level (Nakov et al., 2012).

`chrf-2` and `meteor` emphasize recall more, and the resulting MBR translations overestimate the true length of translations.[2] On the other hand, `chrf-0.5`, a CHRF variant with a bias for precision, leads to the shortest translations overall.

We test whether we can reduce length biases by symmetrizing our utility functions u as follows:

$$u_{sym}(s_i, s_j) = H(u(s_i, s_j), u(s_j, s_i)) \quad (3)$$

where H is the harmonic mean. This should avoid favouring either recall or precision, but in practice even symmetric utility functions lead to translations that are shorter than references on average.

Based on these observations we conclude that **MBR inherits length biases associated with its utility function**.

[2] While Popović (2016) find that the recall-biased CHRF2 achieves the highest correlation with human judgments as an evaluation metric, this does not entail that the same recall bias is optimal in the utility function for MBR.

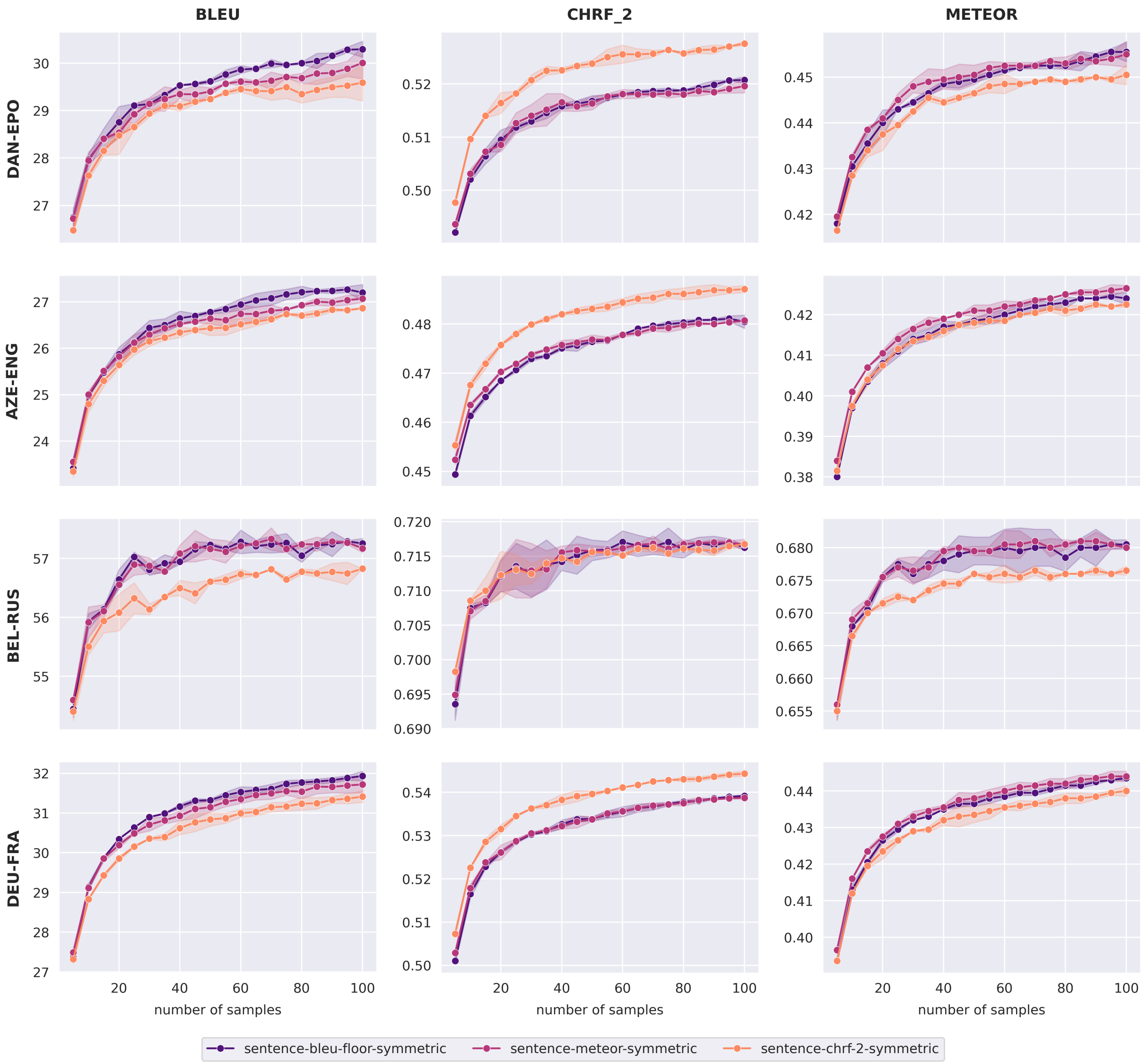

Figure 2: Comparison of MBR utility functions. Different columns show translation quality as measured by a particular evaluation metric. Line colors refer to different utility functions. Shaded areas show standard deviation.

	DAN-EPO	AZE-ENG	BEL-RUS	DEU-FRA
reference	11.91	15.54	8.41	20.19
sample	11.73	15.15	8.29	19.99
beam-normalized	11.61	14.45	8.23	19.62
beam-unnormalized	11.21	13.62	8.20	19.08
bleu-floor	11.51	14.41	8.18	19.55
meteor	12.23	15.29	8.26	20.38
chrf-2	12.50	15.88	8.31	20.89
bleu-floor-symmetric	11.51	14.34	8.19	19.53
meteor-symmetric	11.47	14.12	8.20	19.40
chrf-2-symmetric	11.48	14.16	8.18	19.40
chrf-0.5	10.63	12.99	8.08	18.02

Table 2: Lengths of hypotheses as mean number of tokens.

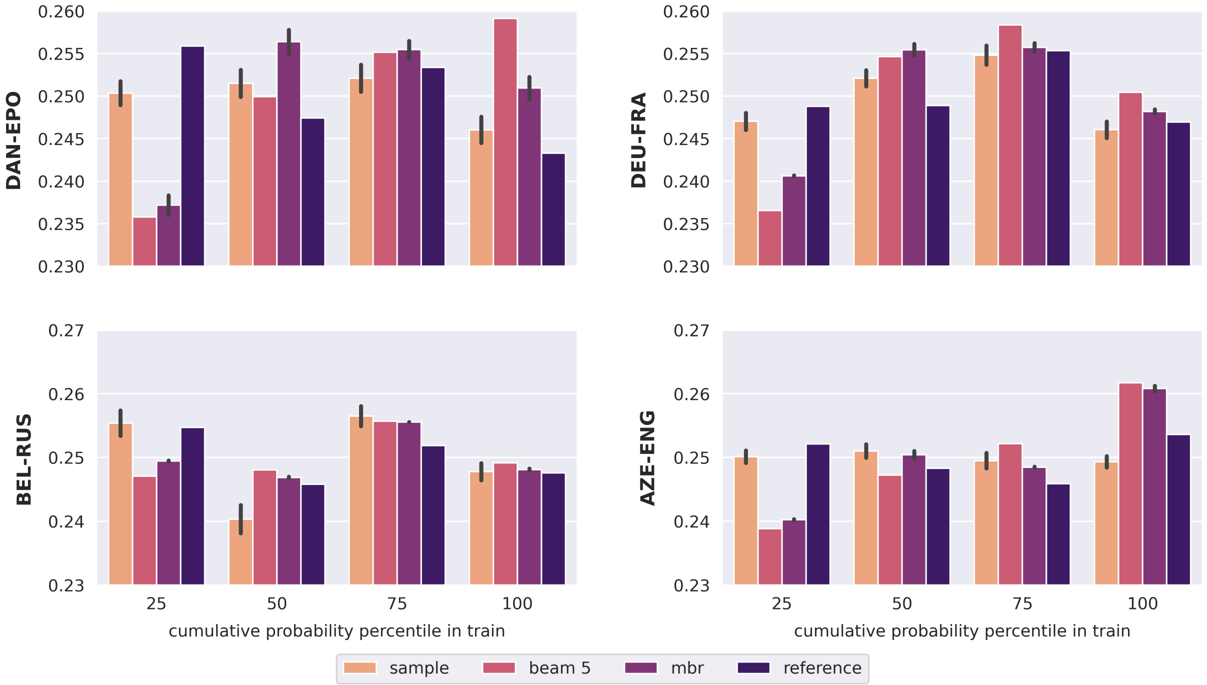

Figure 3: Probability of tokens in translations (x-axis) bucketed by frequency in training data (y-axis). Vertical bars indicate standard deviation for methods that involve sampling.

