Linear-time Minimum Bayes Risk Decoding with Reference Aggregation

Jannis Vamvas and **Rico Sennrich**
Department of Computational Linguistics, University of Zurich
{vamvas,sennrich}@cl.uzh.ch

Abstract

Minimum Bayes Risk (MBR) decoding is a text generation technique that has been shown to improve the quality of machine translations, but is expensive, even if a sampling-based approximation is used. Besides requiring a large number of sampled sequences, it requires the pairwise calculation of a utility metric, which has quadratic complexity. In this paper, we propose to approximate pairwise metric scores with scores calculated against aggregated reference representations. This changes the complexity of utility estimation from $O(n^2)$ to $O(n)$, while empirically preserving most of the quality gains of MBR decoding. We release our source code.[1]

1 Introduction

The idea of generating translations by maximizing a metric of translation quality (Kumar and Byrne, 2004) has recently been revived in the context of neural machine translation. In sampling-based MBR decoding (Eikema and Aziz, 2020), many hypotheses are sampled from the model distribution, and their expected utility is estimated using Monte Carlo (MC) sampling. This approach has been shown to improve translation quality compared to beam search, especially when neural metrics are used for utility estimation (Freitag et al., 2022).

Estimating utility through MC sampling has quadratic complexity in the number of samples, which limits practical application. Previous work suggested pruning the number of samples based on a cheaper metric or a smaller number of references (Eikema and Aziz, 2022; Cheng and Vlachos, 2023). In this paper, we propose *reference aggregation*, an alternative efficiency technique that exploits the fact that most common metrics represent text sequences in averageable form, e.g., as n-gram statistics or as embeddings. Specifically, we combine representations of the references into an aggregate reference representation, which we then use for utility estimation. Our proposed approximation still relies on MC sampling, but on a lower level: Rather than computing an MC estimate of the expected utility, we compute an MC estimate of the "true" reference representation in the feature space of the given utility metric. Since this estimate only needs to be computed once, our approach has linear complexity in the number of sampled hypotheses and references.

We report empirical results for four translation directions and two utility metrics: ChrF (Popović, 2015), which is based on character n-gram overlap, and COMET (Rei et al., 2020), a neural network trained with examples of human translation quality judgments. For ChrF, we find that reference aggregation reduces the time needed for computing the utility of 1024 samples by 99.5%, without affecting translation quality. For COMET, metric accuracy does decrease with aggregation, but to a lesser extent than with simply reducing the number of references. Depending on the COMET model, computation time is reduced by 95–99%, which makes reference aggregation an efficient method for hypothesis pruning with COMET.

2 Background and Related Work

Sampling-based MBR (Eikema and Aziz, 2020) selects a translation hyp^* out of a set of translation hypotheses $hyp_1, \ldots, hyp_n \in hyps$ by maximizing (expected) utility:

$$hyp^* = \arg\max_{hyp \in hyps} \text{utility}(hyp). \quad (1)$$

The set of hypotheses is sampled from the model distribution $p(hyp|src)$. Eikema and Aziz (2020) propose to approximate the utility using MC sampling: sample a set of pseudo-references $refs = \{ref_1, \ldots, ref_m\} \sim p(ref|src)$ from the model and

[1] https://github.com/ZurichNLP/mbr

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 790–801

calculate a metric against each sampled reference:

$$\text{utility}(hyp) \approx \frac{1}{m} \sum_{ref \in refs} \text{metric}(hyp, ref). \quad (2)$$

For machine translation, typical such metrics are CHRF (Popović, 2015) and BLEU (Papineni et al., 2002), which are based on n-gram statistics, or neural metrics such as COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020).

A line of research has focused on improving the efficiency of sampling-based MBR. Eikema and Aziz (2022) propose *coarse-to-fine MBR*, which prunes the hypotheses based on a cheaper metric, and *N-by-S MBR*, which uses fewer references than hypotheses. Cheng and Vlachos (2023) propose *confidence-based pruning*, where the number of hypotheses is iteratively reduced based on an increasing number of references. Jinnai and Ariu (2024) interpret sampling-based MBR as an instance of *medoid identification* and apply an established approximation algorithm to this problem. A line of work uses MBR outputs as a training reward, avoiding the inefficiency of MBR during deployment (Finkelstein et al., 2023; Yang et al., 2023). Finally, alternative reranking approaches that do not require pairwise comparisons have been proposed (Fernandes et al., 2022).

Several other works investigate the aggregation of reference representations to develop a faster variant of MBR decoding. DeNero et al. (2009) perform reference aggregation in the context of statistical machine translation (SMT). Since SMT does not afford random sampling of pseudo-references, they aggregate references from translation forests or k-best lists. Our study shows the effectiveness of reference aggregation from sampled pseudo-references, and for neural metrics such as COMET. Furthermore, concurrent to our work, Deguchi et al. (2024) propose to aggregate the sentence embeddings of COMET, and use k-means to group the references into multiple clusters.

