Document-Level Text Generation with Minimum Bayes Risk Decoding using Optimal Transport

Yuu Jinnai
CyberAgent
jinnai_yu@cyberagent.co.jp

Abstract

Document-level text generation tasks are known to be more difficult than sentence-level text generation tasks as they require the understanding of longer context to generate high-quality texts. In this paper, we investigate the adaption of Minimum Bayes Risk (MBR) decoding for document-level text generation tasks. MBR decoding makes use of a utility function to estimate the output with the highest expected utility from a set of candidate outputs. Although MBR decoding is shown to be effective in a wide range of sentence-level text generation tasks, its performance on document-level text generation tasks is limited as many of the utility functions are designed for evaluating the utility of sentences. To this end, we propose MBR-OT, a variant of MBR decoding using Wasserstein distance to compute the utility of a document using a sentence-level utility function. The experimental result shows that the performance of MBR-OT outperforms that of the standard MBR in document-level machine translation, text simplification, and dense image captioning tasks. Our code is available at https://github.com/jinnaiyuu/mbr-optimal-transport.

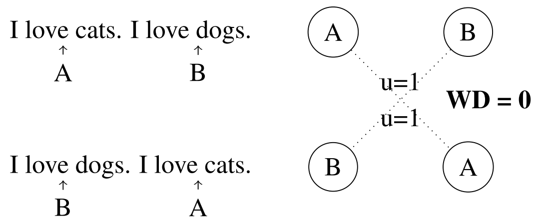


Figure 1: Illustrative example of a metric using Wasserstein distance over two texts "I love cats. I love dogs" and "I love dogs. I love cats.". Each output is segmented into a set of segments (e.g., sentence) and a utility function is used to compute the utility over a pair of segments from each of the outputs. Wasserstein distance is flexible to the change in the structure of the text, making it an adaptive measure for a wide range of tasks.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various natural language processing tasks (Stiennon et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; Dubey et al., 2024; OpenAI et al., 2024). While many text-generation tasks are evaluated at the sentence level, LLMs are also capable of generating text at the document level (Wang et al., 2023; Xia et al., 2024; Zhang et al., 2024). This raises the need for evaluating the performance of decoding algorithms for document-level text generation tasks.

Minimum Bayes Risk (MBR) decoding has been shown to be highly effective for sentence-level directed text generation tasks (Goel and Byrne, 2000; Freitag et al., 2022a; Eikema and Aziz, 2022; Freitag et al., 2023a). However, its effectiveness for generating longer texts is less investigated. In this paper, we evaluate the performance of MBR decoding in document-level text generation tasks. Specifically, we propose MBR-OT, a variant of MBR decoding that uses an optimal transport distance as a utility function. Optimal transport has been used in many fields to measure the dissimilarity of two probability distributions (Peyré and Cuturi, 2020; Villani, 2021). We use optimal transport as a tool to evaluate the utility between documents using sentence-level utility functions. This approach enables the use of the sentence-level utility functions that have been investigated for years to improve their accuracy by many researchers (Freitag et al., 2022b, 2023b, 2024).

We evaluate MBR-OT on multiple directed text generation tasks: document-level machine translation, document-level text simplification, and dense image captioning. Our results show that MBR decoding consistently outperforms the baselines across these tasks, showing its effectiveness for document-level text generation tasks.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28260–28279
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Background

We introduce MBR decoding as one of the algorithms to solve the task. Then, we explain the concept of optimal transport which is used by the proposed method.

2.1 Minimum Bayes Risk (MBR) Decoding

Unlike maximum a posteriori (MAP) decoding (Eikema and Aziz, 2020; Holtzman et al., 2020), which aims to find the most probable output, minimum Bayes risk (MBR) decoding selects the output that maximizes the expected utility, which is equivalent to minimizing risk (Goel and Byrne, 2000; Kumar and Byrne, 2002, 2004).

MBR decoding consists of two key components: a text generation model P_{model} and a utility metric u. The model estimates the probability of an output $\mathbf{y}$ given an input sentence $\mathbf{x}$. The utility metric, $u(\mathbf{h}, \mathbf{y})$, measures the quality of a candidate output $\mathbf{h}$ with respect to a reference output $\mathbf{y}$.

