On the True Distribution Approximation of Minimum Bayes-Risk Decoding

Atsumoto Ohashi[1]* **Ukyo Honda**[2] **Tetsuro Morimura**[2] **Yuu Jinnai**[2]
[1]Nagoya University [2]CyberAgent
ohashi.atsumoto.c0@s.mail.nagoya-u.ac.jp
{honda_ukyo,morimura_tetsuro,jinnai_yu}@cyberagent.co.jp

Abstract

Minimum Bayes-risk (MBR) decoding has recently gained renewed attention in text generation. MBR decoding considers texts sampled from a model as pseudo-references and selects the text with the highest similarity to the others. Therefore, sampling is one of the key elements of MBR decoding, and previous studies reported that the performance varies by sampling methods. From a theoretical standpoint, this performance variation is likely tied to how closely the samples approximate the true distribution of references. However, this approximation has not been the subject of in-depth study. In this study, we propose using anomaly detection to measure the degree of approximation. We first closely examine the performance variation and then show that previous hypotheses about samples do not correlate well with the variation, but our introduced anomaly scores do. The results are the first to empirically support the link between the performance and the core assumption of MBR decoding.[1]

1 Introduction

Minimum Bayes-risk (MBR) decoding has recently re-emerged as a better alternative to beam search in text generation such as neural machine translation (NMT), text summarization, and image captioning (Eikema and Aziz, 2020; Freitag et al., 2022; Fernandes et al., 2022; Suzgun et al., 2023; Bertsch et al., 2023). MBR decoding first samples texts from a model and then selects the text most similar to the others, considering the text samples as substitutes for references. Therefore, sampling plays an important role in MBR decoding, and previous studies have reported that the performance varies by sampling methods (Eikema and Aziz, 2020, 2022; Fernandes et al., 2022; Freitag et al., 2023).

From a theoretical standpoint, the samples are assumed to approximate the *true distribution*, the distribution of human-quality translations (Kumar and Byrne, 2002, 2004). If the approximation deviates, biases can emerge in results of MBR decoding. This implies a significant link between the performance variation and approximation quality. Although previous studies explained the performance variation by some properties of samples, *e.g.*, sampling bias and cumulative probability mass (Eikema and Aziz, 2020; Freitag et al., 2023), those properties have no clear relation with the true distribution. Consequently, the relation between the performance gains by sampling methods and the core assumption remains unclear.

This study aims to empirically support the link between the performance and the approximation of true distribution. To this end, we introduce measures for the degree of approximation. If the assumption for samples holds, references, which are drawn from the true distribution by definition, should not deviate from the majority of the samples. Based on this recasting, we employ *anomaly detection* (also called *outlier or novelty detection*) for the measure. Our hypothesis is that references achieve lower anomaly scores among samples obtained with a higher-performance sampling method. We first closely examine the performance variation by sampling methods. Then, we show that the variation highly correlates with the anomaly scores but not so with the properties based on previous hypotheses. The results are the first to provide empirical evidence for the link between the performance and core assumption, which is an important step to understanding the connection between the actual performance and the theory of MBR decoding.

2 Preliminaries

Let $u(y, r)$ be a utility function to measure the quality of model translation y (**candidate**; Freitag et al.,

*Work done during an internship at CyberAgent.

[1]The code is available at https://github.com/CyberAgentAILab/mbr-anomaly.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 459–468
June 16-21, 2024 ©2024 Association for Computational Linguistics

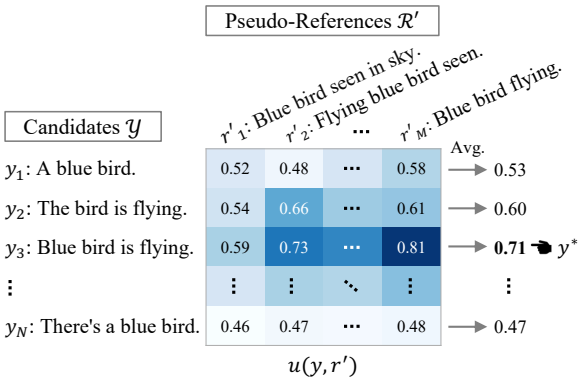

Figure 1: Illustrative example of MBR decoding.

2022) given its reference translation r. Among a set of candidates $\mathcal{Y}$, MBR decoding selects the one that minimizes the expected error or, equivalently, maximizes the expected utility (Kumar and Byrne, 2002, 2004; Freitag et al., 2022):

$$y^* = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{r \sim P_{\text{human}}(\cdot|x)}[u(y, r)]. \quad (1)$$

Here, $P_{\text{human}}(\cdot|x)$ is the **true distribution** over translations of an input text x (Kumar and Byrne, 2002, 2004), which describes human-quality translations in the space of all translations.

Since the true distribution is unknown, MBR decoding approximates Eq. (1) with finite samples drawn from a model $r' \sim P_{\text{model}}(\cdot|x)$. That is, MBR decoding *assumes that the samples drawn from a model approximate the true distribution* of references. The samples are called **pseudo-references** (Freitag et al., 2022), which subsequently serve as alternatives to references in the computation of MBR as follows:

$$y^* = \arg\max_{y \in \mathcal{Y}} \frac{1}{|\mathcal{R}'|} \sum_{r' \in \mathcal{R}'} u(y, r'). \quad (2)$$

In practice, candidates $\mathcal{Y}$ and pseudo-references $\mathcal{R}'$ can be the same or different sets of samples. Figure 1 shows an example of the above procedure.