6 Token frequency bias

Beam search overgenerates tokens that are very common in the training data and undergenerates rare tokens (see Section 2.2). Sampling on the other hand assigns correct probabilities to common and rare tokens. Given that MBR is based on samples, does it share this property with sampling?

In Figure 3 we show that this is not the case. Although the skewedness of probabilities is less severe for MBR than for beam search, MBR still assigns too high a probability to frequent events. A reason for this is that our utility functions are based on surface similarity between samples, so rare tokens, which will be sampled rarely, will thus also have low utility.

Unfortunately, there is a **trade-off between correct probability statistics for very common and very rare words and translation quality**. The most faithful statistics can be obtained from sampling, but sampling leads to the worst overall translation quality.

7 Domain robustness

In general, as the number of samples grows, MBR approaches but does not outperform beam search on our in-domain data (see Figure 1). On our out-of-domain data, the gap between MBR and beam search is smaller. We hypothesize that MBR may be useful for out-of-domain translation.

We evaluate MBR on a domain robustness benchmark by Müller et al. (2020). Figure 4 shows that on this benchmark MBR outperforms beam search on 2 out of 4 unknown test domains. A possible reason why MBR is able to outperform beam search in unknown domains is that it reduces hallucinated translations. To test this hypothesis, we define a *hallucination* as a translation that has a CHRF2 score of less than 0.01 when compared to the reference, inspired by Lee et al. (2018).

Given this definition of hallucination, Figure 5 shows that on average, MBR assigns a lower utility score to hypotheses that are hallucinations. Similarly, MBR reduces the percentage of hallucinations found in the final translations, compared to beam search or sampling. To summarize, we find that **MBR decoding has a higher domain robustness than beam search**.

8 Impact of copy noise in the training data

If copies of source sentences are present on the target side of training data, copies are overrepresented in beam search (Section 2.2). Here we test whether MBR suffers from this copy bias as well.

We create several versions of our training sets where source copy noise is introduced with a proba-

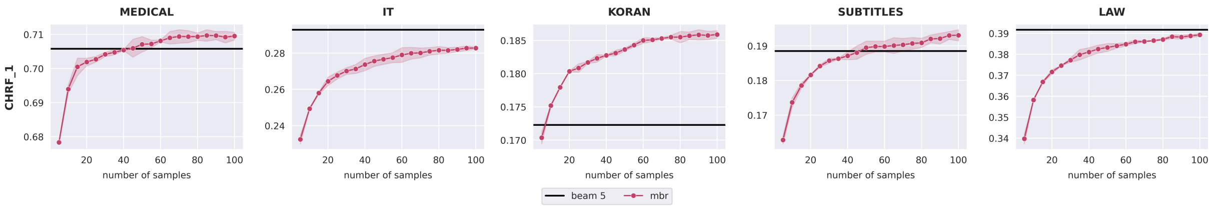

Figure 4: CHRF1 scores of MBR and beam search on the domain robustness benchmark of Müller et al. (2020). The *medical* test set is in-domain, the remaining sets are out-of-domain.