Let $\mathcal{Y}$ be a set of all possible sequences. Given a set of candidate hypotheses $\mathcal{H}_{\text{cand}} \subseteq \mathcal{Y}$, MBR decoding selects the hypothesis that maximizes its expected utility:

$$\mathbf{h}^{\text{human}} = \underset{\mathbf{h} \in \mathcal{H}_{\text{cand}}}{\arg\max} \sum_{\mathbf{y} \in \mathcal{Y}} u(\mathbf{h}, \mathbf{y}) \cdot P_{\text{human}}(\mathbf{y} \mid \mathbf{x}). \tag{1}$$

Since the true human probability distribution, P_{human}, is unknown, MBR approximates it using the model probability P_{model}:

$$\mathbf{h}^{\text{model}} = \underset{\mathbf{h} \in \mathcal{H}_{\text{cand}}}{\arg\max} \sum_{\mathbf{y} \in \mathcal{Y}} u(\mathbf{h}, \mathbf{y}) \cdot P_{\text{model}}(\mathbf{y} \mid \mathbf{x}). \tag{2}$$

For simplicity, we denote P_{model} as P throughout the remainder of this paper unless stated otherwise.

Since integrating over $\mathcal{Y}$ is computationally intractable, Eq. (2) is typically approximated using a Monte Carlo estimate (Eikema and Aziz, 2022; Farinhas et al., 2023). This is done by sampling a set of reference hypotheses $\mathcal{H}_{\text{ref}}$, from the model P:

$$\mathbf{h}^{\text{MBR}} = \underset{\mathbf{h} \in \mathcal{H}_{\text{cand}}}{\arg\max} \frac{1}{N} \sum_{\mathbf{y} \in \mathcal{H}_{\text{ref}}} u(\mathbf{h}, \mathbf{y}), \tag{3}$$

where $N = |\mathcal{H}_{\text{ref}}|$.

A common practice is to use the same set of hypotheses for both the candidate pool ($\mathcal{H}$) and the reference pool ($\mathcal{H}_{\text{ref}}$). We follow the practice in this paper and assume $\mathcal{H}_{\text{cand}} = \mathcal{H}_{\text{ref}}$.

2.2 Optimal Transport (OT)

Optimal Transport (OT; Peyré and Cuturi 2020; Villani 2021) provides a mathematical framework for quantifying the dissimilarity between two distributions. In natural language processing (NLP), OT has been widely used to measure text similarity, often referred to as the Earth Mover's Distance (Kusner et al., 2015; Zhao et al., 2019). It has been applied in various contexts, including document similarity (Kusner et al., 2015) and summary evaluation (Zhao et al., 2019).

While existing metrics for document-level machine translation have been proposed (Vernikos et al., 2022), they are typically designed for tasks where generated documents can be segmented into a fixed sequence of corresponding segments. This limitation makes them unsuitable for scenarios where segment order or count varies across documents.

On the other hand, a key advantage of OT-based metrics is their adaptability to text structure. They can effectively handle variations such as sentence reordering and merging, which frequently occur in machine translation (Hovy and Gerber, 1997; Marcu et al., 2000). For instance, due to structural differences between Japanese and English, professional translators often restructure paragraphs during translation, significantly altering sentence order and merging content (Hovy and Gerber, 1997).

In this paper, we consider multiple OT formulations, including linear assignment, Wasserstein distance, and entropic regularized Wasserstein distance.

Linear assignment (LA). Linear assignment (LA) is a simple formulation of OT where the cost is computed as a linear sum of the cost of each element (Peyré and Cuturi, 2020; Villani, 2021). Let $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ be a set of sentences. Let $p_{\mathbf{h}}$ and $p_{\mathbf{y}}$ be a probability distribution over a set of elements in $\mathbf{h}$ and $\mathbf{y}$. Let C be a non-negative function representing the cost or the dissimilarity between two sentences. Using LA, the cost between two sets of sentences is defined as follows:

$$\begin{aligned} &\text{LA}_C[p_{\mathbf{h}} \| p_{\mathbf{y}}] = \\ &\inf_{\gamma \in \Gamma_L(\mathbf{h}, \mathbf{y})} \sum_{i \in \{1..m\}} p_{\mathbf{h}}(h_i) C(h_i, \gamma(h_i)), \end{aligned} \tag{4}$$

where $\Gamma_L(\mathbf{h}, \mathbf{y})$ is a set of deterministic mappings from $\mathbf{h}$ to $\mathbf{y}$ that is injective if $m \leq n$, and subjective if $m \geq n$ (thus bijective if $m = n$).