3 Reference Aggregation

Our approach is based on the observation that most metrics that are commonly used for MBR make use of feature representations that can be aggregated. For example, the n-gram statistics used by CHRF can be aggregated by averaging the counts of the n-grams across all references; and the sentence embeddings used by COMET can be aggregated by calculating an average sentence embedding.

For simplicity, we re-use the above notation, where *hyp* is a hypothesis and *ref* is a reference, but we now assume that they are represented in an averageable form. We then combine the set of references *refs* into an aggregate representation $\overline{ref}$:

$$\overline{ref} = \frac{1}{m} \sum_{ref \in refs} ref. \quad (3)$$

We approximate the expected the utility of a sampled hypothesis by calculating a single metric score against this aggregate representation:

$$\text{utility}(hyp) \approx \text{metric}(hyp, \overline{ref}). \quad (4)$$

Like with standard sampling-based MBR, it is possible to interpret this approximation as MC sampling: By averaging over representations of sampled references, we estimate a representation of the "true" reference, which we then use for approximating the expected utility of each sampled hypothesis. Importantly, the computational complexity of our approach is in $O(|hyps| + |refs|)$ rather than $O(|hyps| \cdot |refs|)$; see Appendix D for a discussion.

3.1 Application to chrF Metric

CHRF (Popović, 2015) is defined as an F-score over character n-grams:

$$\text{CHRF}_\beta = \frac{(1+\beta^2) \cdot \text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}}, \quad (5)$$

where

$$\text{CHRP} = \frac{|hyp \cap ref|}{|hyp|} \text{ and } \text{CHRR} = \frac{|hyp \cap ref|}{|ref|},$$

and the parameter β controls the relative importance of precision and recall. The representations *hyp* and *ref* are bags of n-grams, i.e., objects that map each n-gram to its count in the string.

We apply reference aggregation to CHRF by averaging the counts of n-grams across all references:

$$\overline{ref} = \frac{1}{m} \biguplus_{ref \in refs} ref, \quad (6)$$

where $\biguplus$ is an operation that sums up the counts of each n-gram. We then approximate the expected utility of a hypothesis by calculating $\text{CHRF}_\beta(hyp, \overline{ref})$. Appendix A provides a more formal definition of reference aggregation for CHRF.

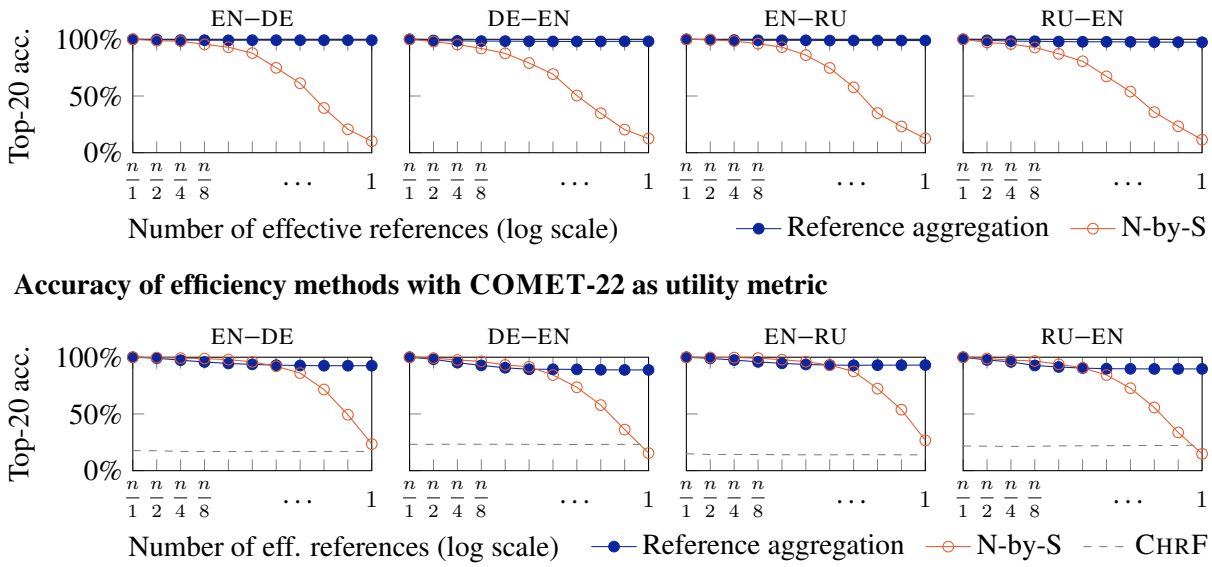


Figure 1: How accurately do MBR efficiency methods approximate standard MBR? In this validation experiment on *newstest21*, we gradually increase efficiency by using fewer references for pairwise utility estimation – either by subsampling the references (N-by-S; Eikema and Aziz, 2022) or by aggregating their representations using partial aggregation (Section 3.3). We report top-20 accuracy, which describes how often an efficiency method ranks the correct hypothesis (as selected by standard MBR) among the top 20 hypotheses. An efficiency method with a high top-20 accuracy could be used for pruning the number of hypotheses to 20 before standard MBR is applied.