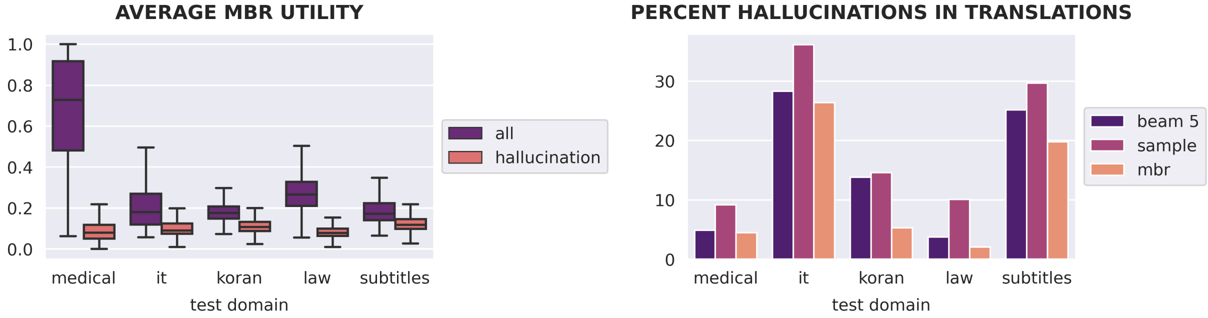

Figure 5: Analysis of hallucinations in MBR and beam translations. Left: Average utility of hallucination hypotheses in pools of samples. Right: how often hallucinations occur in final translations.

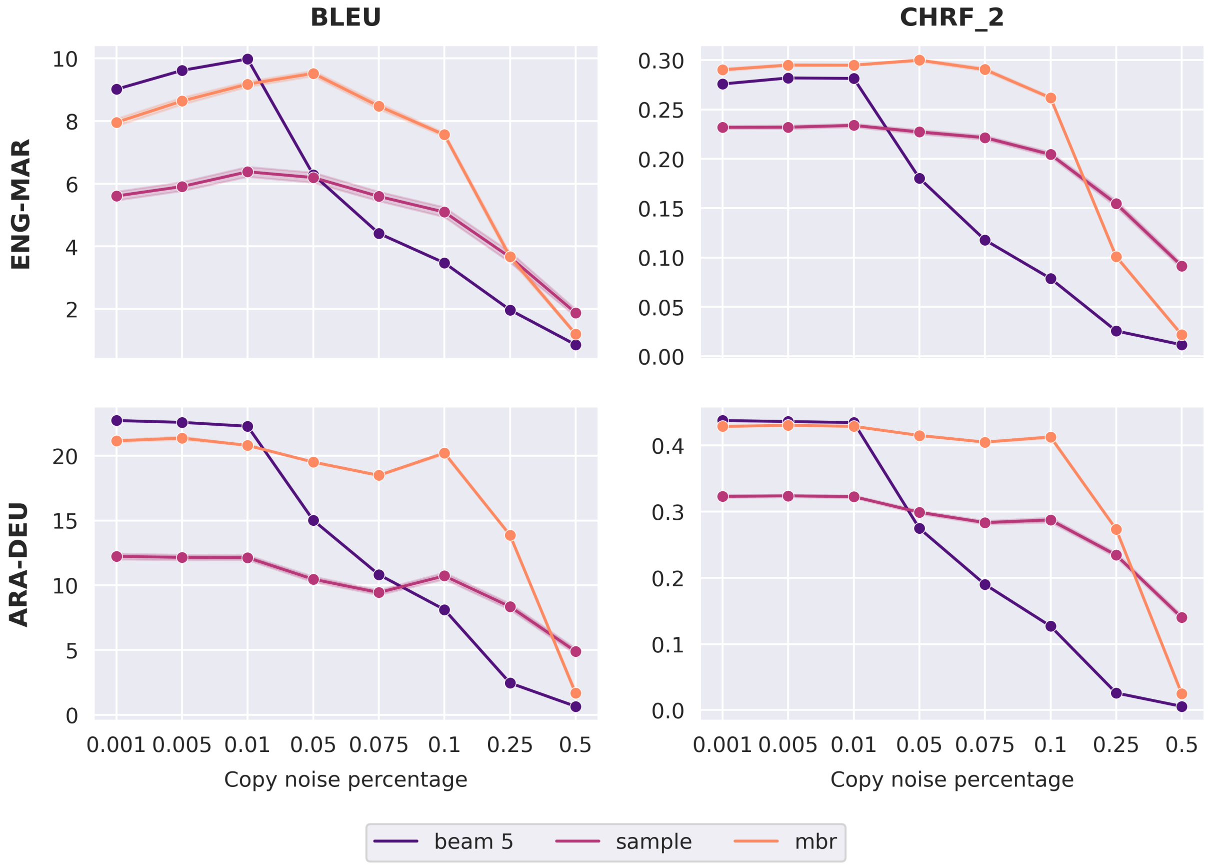

Figure 6: Susceptibility to copy noise in training data.