LA assigns one sentence in a source to exactly one sentence in a target and computes the sum of the cost between them. One of the failure cases of this constraint is when aligning texts with and without *merge*.

(1) I like cats and dogs.

(2) I like cats. I like dogs.

Because LA cannot distribute weights of a source sentences probabilistically, it has to assign "I like cats and dogs." to only one of either "I like cats." or "I like dogs.". Thus, although the two texts have very high similarity, its utility computed by LA gets small.

Wasserstein distance (WD). The shortcoming of LA is that it is constrained to assign the weight of the source element to a single target element, and thus cannot adapt to the change in disclosure. Figure 1 shows the example of sentences where this is undesirable. Wasserstein distance (WD) is a generalization of LA where the source element can divide its weights and assign it to any number of target elements (Peyré and Cuturi, 2020; Villani, 2021). This enables the alignment of the sequence like in Figure 1 where one source sentence is divided into two target sentences.

WD is computed as follows:

$$\mathrm{WD}_C[p_{\mathbf{h}} \| p_{\mathbf{y}}] = \inf_{\gamma \in \Gamma(p_{\mathbf{h}}, p_{\mathbf{y}})} \sum_{(i,j) \in m \times n} \gamma(h_i, y_j)\, C(h_i, y_j), \quad (5)$$

where $\Gamma(p_{\mathbf{h}}, p_{\mathbf{y}})$ is a set of all possible joint distributions γ whose marginals are $p_{\mathbf{h}}$ and $p_{\mathbf{y}}$.

WD is a metric used to quantify the dissimilarity between two probability distributions. Intuitively, it measures the minimum cost required to transform one distribution into the other. This cost is conceptualized as the amount of probability mass that must be moved multiplied by the distance that would be moved.

The advantage of WD over LA is that it can assign multiple reference segments to a single source sentences. In the case of "I like cats and dogs.", one can distribute its weight to both ""I like cats." and "I like dogs.", resulting in high utility score.

Entropic regularized WD (EWD). Entropic regularized WD (EWD) is an extension of WD where the KL regularization is enforced to the joint probability distribution γ (Peyré and Cuturi, 2020; Villani, 2021):

$$\begin{aligned} &\mathrm{WD}_C^{\epsilon}[p_{\mathbf{h}} \| p_{\mathbf{y}}] \\ &= \inf_{\gamma \in \Gamma(p_{\mathbf{h}}, p_{\mathbf{y}})} \sum_{(i,j) \in m \times n} \gamma(h_i, y_j)\, C(h_i, y_j) \\ &\qquad + \epsilon \mathrm{KL}[\gamma \| p_{\mathbf{h}} \oplus p_{\mathbf{y}}], \end{aligned} \quad (6)$$

where KL represents the KL-divergence between the two probability distributions and ϵ is a parameter to choose the weight on the KL-divergence term.

Intuitively, EWD is a WD plus the cost of every sentence pair is considered. The KL-divergence term requires the joint distribution γ to be spread across $\mathbf{y}$. Thus, the KL term is smaller if the two documents are similar to each other overall in addition to the cost of the optimal assignment between pairs of sentences.

EWD is known to be robust under the model uncertainty (Azizian et al., 2023) and is fast to compute using the Sinkhorn algorithm (Cuturi, 2013).

3 MBR-OT: MBR Decoding using OT

The performance of MBR decoding relies on the utility function. However, many of the state-of-the-art utility functions are developed for sentence-level evaluation and are not trained to predict document-level utility.

To this end, we propose **MBR-OT**, a variant of MBR decoding that uses WD as the utility function for MBR decoding. Let the utility function between two sentences be u_{s} where we assume its range to be $[0, 1]$.

Let $p_{\mathbf{h}}$ be a discrete probability distribution over $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_m\}$ and so is $p_{\mathbf{y}}$ for $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$. We propose the WD between the two probability distributions to be the utility function of MBR-OT. Formally, the utility function is defined as follows:

$$u(\mathbf{h}, \mathbf{y}) = 1 - \mathrm{OT}[p_{\mathbf{h}} \| p_{\mathbf{y}}], \quad (7)$$

where

$$C(h_i, y_j) = 1 - u_{\mathrm{s}}(h_i, y_j), \quad (8)$$

where u_{s} is a utility function to be used for computing the utility of two segments. In the following experiments, we use sentence-level utility functions such as BLEU and MetricX for u_{s}. Figure 1 shows the illustrative example of how it is computed.