3.2 Application to COMET Metric

COMET (Rei et al., 2020) is a pre-trained Transformer model (Vaswani et al., 2017) that has been fine-tuned to predict human judgments of translation quality. In this paper, we focus on the Estimator model architecture, which directly estimates a quality score given a hypothesis, a reference and the source sequence. COMET separately encodes these three inputs into fixed-size embeddings:

$$\boldsymbol{hyp}, \boldsymbol{ref}, \boldsymbol{src} = \text{emb}(hyp), \text{emb}(ref), \text{emb}(src).$$

The three embeddings are then fed into a feed-forward module, which outputs a scalar score:

$$\text{comet}(hyp) = \text{score}(\boldsymbol{hyp}, \boldsymbol{ref}, \boldsymbol{src}). \quad (7)$$

We apply reference aggregation to COMET by averaging the reference embeddings:

$$\overline{\boldsymbol{ref}} = \frac{1}{m} \sum_{ref \in refs} \text{emb}(ref), \quad (8)$$

calculating a single score per hypothesis:

$$\text{comet}(hyp) \approx \text{score}(\boldsymbol{hyp}, \overline{\boldsymbol{ref}}, \boldsymbol{src}). \quad (9)$$

3.3 Partial Aggregation

To better understand the loss of accuracy incurred by aggregation, we experiment with partial aggregation, where we vary the number of references that are combined into an average. Given m references and a desired number of references s that should effectively be used for pairwise utility estimation, we partition the set of references into s subsets and create an aggregate reference for each subset. Appendix B presents a formal description of partial aggregation.

3.4 Aggregate-to-fine MBR

Analogously to *coarse-to-fine MBR* (Eikema and Aziz, 2022), we evaluate an *aggregate-to-fine MBR* approach. Specifically, we use the aggregate reference to prune the number of hypotheses to 20 in a first step. In a second step, we use standard MBR to select the best hypothesis from the pruned set. A formal description is provided in Appendix C.

4 Experimental Setup

Data We use *newstest21* (Akhbardeh et al., 2021) as validation data and *newstest22* (Kocmi et al., 2022) as test data.

Generation Parameters As baselines, we evaluate beam search with a beam size of 5 and epsilon sampling (Hewitt et al., 2022) with $\epsilon = 0.02$. For MBR, we generate 1024 samples per segment using epsilon sampling and re-use the same samples as references. While this approach does not guarantee

	EN–DE	DE–EN	EN–RU	RU–EN	Avg.	Time (utility / total)
Beam search (size 5)	76.16	72.56	68.50	75.47	73.17	- / 0.2 s
Epsilon sampling ($\epsilon = 0.02$)	73.39	69.70	65.79	72.13	70.25	- / 0.2 s
MBR with CHRF metric						
– standard MBR	76.03	72.73	69.52	75.51	73.44	15.0 s / 19.8 s
– reference aggregation	75.95	**72.79**	69.46	75.45	73.41	0.1 s / 4.9 s
– aggregate-to-fine MBR	76.02	**72.80**	69.54	75.47	73.46	0.4 s / 5.2 s
MBR with COMET-22 metric						
– standard MBR	77.64	73.57	72.40	76.11	74.93	23.1 s / 27.9 s
– reference aggregation	77.21	73.36	72.05	76.05	74.67	1.1 s / 5.9 s
– aggregate-to-fine MBR	77.54	73.52	72.29	76.13	74.87	1.5 s / 6.3 s

Table 1: Test results on *newstest22*, using BLEURT-20 for automatic evaluation. We use 1024 samples/references for MBR. In the last column, we report the average time needed for translating a segment, measuring (a) the time needed for utility estimation only, and (b) the total, end-to-end time needed for translation. Underline: no significant BLEURT difference to standard MBR; **bold**: significantly better than standard MBR (bootstrap test, $p < 0.05$).

that the estimation of the expected utility is unbiased (Eikema and Aziz, 2022), it has empirically been found to work well (Freitag et al., 2023).

Models We use open-source NMT models trained for the EN–DE, DE–EN, EN–RU and RU–EN translation directions (Ng et al., 2019).[2] The authors provide an ensemble of four models per direction, but we restrict our experiments to one single model per direction. We use the *Fairseq* codebase (Ott et al., 2019) for model inference.