3 Performance Variation by Sampling

Previous studies reported that performance varies by sampling methods in NMT. However, they used the same set of model translations for both candidates and pseudo-references (Eikema and Aziz, 2020; Fernandes et al., 2022; Freitag et al., 2023) or explored sampling methods only for candidates (Eikema and Aziz, 2022). These settings obscure the effect of pseudo-references, for which the true distribution is assumed, on the performance variation. This section shows the effect of pseudo-reference sampling on performance by evaluating pseudo-references separately from candidates.

3.1 Setup

Following Fernandes et al. (2022), we use publicly-available Transformer models (Vaswani et al., 2017) trained by Ng et al. (2019)[2] for the WMT19 news translation task (Barrault et al., 2019). The models were trained in four directions between English (en) and German (de) or Russian (ru). We conducted our experiments on the test set of WMT19 (*newstest19*), which was used as the development set in the previous work (Fernandes et al., 2022). Due to the quadratic computational cost of MBR decoding, we drew 100 samples of $\mathcal{Y}$ and $\mathcal{R}'$ for each of the 1,000 examples of *newstest19*. We employ COMET22 for the utility function u, which is the state-of-the-art evaluation metric in machine translation (Rei et al., 2022, 2020).[3]

For sampling methods, we use those that have been reported to be effective: ancestral sampling (Eikema and Aziz, 2020; Freitag et al., 2022), beam search, nucleus sampling (Eikema and Aziz, 2022; Fernandes et al., 2022), and epsilon sampling (Freitag et al., 2023). Ancestral sampling draws samples from $P_{\text{model}}(\cdot|x)$ without modification, while nucleus sampling restricts sampling to words with top-p probabilities (Holtzman et al., 2020) and epsilon sampling truncates words with probabilities lower than ϵ (Hewitt et al., 2022). We adopt the best hyperparameters reported for p and ϵ (Fernandes et al., 2022; Freitag et al., 2023). The beam size was set to 100 to collect 100 samples.

3.2 Results

Fixing Candidates. Since we focus on sampling for pseudo-references, we first search for the best sampling method *for candidates* and fix it. The objective of searching for the best is to prevent the pseudo-reference's contribution to scores from being capped and obscured by the candidate's quality. To this end, we conduct the search on the same *newstest19* as the subsequent experiments.[4] Following

[2] https://github.com/facebookresearch/fairseq/blob/7409af7f9a7b6ddac4cbfe7cafccc715b3c1b21e/examples/translation/README.md

[3] COMET22 improved robustness to the deviation in numbers and named entities, which was the weakness of the previous COMET (Amrhein and Sennrich, 2022).

[4] If the objective is to find the best combination of sampling methods, which is not our focus, then it is desirable to use

Candidate	de-en	en-de	ru-en	en-ru
Ancestral	85.82	86.32	82.11	86.13
Beam	88.47	89.32	84.16	89.44
Epsilon ($\epsilon = 0.02$)	**88.51**	**89.47**	**84.36**	**90.17**
Nucleus ($p = 0.6$)	88.01	89.12	83.76	89.96
Nucleus ($p = 0.9$)	88.02	89.04	83.98	89.57

Table 1: *Oracle scores* in COMET22 to determine the sampling method *for candidates*. The results are the average of three runs with different seeds except for beam search. The best/worst scores are in **bold**/underlined.

Fernandes et al. (2022), we search for the sampling method that achieves the highest oracle score, $\max_{y \in \mathcal{Y}} u(y, r)$, on average. Table 1 shows that epsilon sampling achieves the highest oracle score across the language pairs. Based on these results, we fixed the sampling method of candidates to epsilon sampling in all the following experiments.

Varying Pseudo-References. Then, we evaluate the effect of pseudo-references on performance by varying their sampling methods. Table 2 shows the results. As expected from previous studies, the performance of MBR decoding varies even when only changing the sampling methods of pseudo-references. The variation is nearly consistent across language pairs, indicating the pervasive effect of pseudo-reference on performance. The best sampling method for candidates (epsilon sampling) is not the best for pseudo-references. This shows that the desirable properties for candidates and pseudo-references are different.

Table 2 also shows the results of beam search just for the comparison with MBR decoding. Here, the beam size was set to 5. MBR decoding significantly outperforms beam search and even outperforms the ensemble model, which was the winner of WMT19 (Barrault et al., 2019). Since the effectiveness of epsilon sampling was reported on the other WMT dataset (Freitag et al., 2023), we have a good reason to use epsilon sampling for this comparison.

4 Hypotheses for Performance Variation

The previous section confirmed that the performance varies by sampling pseudo-references. The question that naturally arises in response to the results is: why does this variation occur?

different splits to explore and test the combination to ensure the generalization. Nevertheless, our subsequent results in Tables 1, 2, 4, and 5 suggest the generalization of the found best combination as it consistently performs the best across almost all language pairs.