We use a uniform distribution for constructing $p(\mathbf{h})$ and $p(\mathbf{y})$ so that each sentence is weighted equally:

$$p(\mathbf{h}_i) = \frac{1}{|\mathbf{h}|}. \tag{9}$$

An alternative is to weight a probability proportional to its length so that each token is weighted equally:

$$p_L(\mathbf{h}_i) = \frac{|\mathbf{h}_i|}{\sum_j |\mathbf{h}_j|}. \tag{10}$$

Incorporating segment informativeness may enhance the performance of our proposed method. This approach may be particularly beneficial in evaluating informal texts where there are large variations in segment informativeness.

The choice of the optimal transport formulation depends on the structure of the language and the objective of the task. We call the MBR-OT using AL, WD, and EWD for the OT in Eq. (7) as MBR-AL, MBR-WD, and MBR-WD$^\epsilon$. In particular, we denote these algorithms using Eq. (10) as MBR-AL$_L$, MBR-WD$_L$, and MBR-WD$^\epsilon_L$.

The strength of the method is that it can make the best use of the sentence-level utility functions. Sentence-level utility functions are meticulously developed by researchers and engineers and they are shown to have a very high correlation with human evaluation (Rei et al., 2022; Guerreiro et al., 2024). Compared to the effort on sentence-level utility functions, document-level utility functions are less investigated. In fact, many of them are based on using sentence-level utility functions efficiently (Liu et al., 2020; Vernikos et al., 2022).

Optimizations for MBR-OT using WD. There are several optimizations applicable to speed up MBR-OT when using WD. First, because WD (and EWD) are metrics, they are guaranteed to be symmetric:

$$\forall \mathbf{h} \forall \mathbf{y} : u(\mathbf{h}, \mathbf{y}) = u(\mathbf{y}, \mathbf{h}). \tag{11}$$

Also, the value of WD is 1 when the two distributions are the same. Thus,

$$\forall \mathbf{h} : u(\mathbf{h}, \mathbf{h}) = 0. \tag{12}$$

Therefore, for N documents, one only needs to compute $\frac{N(N-1)}{2}$ pairs of documents instead of N^2 pairs.

Second, because WD is a metric in a finite-dimensional space, it is likely that the matrix of utility scores can be well approximated by a low-rank matrix (Drineas and Mahoney, 2005; Holodnak and Ipsen, 2015). This enables the optimization by a low-rank approximation, significantly reducing the computational complexity (Trabelsi et al., 2024).

Third, we can train a document-level utility function by distilling the WD metric, or train the language model directly. There are several studies showing that the performance of LLM can be improved by distilling the output of MBR decoding (Yan et al., 2023; Ramos et al., 2024; Wu et al., 2024; Yang et al., 2024; Finkelstein and Freitag, 2024; Guttmann et al., 2024). These approaches require additional training but solve the computational overhead at decoding time.

In the experiments of the paper, we only use the first optimization so that the values are computed exactly.

4 Experiments

We first evaluate the accuracy of the WD metric for machine translation to assess if it is a metric suitable for a document-level evaluation. We evaluate the performance of MBR decoding and MBR-OT on machine translation, document simplification, and dense image captioning tasks.

In all the experiments, we divide the output into a set of sentences for MBR-OT. The default sentencizer in spaPy[1] is used for English and German. GiNZA NLP Library[2] is used for Japanese as the sentencizer in spaPy is incompatible with Japanese. See Appendix E for the hyperparameters used for the experiments and Appendix F for the implementation details.