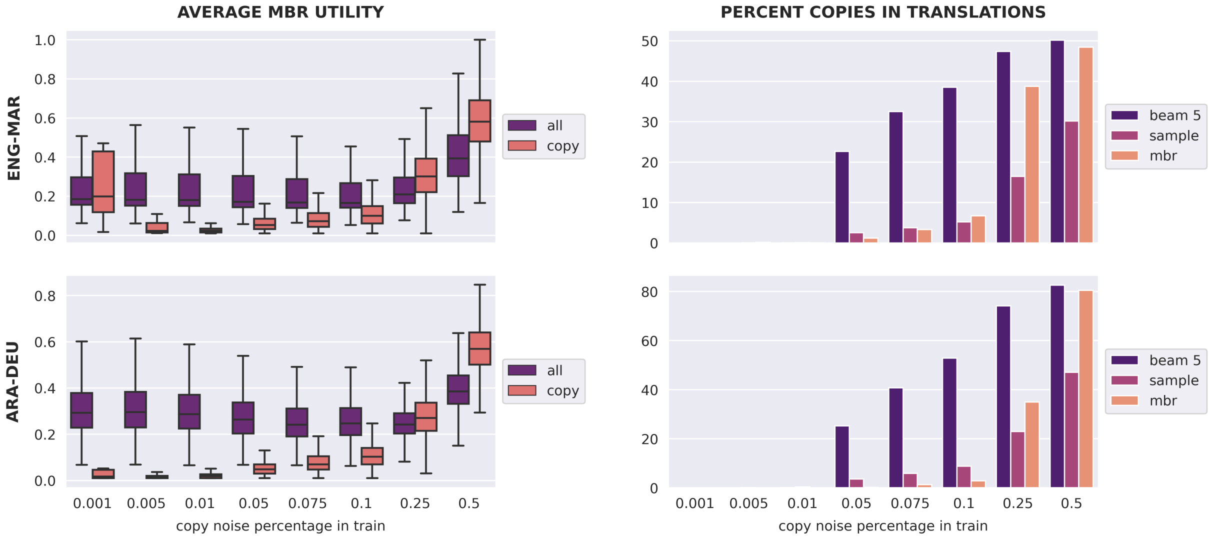

Figure 7: Analysis of copies in MBR and beam translations. Left: Average utility of copy hypotheses in pools of samples. Right: how often copies occur in final translations.

bility between 0.1% and 50%. As shown in Figure 6, MBR and beam search are comparable if there are few copies in the training data. However, if between 5 and 25% of all training examples are copies, then MBR outperforms beam search by a large margin (> 10 BLEU for Arabic-German).

As further evidence for the ability of MBR to tolerate copy noise we present an analysis of copies in Figure 7. We define a *copy* as a translation with a word overlap with the reference of more than 0.9. We show that MBR assigns a much lower utility to copy hypotheses than to all hypotheses taken together. In the final translations, MBR manages to reduce copies substantially. For instance, if around 10% of the training examples are copies, beam search produces around 50% copies, while MBR reduces this number to below 10%.

We conclude from this experiment that **MBR is more robust to copy noise in the training data**. We acknowledge that this setting is artificial because copy noise can easily be removed from data sets. Nonetheless, it is a striking example of a known shortcoming of NMT systems usually attributed to the model or training procedure, when in fact beam search is at least partially to blame.

9 Conclusion and future work

MBR decoding has recently regained attention in MT as a decision rule with the potential to overcome some of the biases of MAP decoding in NMT. We empirically study the properties of MBR decoding with common MT metrics as utility functions, and find it still exhibits a length bias and token frequency bias similar to beam search. The length bias is closely tied to the utility function. However, we also observe that MBR decoding successfully mitigates a number of well-known failure modes of NMT, such as spurious copying, or hallucinations under domain shift. The mechanism by which MBR achieves such robustness is that copies or hallucinated hypotheses in a pool of samples are assigned low utility and never selected as the final translation.