Metrics For estimating the utilities with CHRF, we use a custom implementation of CHRF[3] that is equivalent to SacreBLEU (Post, 2018) with default settings[4]. As COMET model, we use COMET-22 (Rei et al., 2022a); because this model was not trained on annotations of *newstest21* or *newstest22*, a train–test overlap can be ruled out. We estimate wall-clock time based on a part of the segments, using a system equipped with an NVIDIA GeForce RTX 3090 and an AMD EPYC 7742 64-core processor.

[2]The models were trained with a label smoothing of $\epsilon = 0.1$ (Szegedy et al., 2016), which is a common choice in NMT. Some previous studies of MBR trained custom models without label smoothing (e.g., Eikema and Aziz, 2020). We argue that this is only necessary if unbiased utility estimates are sought through ancestral sampling, and should be less of a concern with epsilon sampling.

[3]https://github.com/jvamvas/fastChrF

[4]chrF2|#:1|case:mixed|eff:yes|nc:6|nw:0|space:no|v:2.0.0

5 Results

5.1 Validation results

Figure 1 evaluates how accurately MBR efficiency methods approximate standard MBR. We report top-20 accuracy, motivated by the idea of coarse-to-fine MBR: any method with perfect top-20 accuracy could be used for pruning the hypothesis set to 20 without affecting quality. Results for top-1 accuracy are reported in Appendix I.[5]

For CHRF, we observe that reference aggregation is Pareto superior to N-by-S, maintaining near-perfect top-20 accuracy even if a single aggregate reference is used. For COMET, reference aggregation causes some loss of accuracy, but outperforms N-by-S if the number of effective references is ≤ 16, where efficiency is highest. In addition, we find that reference aggregation approximates standard (pairwise) COMET much better than using CHRF as a coarse metric does, providing a clear motivation for aggregate-to-fine MBR as an alternative to coarse-to-fine MBR.

5.2 Test results

In Table 1, we report test results for *newstest22*, focusing on a comparison between fast baseline algorithms (beam search and sampling) and MBR (with or without reference aggregation). We perform an automatic evaluation using BLEURT-20 (Sellam et al., 2020), chosen because it is unrelated to the utility metrics we use for MBR. CHRF and

[5]Accuracy was proposed by Cheng and Vlachos (2023) as an evaluation metric for MBR efficiency methods.

COMET scores are reported in Appendix F.

The results show that reference aggregation narrows the efficiency gap between MBR and beam search while preserving most of the quality gain of standard MBR. Reference aggregation speeds up utility estimation by 99.5% for CHRF and 95.1% for COMET-22, reducing the total time needed for translation by 75.5% and 78.8%, respectively. Using an aggregate-to-fine approach has a lower loss of quality and still reduces the total translation time by 73.6–77.4%.

Reference aggregation is thus a successful strategy to overcome the quadratic complexity of MBR. However, it is still slower than beam search, as the cost of sampling is now the dominant factor. Future work could focus on sampling efficiency, e.g., by using fewer hypotheses, improved caching, or speculative sampling approaches (Leviathan et al., 2023; Chen et al., 2023).

6 Conclusion

We proposed reference aggregation, a technique that boosts the efficiency of MBR decoding by shifting the MC sampling from the utility estimation to the reference representation. Experiments on machine translation showed that reference aggregation speeds up utility estimation by up to 99.5% while minimally affecting translation quality. This reduces the gap to beam search and makes MBR more practical for large-scale applications.

Limitations

This work has two main limitations:

1. Reference aggregation requires a utility metric based on averageable representations.

2. For trained metrics, the effectiveness of aggregation needs to be evaluated empirically.

We have demonstrated that reference aggregation is a viable technique for MBR with CHRF and COMET, leading to a considerable speed-up with minor quality losses. In the case of CHRF, reference aggregation entails a slight modification of the metric definition, but is otherwise exact and not an approximation. We thus expect that reference aggregation could be applied in a straightforward manner to other lexical overlap metrics such as CHRF++ (Popović, 2017) and BLEU (Papineni et al., 2002).

For COMET, which is a trained metric, reference aggregation involves the averaging of fixed-size sentence embeddings. We empirically studied the loss of accuracy incurred by this averaging and found that there is a favorable trade-off between speed and accuracy for the COMET models we evaluated. We recommend that future work validates the effectiveness of reference aggregation for other trained metrics.

While CHRF and COMET are among the most commonly used metrics for MBR, previous work has also proposed metrics that are not based on averageable reference representations. For example, BLEURT (Sellam et al., 2020), a trained metric that was shown to be effective for MBR (Freitag et al., 2022), is based on a cross-encoder architecture that creates a joint representation for each hypothesis–reference pair. Future work could investigate in what form, if at all, reference aggregation can be applied to cross-encoder architectures.