4.1 Evaluation of WD Metrics for Machine Translation

We evaluate the accuracy of the WD metric using the Metric Shared Task on WMT22 and WMT23 (Freitag et al., 2022b, 2023b). We compare the correlation of the metrics with the human evaluation. As a baseline, we compare 1. using the metric to evaluate each segment (Base), and 2. using the document-level evaluation method by Vernikos et al. (2022) to evaluate each segment but with a context of the document (Doc). On computing the score using the WD, we compute the utility of the entire document without using the segmentation

[1] https://github.com/explosion/spaCy

[2] https://github.com/megagonlabs/ginza

data	lp	Utility	base	doc	ot (Ours)
wmt22	en-de	sacrebleu	0.6420	0.6783	0.7090
		BERTScore	0.7843	0.7670	0.8060
		SentBERT	0.8630	0.8313	0.8681
	en-ru	sacrebleu	0.7861	0.7480	0.7919
		BERTScore	0.8106	0.7864	0.8061
		SentBERT	0.8030	0.8437	0.8114
wmt23	en-de	sacrebleu	0.8911	0.9165	0.9384
		BERTScore	0.8906	0.9212	0.9553
		MetricX-23	0.9876	0.9874	0.9753
	he-en	sacrebleu	0.8751	0.8853	0.8716
		BERTScore	0.8942	0.9449	0.7138
		MetricX-23	0.9357	0.9600	0.8072

Table 1: System level correlation of the metrics with the human evaluation. Base: sentence-level metric, Doc: a document-level metric by Vernikos et al. (2022), WD: a document-level metric using WD.

MetricX-23	En-Ja	En-De
Beam	61.57	79.07
MBR (SentBERT)	56.26	78.66
MBR (COMET-22)	60.55	80.92
MBR (SFR-2)	57.38	76.88
MBR (MetricX)	68.81	82.02
MBR-LA (MetricX)	70.01	80.77
MBR-LA$_L$ (MetricX)	71.01	83.13
MBR-WD$^\epsilon$ (MetricX)	68.07	**83.77**
MBR-WD$^\epsilon_L$ (MetricX)	70.67	83.24
MBR-WD (MetricX)	**75.29**	83.40
MBR-WD$_L$ (MetricX)	72.38	83.24

Table 2: Comparison of MBR-OT methods with LA, WD, and EWD. The evaluation metric is MetricX-23-XXL with EWD.

COMET-22	En-Ja CALM2-DPO	En-De EuroLLM
MBR	79.65	83.03
MBR-WD$^\epsilon$	83.54	**85.13**
MBR-WD$^\epsilon_L$	**84.08**	84.25

Table 3: Evaluation of MBR-WD$^\epsilon$ on WMT24 with 32 samples on LLMs specifically trained for the languages (`cyberagent/calm2-7b-chat-dpo-experimental` and `utter-project/EuroLLM-1.7B-Instruct`). The evaluation metric is MetricX-23-XXL with EWD.

provided. Then, we use the average score over a set of documents as the system score. We set $\epsilon = 0$ and use a uniform weight (Eq. 9) for the probability distribution of WD.

We evaluate BLEU score (Papineni et al., 2002) using sacrebleu library (Post, 2018), BERTScore (Zhang et al., 2020) (`bert-base-multilingual-cased`),[3] and a cosine distance of the sentence embedding model (SentBERT; `sentence-transformers/all-mpnet-base-v2`), and MetricX-23 (`google/metricx-23-xl-v2p0`; Juraska et al., 2023).

Table 1 shows the correlation of the metrics with human evaluation. Because MetricX-23 is trained with the WMT22 dataset, we evaluate SentBERT instead. Overall, we observe the WD metric to be on par with the accuracy of segment-level and document-level evaluation metrics, except for the WMT23 he-en. The authors are not familiar with Hebrew so are not aware of the reason.

Note that the WD metric does not exploit the fact that the sentences of the documents are aligned in the same order. Thus, it uses less information than the baselines. The result shows that the proposed method is on par with the state-of-the-art metric, while it is a robust metric that is applicable to a wide range of tasks where the disclosure structures can vary.

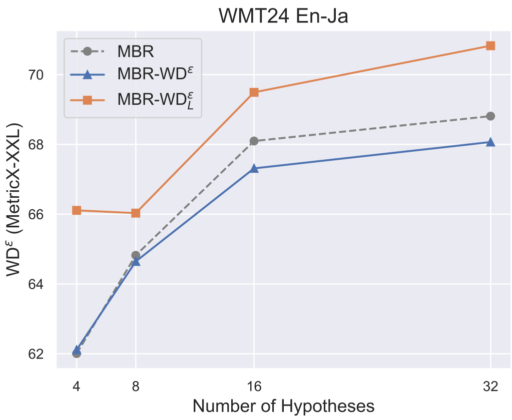
4.2 Document-Level Translation

The use of LLMs for document-level translation is shown to be effective due to their ability to comprehend long-context texts (Wang et al., 2023).