	Pseudo-Reference	de-en	en-de	ru-en	en-ru
Epsilon ($\epsilon = 0.02$)	Ancestral	85.82	87.51	82.02	88.41
	Beam	85.62	87.40	81.64	87.78
	Epsilon ($\epsilon = 0.02$)	85.89	87.74	82.01	88.46
	Epsilon ($\epsilon = 0.02$)*	85.87	87.74	81.98	88.46
	Nucleus ($p = 0.6$)	85.69	87.57	81.76	88.26
	Nucleus ($p = 0.9$)	**86.04**	**87.82**	**82.18**	**88.61**
	Beam Search	84.38	86.13	80.76	85.69
	Beam Search (ensemble)	84.30	86.06	80.91	85.74

Table 2: COMET22 scores of MBR decoding with different pseudo-references. Candidates are sampled with epsilon sampling ($\epsilon = 0.02$). Epsilon ($\epsilon = 0.02$)* shows the results of sampling candidates and pseudo-references with the same epsilon sampling but with different seeds. The results are the average of three runs with different seeds except for beam search. The best/worst scores are in **bold**/underlined.

4.1 Previous Hypotheses

Eikema and Aziz (2022) hypothesized that unbiased sampling is desirable for pseudo-references. Since the biased sampling methods limit the sampling to words of high probability, we use the average log probability (**Avg. Prob.**) of samples as a continuous proxy of bias existence in sampling. Eikema and Aziz (2020) and Freitag et al. (2023) did not distinguish between candidates and pseudo-references but referred to the larger cumulative probability mass (**Cum. Prob.**) of unique samples as a desirable property because it indicates diverse and probable samples. Eikema and Aziz (2022) employed candidate sampling that achieved high expected utility. If this criterion applies to pseudo-references, performance should be higher when the expected utility against candidates (**Cand. Sim.**) or references (**Ref. Sim.**) is high.

4.2 Our Hypothesis

Given the relaxation from Eq. (1) to Eq. (2), a better approximation of the true distribution by pseudo-references should be associated with higher performance. To examine the relation, we propose using anomaly detection to quantitatively evaluate the approximation. If a better approximation is achieved, references should deviate less from the majority of the samples since references are drawn from the true distribution by definition. This recasting allows us to use anomaly scores of anomaly detection for measuring the degree of approximation. We then hypothesize that *a higher-performance sampling method forms samples where references achieve lower anomaly scores*.

5 Experiments

We test the hypotheses discussed in the previous section by evaluating the correlation between the performance variation and the properties or anomaly scores.

5.1 Setup

The setup is the same as described in Section 3.1. We run each sampling method with three different seeds and then calculate the Spearman's rank correlation coefficient ρ between their averaged properties or anomaly scores (see Section 4) with the COMET22 scores reported in Table 2.

5.2 Anomaly Scores

To test our hypothesis, we employ three popular methods used in *unsupervised* anomaly detection (Kriegel et al., 2011; Gu et al., 2019; Ruff et al., 2021). The first uses **Mahalanobis distance** (d_M; Mahalanobis, 1936) as an anomaly score, a classical distance measure between a data point and a distribution. In our context, the distance is between reference r and pseudo-references $\mathcal{R}'$ in a feature space: $d_M(r, \mathcal{R}') = \sqrt{(r-\mu)\Sigma^{-1}(r-\mu)}$, where μ and Σ^{-1} are mean and inverse covariance matrix of $\mathcal{R}'$, respectively. Mahalanobis distance assumes that the data is normally distributed, but this assumption does not necessarily hold. ***k*-nearest neighbors** (***k*NN**; Angiulli and Pizzuti, 2002) does not have the assumption and is applicable to other data, such as the one with multimodal distribution. kNN is a simple algorithm to consider the local density of a given data point. Still, it is known to perform favorably to some state-of-the-art algorithms for anomaly detection (Gu et al., 2019). Let $N_i(r, \mathcal{R}')$ be the ith nearest pseudo-reference to r in Euclidean distance. kNN takes the average of Euclidean distance from r to its k-nearest pseudo-references $\{N_i(r, \mathcal{R}')\}_{i=1}^k$: $k\mathrm{NN}(r, \mathcal{R}') = \frac{1}{k}\sum_{i=1}^k \|r - N_i(r, \mathcal{R}')\|$. **Local outlier factor** (**LOF**; Breunig et al., 2000) additionally considers the local density of the k-nearest data points themselves. LOF in our setting measures how the local density of r deviates from that of its k-nearest pseudo-references. To better illustrate the relationship with kNN, we show a simplified version of LOF (Schubert et al., 2014) here: $\mathrm{LOF}_k(r, \mathcal{R}') = \frac{1}{k}\sum_{r' \in \{N_i(r,\mathcal{R}')\}_{i=1}^k} \frac{\|r - N_k(r,\mathcal{R}')\|}{\|r' - N_k(r',\mathcal{R}')\|}$. See Breunig et al. (2000) for the complete formula we used.