Finally, this work studies MBR decoding with a classical sequence-to-sequence NMT model and in the context of sentence-level MT. While MBR decoding has also been successfully applied to MT with large language models (Farinhas et al., 2023), more research is needed on MBR decoding with large language models, especially on the document level.

Acknowledgments

We thank Clara Meister and Bryan Eikema for helpful discussions and feedback. This work was funded by the Swiss National Science Foundation (project MUTAMUR; no. 213976).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

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

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling.

Julius Cheng and Andreas Vlachos. 2023. Faster minimum Bayes risk decoding with confidence-based pruning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12473–12480, Singapore. Association for Computational Linguistics.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. 2024. Centroid-based efficient minimum bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. 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.

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

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

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

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

Mara Finkelstein, Subhajit Naskar, Mehdi Mirzazadeh, Apurva Shah, and Markus Freitag. 2023. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuu Jinnai and Kaito Ariu. 2024. Hyperparameter-free approach for faster minimum bayes risk decoding.

Donald E Knuth. 1997. *Art of computer programming, Volume 2: Seminumerical algorithms*, 3rd edition. Addison-Wesley.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. 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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

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

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

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2023. Direct preference optimization for neural machine translation with minimum bayes risk decoding.

A Formal Definition of Reference Aggregation for ChrF

The CHRF metric (Popović, 2015) is a harmonic mean of precision and recall scores:

$$\text{CHRF}_\beta = \frac{(1+\beta^2)\cdot \text{CHRP}\cdot \text{CHRR}}{\beta^2\cdot \text{CHRP} + \text{CHRR}}. \quad (10)$$

Internally, CHRF converts hypotheses and references into bags of character n-grams. Such bags can be represented as multisets (Knuth, 1997, Section 4.6.3) or as (sparse) vectors. We will use vector notation in this formal definition, which allows us to define reference aggregation with standard vector operations.

Let $\boldsymbol{hyp} \in \mathbb{R}^{|\mathcal{V}|}$ and $\boldsymbol{ref} \in \mathbb{R}^{|\mathcal{V}|}$ be bags representing a hypothesis and a reference, where $\mathcal{V}$ is the vocabulary of all character n-grams up to maximum order n, and the entries $\boldsymbol{hyp}_j$ and $\boldsymbol{ref}_j$ are the counts of n-gram $j \in \mathcal{V}$ in the hypothesis and reference, respectively.

For a given n-gram order $i \in \{1,\ldots,n\}$, precision and recall are defined as:

$$\text{CHRP}_i(\boldsymbol{hyp}, \boldsymbol{ref}) = \frac{\sum_{j\in\mathcal{V}_i} \min(\boldsymbol{hyp}_j, \boldsymbol{ref}_j)}{\sum_{j\in\mathcal{V}_i} \boldsymbol{hyp}_j}, \quad (11)$$

$$\text{CHRR}_i(\boldsymbol{hyp}, \boldsymbol{ref}) = \frac{\sum_{j\in\mathcal{V}_i} \min(\boldsymbol{hyp}_j, \boldsymbol{ref}_j)}{\sum_{j\in\mathcal{V}_i} \boldsymbol{ref}_j}, \quad (12)$$

where $\mathcal{V}_i$ is the set of all character n-grams of order i. Overall precision and recall are calculated as the arithmetic mean of the precision and recall scores for each n-gram order:

$$\text{CHRP}(\boldsymbol{hyp}, \boldsymbol{ref}) = \frac{1}{n}\sum_{i=1}^{n} \text{CHRP}_i(\boldsymbol{hyp}, \boldsymbol{ref}), \quad (13)$$

$$\text{CHRR}(\boldsymbol{hyp}, \boldsymbol{ref}) = \frac{1}{n}\sum_{i=1}^{n} \text{CHRR}_i(\boldsymbol{hyp}, \boldsymbol{ref}). \quad (14)$$

When CHRF is used as a utility metric in a standard MBR setting, the expected utility of a hypothesis is estimated based on a set $\{\boldsymbol{ref}^{(1)},\ldots,\boldsymbol{ref}^{(m)}\}$ of m references:

$$\text{utility}(\boldsymbol{hyp}) = \frac{1}{m}\sum_{k=1}^{m} \text{CHRF}_\beta(\boldsymbol{hyp}, \boldsymbol{ref}^{(k)}). \quad (15)$$

In contrast, reference aggregation first calculates the arithmetic mean of the reference bags:

$$\overline{\boldsymbol{ref}} = [\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{ref}_1^{(k)}, \ldots, \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{ref}_{|\mathcal{V}|}^{(k)}], \quad (16)$$

and estimates the utility as:

$$\text{utility}_{\text{agg}}(\boldsymbol{hyp}) = \text{CHRF}_\beta(\boldsymbol{hyp}, \overline{\boldsymbol{ref}}). \quad (17)$$

Note that the only mathematical difference between pairwise calculation of chrF and using the aggregate reference is that the F-score is averaged across sentences in the pairwise calculation, and computed over the global precision and recall with reference aggregation.