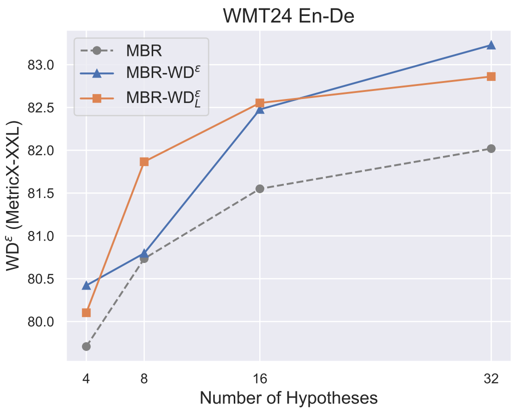
We use WMT24 En-De and En-Ja language pairs for the evaluation (Kocmi et al., 2024). We generate 32 samples using Llama-3.1 (`meta-llama/Llama-3.1-8B-Instruct`) as a language model (Dubey et al., 2024). See Appendix D for the prompt used.

We use MetricX-23 (`google/metricx-23-xl-v2p0`) as the utility function (Juraska et al., 2023). Because the output from the MetricX-23 models is a score in the range $[0, 25]$ where lower is better, we inverted and rescaled it to $[0, 1]$. For an evaluation, we use MetricX-23 with EWD ($\epsilon = 0.1$) to evaluate document-level text generation. Given that MetricX-23 with EWD shows a high correlation with human evaluation in Table 1, we foresee it to be a valid metric for the task. In addition to using MetricX as a utility function, we also evaluate using SentBERT, COMET-22, and SFR-2 (`Salesforce/SFR-Embedding-2_R`) as a reference. We use a bigger MetricX-23 model

[3] `https://github.com/Tiiiger/bert_score`



(a) WMT24 En-Ja



(b) WMT24 En-De

Figure 2: Evaluation of MBR-OT on document-level machine translation tasks. WD metric with MetricX-23 as a sentence-level utility function is used as the evaluation metric.

(`google/metricx-23-xxl-v2p0`) to alleviate the overfitting problem of MBR decoding (Kovacs et al., 2024). We additionally evaluate with COMET-22 (`Unbabel/wmt22-comet-da`; Rei et al. 2022) in Appendix A.

Comparison of LA, WD, and EWD. Table 2 shows the performance of MBR-OT using LA, WD, and EWD as the formulation of the optimal transport. We observe that MBR-WD$^{\epsilon}$ and MBR-WD outperform the baselines.

We observe the BLEU scores of En-Ja are low compared to the state-of-the-art models (Table 4). Llama-3 tends to generate a shorter summary of the original English document rather than a precise translation. The average length of the generated texts is 642.34 characters, whereas the reference texts average 860.61 characters. In several cases, the generated outputs are less than one-third the length of the reference translations. Table 5 summarizes the number of characters, sentences, and the average number of characters per sentences generated by MBR and MBR-OT.

BLEU	En-Ja	En-De
Beam	7.75	17.44
MBR	10.01	18.42
MBR-WD$^{\epsilon}$	9.66	18.26
MBR-WD$^{\epsilon}_{L}$	9.82	18.42

Table 4: BLEU scores of the MBR-WD$^{\epsilon}$ and MBR on WMT24 with 32 samples. `mecab-python3` library (Kudo et al., 2004) is used for tokenizing Japanese texts. Note that the lexical metric is shown to have little correlation with human evaluation. The values are reported for reference.

For the rest of the paper, we conduct experiments using EWD with $\epsilon = 0.10$, as its performance is only marginal below WD, and EWD is known to be robust to model noises.

Evaluation of MBR-OT. Figure 2 shows the performance of the algorithms on Llama-3.1 and Table 3 on LLMs specifically trained for the target languages. The result shows that the performance of MBR-OT is outperforming the baselines in both language pairs.