To calculate the anomaly scores, samples needs to be represented in a feature space. We obtain the representation in the space of utility by measuring the utility of references or pseudo-references given a set of candidates $\mathcal{Y}$. A reference r in the space is then defined as $\left[u(r, y_1), \ldots, u(r, y_{|\mathcal{Y}|})\right]^\top$. Same for a pseudo-reference r'.

	de-en	en-de	ru-en	en-ru
Avg. Prob.$_{(-)}$	0.580$_{(\checkmark)}$	0.290$_{(\checkmark)}$	0.870$_{(\checkmark)}$	0.638$_{(\checkmark)}$
Cum. Prob.$_{(+)}$	<u>0.058</u>$_{(\times)}$	<u>0.116</u>$_{(\times)}$	<u>0.348</u>$_{(\times)}$	<u>0.058</u>$_{(\times)}$
Cand. Sim.$_{(+)}$	0.543$_{(\times)}$	0.314$_{(\times)}$	0.829$_{(\times)}$	0.657$_{(\times)}$
Ref. Sim.$_{(+)}$	0.580$_{(\times)}$	0.290$_{(\times)}$	0.870$_{(\times)}$	0.638$_{(\times)}$
$d_{M(-)}$	0.771$_{(\checkmark)}$	0.486$_{(\checkmark)}$	0.886$_{(\checkmark)}$	0.771$_{(\checkmark)}$
$k\mathrm{NN}_{(-)}$				
k = 5	0.771$_{(\checkmark)}$	0.829$_{(\checkmark)}$	0.886$_{(\checkmark)}$	0.829$_{(\checkmark)}$
k = 25	0.943$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	0.886$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$
k = 50	0.771$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	0.829$_{(\checkmark)}$
k = 75	0.771$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	0.371$_{(\checkmark)}$	0.829$_{(\checkmark)}$
k = 100	0.086$_{(\checkmark)}$	0.314$_{(\checkmark)}$	0.371$_{(\checkmark)}$	0.029$_{(\checkmark)}$
$\mathrm{LOF}_{(-)}$				
k = 5	0.829$_{(\checkmark)}$	0.600$_{(\checkmark)}$	0.943$_{(\checkmark)}$	0.771$_{(\checkmark)}$
k = 25	0.829$_{(\checkmark)}$	0.714$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	0.829$_{(\checkmark)}$
k = 50	**1.000**$_{(\checkmark)}$	0.886$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	0.829$_{(\checkmark)}$
k = 75	**1.000**$_{(\checkmark)}$	0.886$_{(\checkmark)}$	**0.943**$_{(\checkmark)}$	0.829$_{(\checkmark)}$
k = 100	0.600$_{(\checkmark)}$	0.371$_{(\checkmark)}$	0.886$_{(\checkmark)}$	0.657$_{(\checkmark)}$

Table 3: Correlation coefficients (Spearman's ρ) between COMET22 performance variation and pseudo-references' properties or anomaly scores. We show the absolute value of ρ. The subscript signs $_{(+/-)}$ are the expected signs of ρ (see Section 4), and the subscript marks $_{(\checkmark/\times)}$ show whether the actual signs match/mismatch the expected ones. The best/worst scores are in **bold**/<u>underlined</u>.

5.3 Results

Table 3 shows the results. As expected, the anomaly scores are clearly more correlated than the properties based on previous hypotheses. Except for Cum. Prob., Cand. Sim., and Ref. Sim., the signs of ρ are all as expected, including the anomaly scores. See Table 7 in Appendix for the results used to calculate ρ.

Among the anomaly scores,[5] kNN and LOF with k = 50 stably correlate with the performance variation better than those with k = 100 and d_M. We speculate that the significant degradation of kNN with k = 100 is caused by outliers in pseudo-references. While kNN with k < 100 can effec-

[5]We took the median of d_M and LOF scores instead of the mean because they are unstable due to the inverse covariant matrix Σ^{-1} and division, respectively. For d_M, we removed duplicated y from the position vector and added an identity matrix not to drop the rank of Σ and stabilize the computation of Σ^{-1}. The value of the elements of the identity matrix was set to 1e-5, taking into account that the average value of the diagonal components of Σ was 1e-3.

tively avoid including these outliers in the calculation of anomaly scores, kNN with $k = 100$ cannot, and its anomaly scores are likely to be distorted by the outliers. These results suggest that even if some pseudo-references are outliers against a reference, the performance tends to be higher if the rest of the pseudo-reference is close to the reference. In other words, pseudo-references do not have to be close to references in entirety to perform well.

6 Related Work

MBR decoding has been used in automatic speech recognition (Goel and Byrne, 2000), statistical machine translation (Kumar and Byrne, 2002, 2004), and NMT (Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017). Recently, MBR decoding has gained prominence again in NMT because of the following two innovations. (1) Eikema and Aziz (2020) showed that MBR decoding with stochastic sampling has a potential to outperform MAP decoding methods, including beam search; (2) Freitag et al. (2022) and Fernandes et al. (2022) explored utility functions and found that using neural reference-based metrics as the utility function significantly enhances the quality of output texts. Müller and Sennrich (2021) reported domain robustness and less hallucination in the outputs of MBR decoding. Other text generation tasks such as text summarization, image captioning, and diversity-aware text generation also benefit from MBR decoding (Suzgun et al., 2023; Borgeaud and Emerson, 2020; Jinnai et al., 2024). Recent studies have focused on improving the efficiency of MBR decoding (Cheng and Vlachos, 2023; Finkelstein and Freitag, 2023; Yang et al., 2023; Jinnai et al., 2023; Jinnai and Ariu, 2024).

The most related studies explored sampling methods for MBR decoding and raised hypotheses to explain the difference in performance by sampling methods (Eikema and Aziz, 2020, 2022; Fernandes et al., 2022; Freitag et al., 2023). We also explored sampling methods but differed in that we did it more closely by focusing on pseudo-references. Furthermore, we introduced anomaly scores that correlate with the performance variation better than previous hypotheses.