In our experiments, MBR did not generally outperform beam search according to automatic metrics, but we still deem it a promising alternative to MAP decoding due to its robustness. For future work, we are interested in exploring more sophisticated similarity metrics to be used as utility functions, including trainable metrics such as COMET (Rei et al., 2020), and investigating how these utility functions affect the overall quality and biases of translations.

10 Note on reproducibility

We will not only release the source code used to train our models (as is common in NLP papers at the moment), but a complete pipeline of code that can be run on any instance in a fully automated fashion. This will allow to reproduce our results, including the graphs and tables shown in this paper, in a consistent way with minimal changes. We encourage the community to attempt to reproduce our results and publish the results.

Acknowledgements

This work has received funding from the Swiss National Science Foundation (grant numbers 105212-169888 and 176727). Also, we have been assisted by the computing services of the University of Zurich (S3IT).

We would like to thank Bryan Eikema for his help with our implementation of MBR. We also thank Jörg Tiedemann, Annette Rios and Tannon Kew for helpful comments and discussion.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA. MIT Press.

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

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2013. Audio chord recognition with recurrent neural networks. In *ISMIR*, pages 335–340. Citeseer.

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.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

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.

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

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

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.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. 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. 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.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

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 (AMTA 2020)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

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.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

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

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.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. 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.

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.

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.

Raphael Shu and Hideki Nakayama. 2017. Later-stage minimum bayes-risk decoding for neural machine translation. *arXiv preprint arXiv:1704.03169*.

Noah A. Smith. 2011. *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1172–1180, Online. Association for Computational Linguistics.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference* *on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

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

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

A Data set details

ISO3 abbreviation	language pair	size	scripts
DAN-EPO	Danish-Esperanto	110k	Roman-Roman
AZE-ENG	Azerbaijani-English	680k	Roman*-Roman
BEL-RUS	Belarusian-Russian	70k	Cyrillic-Cyrillic
DEU-FRA	German-French	47m	Roman-Roman
ENG-MAR	English-Marathi	370k	Roman-Devanagari
ARA-DEU	Arabic-German	12m	Arabic-Roman
DEU-ENG	German-English	1m	Roman-Roman

Table 3: Details about data sets. Size refers to the number of sentence pairs in the training data. Roman* = Roman script with some modifications.

B Evaluation details

For evaluation metrics that require tokenization (BLEU and METEOR), we use the standard `mteval13a` tokenization implemented in SacreBLEU. We do not use any language-specific tokenization rules even if they are available for the target language. The SacreBLEU signatures for our CHRF and BLEU evaluation metrics are listed in Table 4.

evaluation metric	SacreBLEU signature
CHRF_1	chrF1+numchars.6+space.false+version.1.4.14
CHRF_2	chrF2+numchars.6+space.false+version.1.4.14
CHRF_3	chrF3+numchars.6+space.false+version.1.4.14
BLEU	BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.14

Table 4: SacreBLEU signatures of evaluation metrics.

C Comments on the development sets distributed with the Tatoeba challenge

The Tatoeba Challenge (Tiedemann, 2020) distributes training, development and test data for a large number of language pairs. What is peculiar about the challenge is that the training data is assembled from various sources through OPUS (Tiedemann, 2012), while the development and test data are contributed by users of Tatoeba[3]. This means that the development and test set can be considered out-of-domain material.

We investigated this issue and conclude that it does not constitute a problem. When both the development and test data are sampled from the training data, the results are similar to the ones we present in this paper, except for a small overall shift.

D Additional comparisons between utility functions

Figures 8 and 9 show additional results for MBR decoding with utility functions that are variants of CHRF and BLEU.

E Additional length tables

We provide additional length statistics for utility functions used with MBR in Table 5.