4.3 Document-Level Simplification

Document-level simplification task is a combination of document summarization and a text simplification task (Sun et al., 2021; Blinova et al., 2023). The goal is to generate a short and easily readable summary of the given document so that the information is accessible to children and those who are learning the language. We use the first 300 entries of the JADOS dataset for the task (Nagai et al., 2024). The dataset contains articles in Japanese and their summaries, which were manually written by native Japanese speakers. We use the Wikipedia subset of the JADOS dataset as it is open-sourced.

We generate 32 samples using Llama-3.1 as a language model. We use the SentBERT as a utility function. D-SARI (Sun et al., 2021; Blinova et al., 2023) is used as the document-level evaluation metric following Nagai et al. (2024).

Figure 3a shows the D-SARI scores of the decoding algorithms with varying number of samples. Overall, we observe MBR-OT outperforming the baselines. We additionally evaluate the readability of the texts using JReadability (Hasebe and Lee, 2015) in Table 6 which shows that the readability

Dataset	MBR			MBR-WD$^{\epsilon}$			MBR-WD$^{\epsilon}_{L}$		
	|C|	|S|	|C|/|S|	|C|	|S|	|C|/|S|	|C|	|S|	|C|/|S|
WMT24 En-Ja	713.6	28.4	25.2	633.9	29.4	21.6	625.0	29.6	21.1
WMT24 En-De	1435.0	12.3	116.4	1366.9	11.8	115.3	1366.0	11.6	117.4
JADOS	338.7	7.8	43.6	317.9	7.3	43.5	319.3	7.2	44.2
CNNDM	488.5	3.0	162.8	489.7	2.9	168.9	471.1	2.9	165.3
PP-cc12m	660.5	5.9	111.4	589.0	6.9	85.5	625.9	7.2	86.7
PP-commonpool	565.9	5.1	110.3	517.8	6.1	84.9	508.0	5.9	86.5
PP-redcaps	710.5	6.5	109.3	572.7	6.8	84.5	586.7	6.5	90.1

Table 5: The average number of characters (|C|), sentences (|S|), and the average number of characters per sentence (|C|/|S|).

	jReadability
Beam	4.46
MBR	3.30
MBR-WD$^{\epsilon}$	**3.25**
MBR-WD$^{\epsilon}_{L}$	**3.25**

Table 6: Evaluation of readability of the generated text using jReadability (Hasebe and Lee, 2015). The lower score shows better readability.

	CNNDM	JADOS
Beam	16.32	16.39
MBR	17.58	25.39
MBR-WD$^{\epsilon}$	18.23	26.63
MBR-WD$^{\epsilon}_{L}$	18.88	26.56

Table 7: ROUGE scores of the MBR-WD$^{\epsilon}$ and MBR on CNNDM and JADOS with 32 samples. Note that the lexical metric is shown to have little correlation with human evaluation. The values are reported for reference.

of MBR-OT is on par with MBR. Table 7 shows the ROUGE scores as a reference.

4.4 Document-Level Summarization

Although the proposed method is targeted for document-level text generation tasks, it is applicable to generating short paragraphs with a couple of sentences. We evaluate the performance of MBR-OT on the first 300 entries of the CNNDM dataset where the output is around 2 to 4 sentences. CNNDM is a summarization task where the goal is to generate a short summary of the given news article. We use Llama-3.1 as a language model and SentBERT as a utility function. The result shows that the present approach has a positive impact even when the number of segments in the output is relatively small.

	cc12m	commonpool	redcaps
Beam	28.26	24.77	27.57
MBR	28.38	25.53	29.27
MBR-WD$^{\epsilon}$	27.99	25.20	29.03
MBR-WD$^{\epsilon}_{L}$	28.28	24.33	28.70

Table 8: Evaluation of the MBR-WD$^{\epsilon}$ and MBR using METEOR with 32 samples. Note that the lexical metric is shown to have little correlation with human evaluation. The values are reported for reference.