7 Conclusion

This study investigated the relation between the performance of MBR decoding and the core assumption about samples: samples follow the true distribution of references. We introduced anomaly scores used in anomaly detection to evaluate the approximation of the true distribution. Experimental results demonstrated that the anomaly scores correlate with the performance significantly better than the properties hypothesized to explain the performance variation in prior literature. The previous hypotheses assumed that unbiased sampling (Avg. Prob.), diverse and probable samples (Cum. Prob.), or high expected utility (Cand. and Ref. Sim.) are the key properties of samples to achieve high performance. However, these properties do not have an obvious relationship to approximating the true distribution of references, in contrast to the anomaly scores we employed.

These results show the insufficiency of existing hypotheses about the properties that samples should possess. The results are also the first to empirically support the link between the actual performance and the key assumption of MBR decoding. We believe this serves as an essential step to understanding the connection between the actual performance and the theory of MBR decoding.

8 Limitations and Risks

The limitation of the study is that it is solely a thorough analysis of MBR decoding, not accompanied by an algorithm to improve the performance of MBR decoding. However, our analysis empirically shows that previous hypotheses about the properties of samples are insufficient and that following the assumption of the MBR decoding is the key to improving performance. We believe this is an important contribution that modifies the direction of future development of MBR decoding.

Our investigation is limited to Transformer models provided by Ng et al. (2019) and the task is limited to machine translation. Future work will extend the analysis to a wider range of models and text generation tasks. However, it is worth noting that some studies support the general applicability of MBR decoding findings obtained in NMT to other text generation tasks and models. Some hyperparameters (Suzgun et al., 2023), efficiency-boosting techniques (Jinnai et al., 2023; Jinnai and Ariu, 2024), or diversity-aware extensions (Jinnai et al., 2024) for MBR decoding consistently perform well across machine translation, summarization, image captioning, and data-to-text generation with different models. Bertsch et al. (2023) shows that MBR decoding works well even in open-ended

text generation tasks.

We do not foresee any ethical concerns in our analysis.

References

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

Fabrizio Angiulli and Clara Pizzuti. 2002. Fast outlier detection in high dimensional spaces. In *Principles of Data Mining and Knowledge Discovery*, pages 15–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew R. Gormley. 2023. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. *arXiv preprint arXiv:2310.01387v1*.

Frédéric Blain, Lucia Specia, and Pranava Madhyastha. 2017. Exploring hypotheses spaces in neural machine translation. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 282–298, Nagoya Japan.

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.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104.

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.

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

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

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

Mara Finkelstein and Markus Freitag. 2023. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. *arXiv preprint arXiv:2309.10966v1*.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. *arXiv preprint arXiv:2305.09860v2*.

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.

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

Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. 2019. Statistical analysis of nearest neighbor methods for anomaly detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

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

Yuu Jinnai and Kaito Ariu. 2024. Hyperparameter-free approach for faster minimum bayes risk decoding. *arXiv preprint arXiv:2401.02749v1*.

Yuu Jinnai, Ukyo Honda, Tetsuro Morimura, and Peinan Zhang. 2024. Generating diverse and high-quality texts by minimum bayes risk decoding. *arXiv preprint arXiv:2401.05054v1*.

Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. 2023. Model-based minimum bayes risk decoding. *arXiv preprint arXiv:2311.05263v1*.

Hans-Peter Kriegel, Peer Kroger, Erich Schubert, and Arthur Zimek. 2011. Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining (SDM)*, pages 13–24. SIAM.

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.

Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statistics. In *Proceedings of the National Institute of Science of India*, volume 12, pages 49–55. National Institute of Science of India.

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.

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.

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. 2022. 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, 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.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. 2021. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795.

Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. 2014. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data mining and knowledge discovery*, 28:190–237.

Raphael Shu and Hideki Nakayama. 2017. Later-stage minimum Bayes-risk decoding for neural machine translation. *arXiv preprint arXiv:1704.03169v2*.

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.

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

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. *arXiv preprint arXiv:2311.08380v1*.

Candidate	de-en	en-de	ru-en	en-ru
Ancestral	62.31	61.38	49.83	64.40
Beam	73.27	71.82	58.08	78.19
Epsilon ($\epsilon = 0.02$)	**73.65**	**72.17**	**59.46**	**80.86**
Nucleus ($p = 0.6$)	71.11	70.93	56.56	80.12
Nucleus ($p = 0.9$)	71.87	70.84	58.01	78.33

Table 4: *Oracle scores* in COMET20 to determine the sampling method *for candidates*. The results are the average of three runs with different seeds except for beam search. The best/worst scores are in **bold**/underlined.

	Pseudo-Reference	de-en	en-de	ru-en	en-ru
Epsilon ($\epsilon = 0.02$)	Ancestral	58.77	64.21	**47.92**	71.19
	Beam	57.24	63.49	46.56	68.26
	Epsilon ($\epsilon = 0.02$)	59.07	65.32	46.56	71.85
	Epsilon ($\epsilon = 0.02$)*	59.19	65.25	46.59	71.79
	Nucleus ($p = 0.6$)	57.82	64.31	45.42	71.08
	Nucleus ($p = 0.9$)	**59.65**	**65.44**	47.77	**72.33**

Table 5: COMET20 scores of MBR decoding with different pseudo-references. Candidates are sampled with epsilon sampling ($\epsilon = 0.02$). Epsilon ($\epsilon = 0.02$)* shows the results of sampling candidates and pseudo-references with the same epsilon sampling but with different seeds. The results are the average of three runs with different seeds except for beam search. The best/worst scores are in **bold**/underlined.