[3] https://tatoeba.org

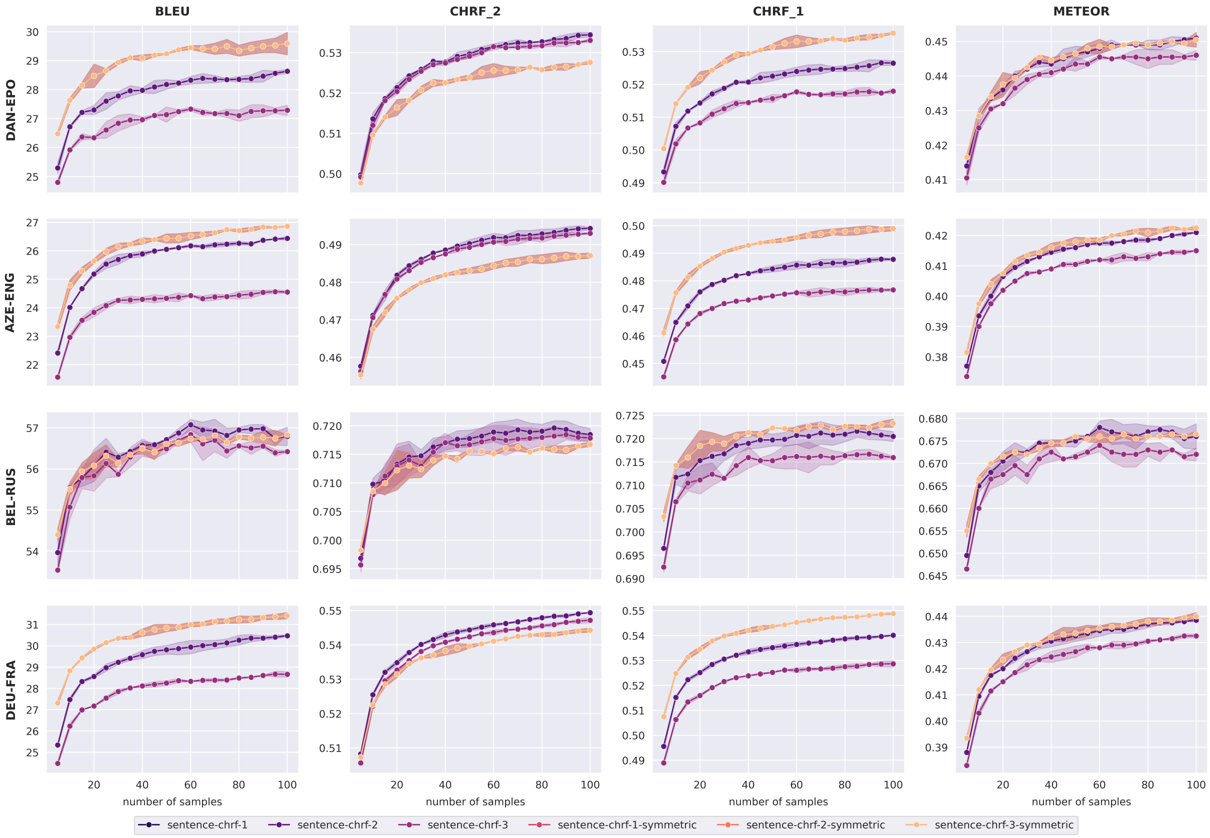

Figure 8: Comparison of utility functions that are variants of CHRF.