B Formal Definition of Partial Aggregation

We conceptualize partial aggregation as follows:

1. The set of individual references contains m references.
2. We randomly partition the set of references into s groups of equal size.
3. Each group is combined into an average reference representation, resulting in s aggregate references $\overline{ref}^{(i)},\ldots,\overline{ref}^{(s)}$.

The expected utility of each sampled hypothesis is then approximated as the average metric score over all aggregate references:

$$\text{utility}(hyp) \approx \frac{1}{s}\sum_{i=1}^{s} \text{metric}(hyp, \overline{ref}^{(i)}). \quad (18)$$

Like with N-by-S MBR, the parameter s can be seen as the *number of effective references* that determines the computational complexity of the utility estimation. The case $s = m$ corresponds to standard MBR, where each sampled hypothesis is compared to each reference in a pairwise fashion. The case $s = 1$ corresponds to the full aggregation approach, where a single aggregate reference is created from all references.

C Formal Definition of Aggregate-to-fine MBR

Aggregate-to-fine MBR is a special case of coarse-to-fine MBR (Eikema and Aziz, 2022), which uses a cheap proxy utility function to prune the number of hypotheses. In the case of aggregate-to-fine MBR, the proxy utility function is based on an aggregate reference representation.

The general definition of coarse-to-fine MBR is as follows: Given the original set of sampled hypotheses $\bar{\mathcal{H}}(x)$ and a proxy utility function u_{proxy}, coarse-to-fine MBR selects a subset of T hypotheses:

$$\bar{\mathcal{H}}_T(x) \coloneqq \underset{hyp \in \bar{\mathcal{H}}(x)}{\text{top-T}} \; u_{\text{proxy}}(hyp). \tag{19}$$

In the second step, the utility of each hypothesis in the pruned set is estimated using the fine-grained utility function u_{target}:

$$y^{\text{C2F}} \coloneqq \underset{hyp \in \bar{\mathcal{H}}_T(x)}{\arg\max} \; u_{\text{target}}(hyp). \tag{20}$$

When experimenting with aggregate-to-fine MBR, we re-use the same utility metric for both steps, but first with an aggregate reference and then with the full set of references:

$$u_{\text{proxy}}(hyp) = \text{metric}(hyp, \overline{ref}), \tag{21}$$

$$u_{\text{target}}(hyp) = \frac{1}{m} \sum_{ref \in refs} \text{metric}(hyp, ref). \tag{22}$$

Note that using the same metric in both steps is not strictly necessary, but has the advantage that the features (e.g., embeddings) only need to be computed once.

D Complexity Analysis

Generally, reference aggregation reduces the complexity of utility estimation from $O(nm)$ to $O(n + m)$, where n is the number of hypotheses and m is the number of references. The exact complexity depends on the specifics of the utility metric. Here, we provide a more detailed analysis for ChrF and COMET.

Above, we stated that utility estimation with these metrics usually has two stages: feature extraction and scoring. The feature extraction stage is not affected by reference aggregation, and previous work has already remarked that reference features can be extracted once and re-used for all hypotheses (Amrhein and Sennrich, 2022). If the reference set is identical to the set of hypotheses, the feature extraction stage is in $O(n)$, otherwise $O(n + m)$.

The scoring stage of ChrF is dominated by the element-wise minimum function in Eqs. 11 and 12 (or, if the bags of n-grams are represented as multisets, by the intersection operation $hyp \cap ref$). Because this operation is performed separately for each hypothesis–reference pair, the complexity is in $O(nm)$. Reference aggregation reduces the complexity to $O(n + m)$, given that the aggregate reference can be computed once and then re-used for all hypotheses.[6]

The same analysis applies to COMET. With standard MBR, Eq. 7 is evaluated for each hypothesis–reference pair; with reference aggregation, it is only evaluated once for each hypothesis. The aggregate reference embeddings can be computed once and re-used for all hypotheses.

In practice, the runtime of utility estimation is affected by additional factors. There may be duplicates among the samples, so the number of scores that effectively need to be computed can vary. In addition, most aspects of utility estimation can be computed in parallel, which makes the effective runtime highly implementation-dependent.

[6] For ChrF, reference aggregation can result in an aggregate bag of n-grams that is larger that the bags of the individual references; in the theoretical worst case, where all the references are disjoint, even in an aggregate bag that is m times larger. However, this is a highly unlikely scenario in practice, since different translations of the same source will have substantial overlap, and even if $|\overline{ref}| \gg |ref|$, the cost of intersection only depends on $|hyp|$, assuming that a constant-time hash table is used to check whether each item in hyp is contained in $\overline{ref}$.