A Results with COMET20

To support the analysis of this study with a different utility function, we conducted the same experiments as in Tables 1, 2, and 3 using COMET20 (Rei et al., 2020). Tables 4, and 5, 6 show the same tendency: the performance varies by sampling methods almost consistently, and the anomaly scores achieve the best correlations with the performance variation. These results confirm the validity of our analysis even with other utility functions.

B Detailed Results

Tables 7 and 8 show the results used to calculate the Spearman's ρ in Tables 3 and 6, respectively.

	de-en	en-de	ru-en	en-ru
Avg. Prob.$_{(-)}$	$0.580_{(\checkmark)}$	$0.290_{(\checkmark)}$	$0.928_{(\checkmark)}$	$0.638_{(\checkmark)}$
Cum. Prob.$_{(+)}$	$0.058_{(\times)}$	$0.116_{(\times)}$	$0.406_{(\times)}$	$0.058_{(\times)}$
Cand. Sim.$_{(+)}$	$0.600_{(\times)}$	$0.257_{(\times)}$	$0.943_{(\times)}$	$0.600_{(\times)}$
Ref. Sim.$_{(+)}$	$0.580_{(\times)}$	$0.290_{(\times)}$	$0.928_{(\times)}$	$0.638_{(\times)}$
$d_{M(-)}$	$0.714_{(\checkmark)}$	$0.543_{(\checkmark)}$	$0.886_{(\checkmark)}$	$0.829_{(\checkmark)}$
kNN$_{(-)}$				
$k = 5$	$0.829_{(\checkmark)}$	$0.714_{(\checkmark)}$	$\mathbf{1.000}_{(\checkmark)}$	$0.771_{(\checkmark)}$
$k = 25$	$\mathbf{1.000}_{(\checkmark)}$	$\mathbf{0.886}_{(\checkmark)}$	$\mathbf{1.000}_{(\checkmark)}$	$\mathbf{0.886}_{(\checkmark)}$
$k = 50$	$0.829_{(\checkmark)}$	$\mathbf{0.886}_{(\checkmark)}$	$0.943_{(\checkmark)}$	$0.771_{(\checkmark)}$
$k = 75$	$0.829_{(\checkmark)}$	$\mathbf{0.886}_{(\checkmark)}$	$0.143_{(\checkmark)}$	$0.771_{(\checkmark)}$
$k = 100$	$0.143_{(\checkmark)}$	$0.257_{(\checkmark)}$	$0.086_{(\checkmark)}$	$0.086_{(\checkmark)}$
LOF$_{(-)}$				
$k = 5$	$0.771_{(\checkmark)}$	$0.829_{(\checkmark)}$	$0.886_{(\checkmark)}$	$0.829_{(\checkmark)}$
$k = 25$	$0.771_{(\checkmark)}$	$0.829_{(\checkmark)}$	$0.943_{(\checkmark)}$	$0.829_{(\checkmark)}$
$k = 50$	$0.771_{(\checkmark)}$	$0.829_{(\checkmark)}$	$0.943_{(\checkmark)}$	$0.829_{(\checkmark)}$
$k = 75$	$0.829_{(\checkmark)}$	$0.829_{(\checkmark)}$	$0.943_{(\checkmark)}$	$0.829_{(\checkmark)}$
$k = 100$	$0.657_{(\checkmark)}$	$0.486_{(\checkmark)}$	$\mathbf{1.000}_{(\checkmark)}$	$0.486_{(\checkmark)}$

Table 6: Correlation coefficients (Spearman's ρ) between COMET20 performance variation and pseudo-references' properties or anomaly scores. We show the absolute value of ρ. The subscript signs $_{(+/-)}$ are the expected signs of ρ (see Section 4), and the subscript marks $_{(\checkmark/\times)}$ show whether the actual signs match/mismatch the expected ones. The best/worst scores are in **bold**/underlined.