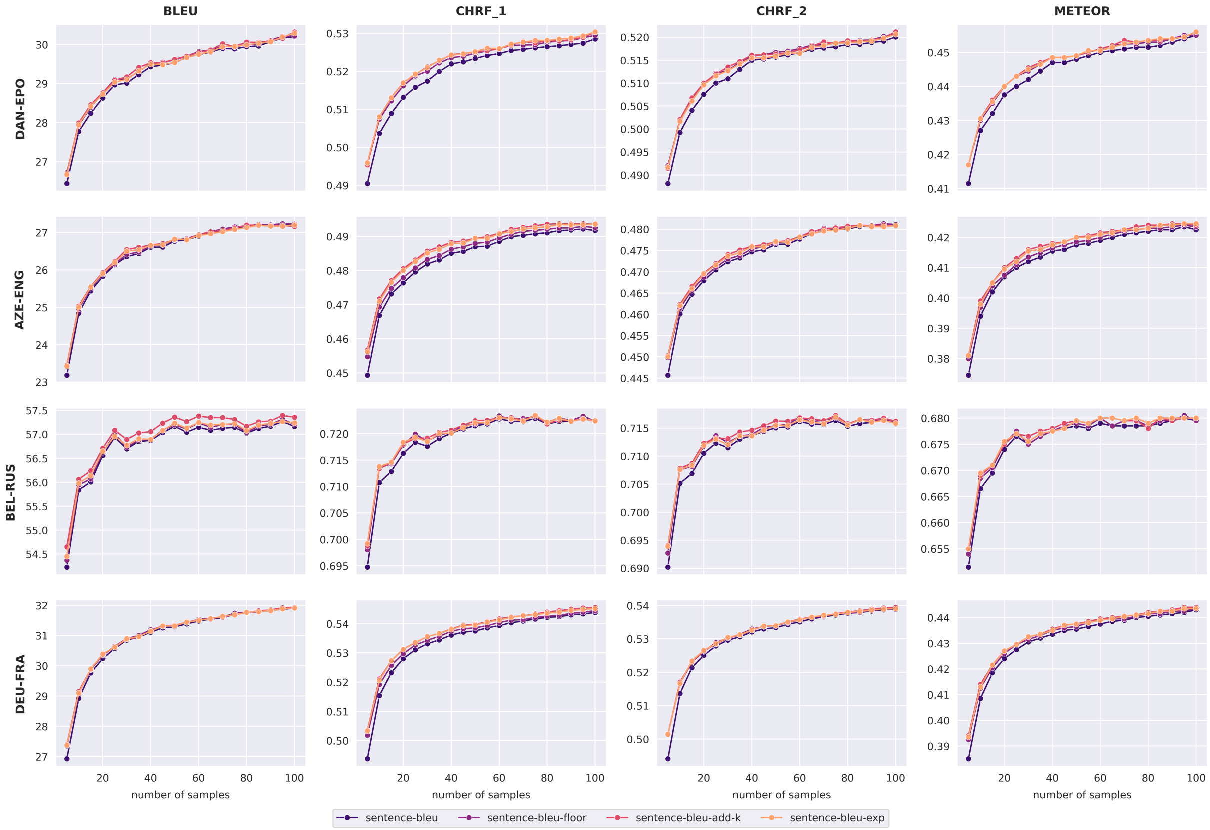

Figure 9: Comparison of utility functions that are variants of BLEU.

	DAN-EPO	**AZE-ENG**	**BEL-RUS**	**DEU-FRA**
reference	11.91	15.54	8.41	20.19
sample	11.73	15.15	8.29	19.99
beam-normalized	11.61	14.45	8.23	19.62
beam-unnormalized	11.21	13.62	8.20	19.08
bleu	11.54	14.45	8.17	19.59
bleu-floor	11.51	14.41	8.18	19.55
bleu-add-k	11.46	14.29	8.20	19.40
bleu-exp	11.42	14.29	8.18	19.41
bleu-symmetric	11.55	14.39	8.19	19.58
bleu-floor-symmetric	11.51	14.34	8.19	19.53
bleu-add-k-symmetric	11.39	14.14	8.19	19.25
bleu-exp-symmetric	11.41	14.21	8.18	19.37
chrf-1	11.48	14.16	8.18	19.40
chrf-2	12.50	15.88	8.31	20.89
chrf-3	13.01	16.92	8.45	21.93
chrf-1-symmetric	11.48	14.16	8.18	19.40
chrf-2-symmetric	11.48	14.16	8.18	19.40
chrf-3-symmetric	11.48	14.16	8.18	19.40
chrf-0.5	10.63	12.99	8.08	18.02
meteor	12.23	15.29	8.26	20.38
meteor-symmetric	11.47	14.12	8.20	19.40

Table 5: Lengths of hypotheses as mean number of tokens.