E Data Statistics

	# Segments	# Samples per segment	# Unique samples per segment
newstest21 EN–DE	1002	1024	874.2
newstest21 DE–EN	1000	1024	716.9
newstest21 EN–RU	1002	1024	896.7
newstest21 RU–EN	1000	1024	727.3
newstest22 EN–DE	2037	1024	697.5
newstest22 DE–EN	1984	1024	671.4
newstest22 EN–RU	2037	1024	750.2
newstest22 RU–EN	2016	1024	726.3

Table 2: Statistics for the datasets used in this paper. We sample 1024 hypotheses per source segment using epsilon sampling and find that most of the samples are unique.

F Extended Test Results

	CHRF	Cometinho	COMET-22	XCOMET-XL	BLEURT-20
Beam search (size 5)	58.6	56.0	84.3	92.2	73.2
Epsilon sampling ($\epsilon = 0.02$)	52.6	45.3	81.9	89.4	70.3
MBR with CHRF metric					
– standard MBR	59.8	58.3	84.5	91.8	73.4
– reference aggregation	59.8	58.2	84.5	91.7	73.4
– aggregate-to-fine MBR	59.8	58.3	84.5	91.8	73.5
MBR with Cometinho metric					
– standard MBR	57.5	65.1	85.1	92.5	74.0
– reference aggregation	**57.8**	64.5	85.0	92.4	73.9
– aggregate-to-fine MBR	57.5	65.0	85.1	92.5	74.0
MBR with COMET-22 metric					
– standard MBR	57.3	60.8	87.1	93.7	74.9
– reference aggregation	**57.7**	60.8	86.8	93.4	74.7
– aggregate-to-fine MBR	**57.4**	60.8	87.0	93.7	74.9
Coarse-to-fine MBR					
– standard CHRF to COMET-22	**59.3**	60.1	85.8	93.0	74.4
– aggregate CHRF to COMET-22	**59.4**	60.2	85.8	93.0	74.4

Table 3: Extended results on *newstest22* with 1024 samples/references for MBR. In this table, we include Cometinho (Rei et al., 2022b) as utility metric, which is a distilled COMET model. Furthermore, as an additional evaluation metric, we report XCOMET-XL (Guerreiro et al., 2023). We average the evaluation scores across the four translation directions. Underline: no significant difference to standard MBR; **bold**: significantly better than standard MBR (bootstrap test, $p < 0.05$).

G Test Results with 256 Samples

	EN–DE	DE–EN	EN–RU	RU–EN	Avg.	Time (utility / total)
Beam search (size 5)	76.16	72.56	68.50	75.47	73.17	- / 0.2 s
Epsilon sampling ($\epsilon = 0.02$)	73.39	69.70	65.79	72.13	70.25	- / 0.2 s
MBR with CHRF metric						
– standard MBR	75.90	72.66	69.27	75.60	73.36	0.8 s / 2.1 s
– reference aggregation	75.83	72.69	69.19	75.53	73.31	$<$ 0.1 s / 1.3 s
– aggregate-to-fine MBR	75.90	72.67	69.29	75.58	73.36	0.1 s / 1.4 s
MBR with COMET-22 metric						
– standard MBR	77.44	73.38	72.15	76.07	74.76	1.6 s / 2.9 s
– reference aggregation	77.18	73.24	71.85	75.98	74.56	0.3 s / 1.6 s
– aggregate-to-fine MBR	77.42	73.36	71.98	76.05	74.70	0.4 s / 1.7 s

Table 4: Version of Table 1 that uses 256 samples/references for MBR.

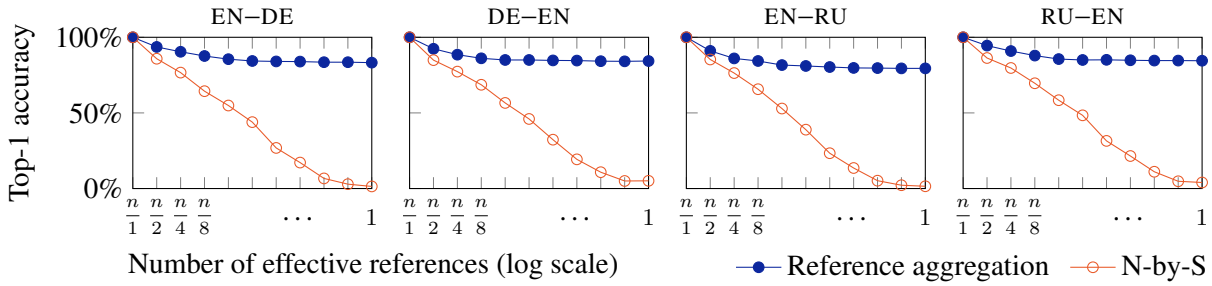
H Effect of Larger Beam Size

Beam size	EN–DE	DE–EN	EN–RU	RU–EN	Avg.
5	76.16	72.56	68.50	75.47	73.17
10	76.20	72.57	67.92	75.51	73.05
15	76.19	72.53	68.10	75.48	73.08
20	76.18	72.54	67.84	75.49	73.01
25	76.19	72.50	67.82	75.46	72.99