	Pseudo-Reference	COMET22↑	Prob. Avg.↓	Prob. Cum.↑	Sim. Cand.↑	Sim. Ref.↑	d_M↓	kNN↓ 5	kNN↓ 25	kNN↓ 50	kNN↓ 75	kNN↓ 100	LOF↓ 5	LOF↓ 25	LOF↓ 50	LOF↓ 75	LOF↓ 100
	de-en																
Epsilon (ϵ = 0.02)	Ancestral	85.82	**-3.59**	0.87	71.20	60.45	8.02	0.22	0.39	0.62	0.88	1.30	1.05	1.07	1.10	1.03	1.00
	Beam	85.62	-0.76	**1.02**	85.44	83.01	15.10	0.33	0.41	0.46	0.50	0.54	1.75	1.55	1.47	1.25	1.00
	Epsilon (ϵ = 0.02)	85.89	-0.89	0.97	84.81	82.18	**6.94**	0.26	0.35	0.41	0.46	0.53	1.08	1.09	1.08	1.01	1.00
	Epsilon (ϵ = 0.02)*	85.87	-0.89	0.97	84.70	82.18	8.61	0.23	0.33	0.39	0.45	**0.51**	1.10	1.10	1.09	1.02	1.00
	Nucleus (p = 0.6)	85.69	-0.70	0.83	**85.63**	**83.24**	16.47	0.32	0.39	0.45	0.49	0.53	1.62	1.43	1.37	1.16	1.00
	Nucleus (p = 0.9)	**86.04**	-1.50	0.95	81.66	78.02	7.38	**0.18**	**0.27**	**0.35**	**0.42**	0.57	**1.04**	**1.02**	**1.01**	**0.99**	1.00
	en-de																
Epsilon (ϵ = 0.02)	Ancestral	87.51	**-3.60**	0.65	68.41	51.56	8.16	0.22	0.44	0.74	1.09	1.71	1.08	1.11	1.11	1.03	1.00
	Beam	87.40	-0.65	**0.74**	87.02	85.06	15.86	0.30	0.37	0.41	0.44	0.47	2.05	1.80	1.64	1.36	1.00
	Epsilon (ϵ = 0.02)	87.74	-0.80	0.71	86.33	83.96	**6.52**	0.23	0.31	0.36	0.40	0.45	1.09	1.11	1.09	1.01	1.00
	Epsilon (ϵ = 0.02)*	87.74	-0.80	0.71	86.24	83.96	8.22	0.21	0.29	0.35	0.39	**0.44**	1.11	1.13	1.11	1.02	1.00
	Nucleus (p = 0.6)	87.57	-0.61	0.62	**87.12**	**85.10**	14.95	0.29	0.35	0.39	0.43	0.46	1.68	1.49	1.42	1.21	1.00
	Nucleus (p = 0.9)	**87.82**	-1.30	0.70	83.84	79.38	7.08	**0.16**	**0.24**	**0.31**	**0.37**	0.49	**1.04**	**1.04**	**1.02**	**1.00**	1.00
	ru-en																
Epsilon (ϵ = 0.02)	Ancestral	88.41	**-3.75**	0.60	70.67	59.20	8.13	**0.20**	**0.35**	0.56	0.83	1.25	**1.04**	**1.03**	**1.03**	**1.00**	**1.00**
	Beam	87.78	-0.74	**0.74**	85.71	79.69	24.41	0.60	0.68	0.73	0.77	0.81	3.02	2.65	2.40	1.99	1.28
	Epsilon (ϵ = 0.02)	88.46	-0.89	0.69	85.18	79.03	11.26	0.48	0.60	0.67	0.72	0.77	1.30	1.40	1.44	1.34	**1.00**
	Epsilon (ϵ = 0.02)*	88.46	-0.89	0.69	85.09	79.03	11.19	0.46	0.58	0.65	0.71	0.76	1.33	1.46	1.51	1.39	**1.00**
	Nucleus (p = 0.6)	88.26	-0.69	0.60	**85.96**	**79.91**	25.16	0.59	0.67	0.73	0.77	0.81	2.89	2.40	2.18	1.83	1.15
	Nucleus (p = 0.9)	**88.61**	-1.52	0.67	82.07	75.17	**7.85**	0.24	0.41	**0.52**	**0.61**	**0.75**	1.05	1.09	1.14	1.02	**1.00**
	en-ru																
Epsilon (ϵ = 0.02)	Ancestral	82.02	**-3.85**	0.37	71.25	55.20	8.91	0.23	0.42	0.67	0.97	1.40	1.07	1.13	1.13	1.05	**1.00**
	Beam	81.64	-0.72	**0.44**	**87.47**	**84.73**	20.91	0.38	0.44	0.48	0.51	0.54	2.42	2.04	1.85	1.51	1.02
	Epsilon (ϵ = 0.02)	82.01	-0.94	0.42	86.68	83.57	**7.27**	0.28	0.37	0.42	0.46	0.51	1.10	1.12	1.11	1.03	**1.00**
	Epsilon (ϵ = 0.02)*	81.98	-0.94	0.42	86.58	83.57	9.04	0.26	0.35	0.40	0.45	**0.50**	1.11	1.14	114	1.05	**1.00**
	Nucleus (p = 0.6)	81.76	-0.80	0.37	87.06	84.26	13.17	0.30	0.38	0.43	0.47	0.51	1.35	1.36	1.35	1.17	**1.00**
	Nucleus (p = 0.9)	**82.18**	-1.69	0.40	83.10	77.69	7.76	**0.18**	**0.26**	**0.33**	**0.40**	0.55	**1.03**	**1.02**	**1.01**	**0.99**	**1.00**

Table 7: Results used to calculate the Spearman's ρ in Table 3. Candidates are sampled with epsilon sampling (ϵ = 0.02). Epsilon (ϵ = 0.02)* shows the results of sampling candidates and pseudo-references with the same epsilon sampling but with different seeds. Avg. Prob. is the log probability. ↑ and ↓ denote that the values are considered to be better when they are higher and lower, respectively. The best/worst scores are in **bold**/underlined.

			Prob.		Sim.			kNN↓					LOF↓				
	Pseudo-Reference	COMET20↑	Avg.↑	Cum.↑	Cand.↑	Ref.↑	d_M↓	5	25	50	75	100	5	25	50	75	100
	de-en																
Epsilon (ϵ = 0.02)	Ancestral	58.77	-3.59	0.87	-0.11	-0.51	17.65	1.13	1.98	3.05	4.41	6.56	**1.05**	1.05	1.07	1.01	1.00
	Beam	57.24	-0.76	**1.02**	0.59	0.46	46.37	1.67	2.08	2.34	2.55	2.81	1.82	1.62	1.51	1.28	1.00
	Epsilon (ϵ = 0.02)	59.07	-0.89	0.97	0.56	0.43	21.33	1.27	1.76	2.10	2.37	2.69	1.12	1.13	1.11	1.01	1.00
	Epsilon (ϵ = 0.02)*	59.19	-0.89	0.97	0.56	0.43	23.26	1.19	1.69	2.03	2.31	**2.64**	1.13	1.14	1.11	1.02	1.00
	Nucleus (p = 0.6)	57.82	**-0.70**	0.83	**0.60**	**0.48**	55.60	1.59	1.98	2.26	2.48	2.71	1.71	1.52	1.44	1.18	1.00
	Nucleus (p = 0.9)	**59.65**	-1.50	0.95	0.41	0.26	**17.29**	**0.93**	**1.39**	**1.79**	**2.19**	2.95	**1.05**	**1.03**	**1.01**	**0.99**	1.00
	en-de																
Epsilon (ϵ = 0.02)	Ancestral	64.21	-3.60	0.65	-0.14	-0.80	14.50	0.87	1.62	2.65	4.15	7.24	1.06	1.07	1.05	**1.00**	1.00
	Beam	63.49	-0.65	**0.75**	0.62	**0.56**	40.90	1.15	1.38	1.54	1.66	1.82	2.06	1.75	1.57	1.29	1.00
	Epsilon (ϵ = 0.02)	65.32	-0.80	0.71	0.60	0.52	16.66	0.90	1.19	1.40	1.57	1.78	1.10	1.10	1.07	1.01	1.00
	Epsilon (ϵ = 0.02)*	65.25	-0.80	0.71	0.60	0.52	18.18	0.84	1.14	1.36	1.52	**1.73**	1.13	1.13	1.09	1.01	1.00
	Nucleus (p = 0.6)	64.31	**-0.61**	0.62	**0.63**	**0.56**	42.62	1.11	1.34	1.51	1.65	1.79	1.84	1.47	1.34	1.14	1.00
	Nucleus (p = 0.9)	**65.44**	-1.30	0.70	0.51	0.35	**14.21**	**0.67**	**0.99**	**1.24**	**1.48**	1.97	**1.05**	**1.05**	**1.02**	**1.00**	1.00
	ru-en																
Epsilon (ϵ = 0.02)	Ancestral	**47.92**	-3.75	0.60	-0.05	-0.45	**17.37**	**0.96**	**1.63**	2.49	3.65	5.68	**1.04**	**1.03**	**1.02**	**1.00**	**1.00**
	Beam	44.90	-0.74	**0.74**	0.62	0.36	68.23	2.56	2.94	3.18	3.36	3.55	3.01	1.54	2.40	1.98	1.15
	Epsilon (ϵ = 0.02)	46.56	-0.89	0.69	0.59	0.34	29.48	1.99	2.54	2.88	3.13	3.38	1.31	1.44	1.45	1.28	**1.00**
	Epsilon (ϵ = 0.02)*	46.59	-0.89	0.69	0.59	0.34	30.97	1.89	2.47	2.82	3.07	**3.33**	1.33	1.47	1.49	1.31	**1.00**
	Nucleus (p = 0.6)	45.42	**-0.69**	0.60	**0.63**	**0.37**	87.59	2.49	2.88	3.15	3.35	3.53	3.12	2.40	2.18	1.76	1.08
	Nucleus (p = 0.9)	47.77	-1.52	0.67	0.45	0.18	19.08	1.10	1.79	**2.27**	**2.67**	**3.33**	1.06	1.11	1.13	1.01	**1.00**
	en-ru																
Epsilon (ϵ = 0.02)	Ancestral	71.19	-3.85	0.37	0.02	-0.57	16.23	1.03	1.71	2.60	3.81	5.85	1.06	1.08	1.08	1.02	1.00
	Beam	68.26	**-0.72**	**0.44**	**0.69**	**0.57**	58.19	1.64	1.96	2.16	2.32	2.50	2.33	1.91	1.73	1.41	1.00
	Epsilon (ϵ = 0.02)	71.85	-0.94	0.42	0.65	0.52	16.62	1.21	1.61	1.88	2.10	2.36	1.09	1.12	1.10	1.02	1.00
	Epsilon (ϵ = 0.02)*	71.79	-0.94	0.42	0.65	0.52	20.76	1.12	1.53	1.80	2.03	**2.28**	1.11	1.13	1.12	1.03	1.00
	Nucleus (p = 0.6)	71.08	-0.80	0.37	0.67	0.56	36.80	1.30	1.67	1.92	2.12	2.34	1.37	1.32	1.32	1.14	1.00
	Nucleus (p = 0.9)	**72.33**	-1.69	0.41	0.49	0.30	**15.40**	**0.83**	**1.21**	**1.52**	**1.84**	2.52	**1.03**	**1.02**	**1.01**	**0.99**	1.00

Table 8: Results used to calculate the Spearman's ρ in Table 6. Candidates are sampled with epsilon sampling (ϵ = 0.02). Epsilon (ϵ = 0.02)* shows the results of sampling candidates and pseudo-references with the same epsilon sampling but with different seeds. Avg. Prob. is the log probability. ↑ and ↓ denote that the values are considered to be better when they are higher and lower, respectively. The best/worst scores are in **bold**/underlined.