Table 5: Increasing the beam size to values larger than 5 does not tend to improve translation quality of beam search on *newstest22* in terms of BLEURT-20.

I Top-1 Accuracy of Efficiency Methods

Utility metric: CHRF



Utility metric: COMET-22

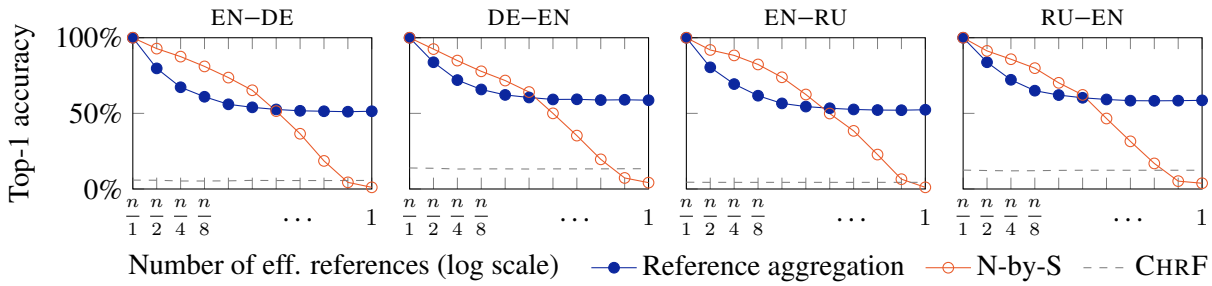
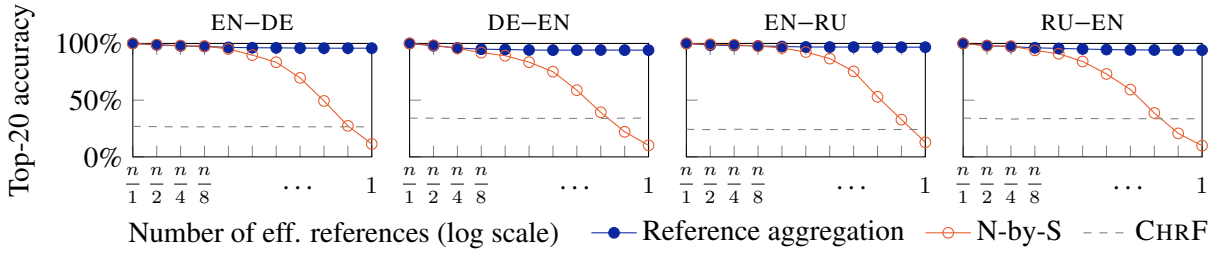


Figure 2: Top-1 accuracy of MBR efficiency methods on *newstest21*, analogous to Figure 1.

J Validation Results for Cometinho

Top-20 accuracy



Top-1 accuracy

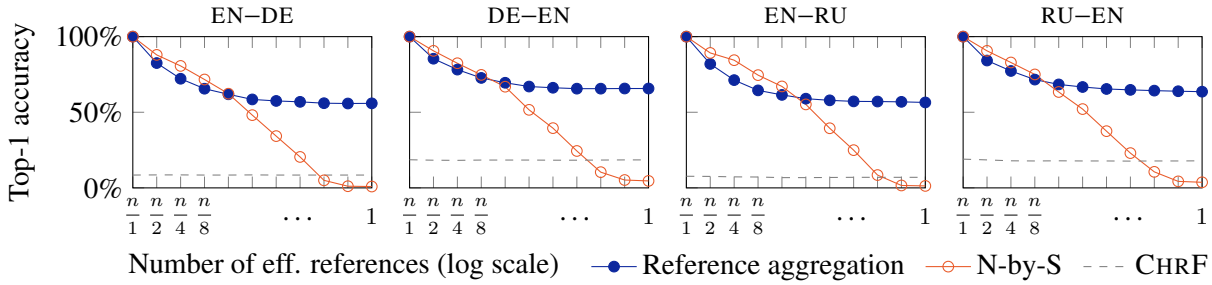


Figure 3: Accuracy of MBR efficiency methods on *newstest21* when using the Cometinho model (Rei et al., 2022b) as utility metric.