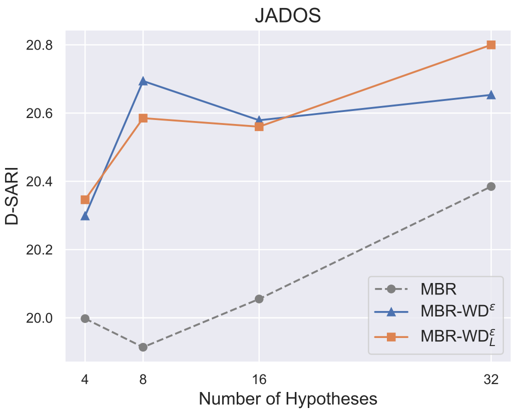
4.5 Dense Image Captioning

The task of dense image captioning is to generate a caption of an image containing types, attributes, and counts of objects in an image, in addition to spatial relations between objects, the presence of text, various broad image categorizations, etc (Johnson et al., 2016; Krishna et al., 2017; Urbanek et al., 2024; Singla et al., 2024).

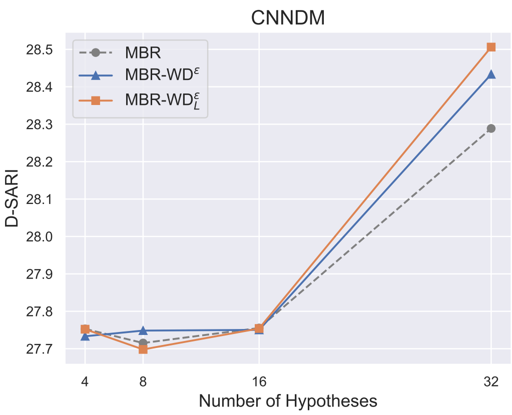
Unlike a traditional image captioning task, the caption wants to contain as much information about the image as possible.

We use the PixelProse dataset for the evaluation (Singla et al., 2024). PixelProse is a collection of images crawled web pages with curation to filter out harmful images and images that can violate privacy. We use 200 images from each of the three subsets of the dataset (cc12m, commonpool, and redcaps). We use PaliGemma-2 10B (Steiner et al., 2024b) fine-tuned on DOCCI dataset (Onoe et al., 2024) to generate 32 captions for each image (`google/paligemma2-10b-ft-docci-448`).

We use CLIPText as a utility function (`openai/clip-vit-large-patch14`; Qin et al., 2023). CLAIR (Chan et al., 2023a) is used as the

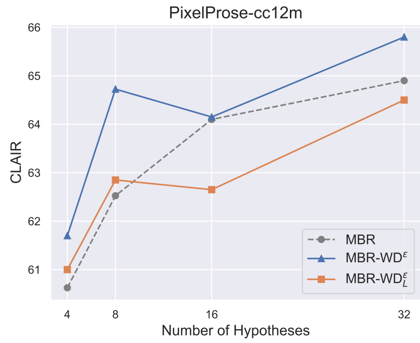


(a) JADOS

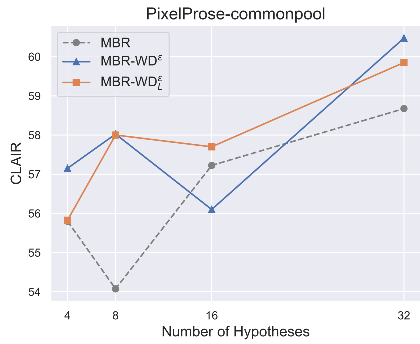


(b) CNNDM

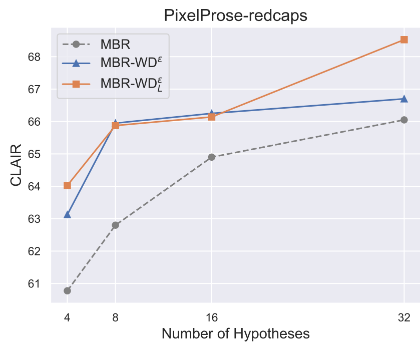
Figure 3: Evaluation of MBR-OT on document-level summarization and simplification tasks.



(a) PP-cc12m



(b) PP-commonpool



(c) PP-redcaps

Figure 4: Evaluation of MBR-OT on dense image captioning tasks.

metric to evaluate the outputs. CLAIR uses GPT-4 as a judge to compute the utility of the caption.

Figure 4 shows the performance of the algorithms. Overall, the performance of MBR-WD outperforms MBR decoding in all three subsets, showing that it consistently achieves high quality captioning in a wide range of images. We observe little differences between the algorithms on METEOR (Banerjee and Lavie, 2005) scores (Table 8).

5 Related Work

MBR decoding. The performance of MBR decoding is known to be dependent on the quality of the utility function (Fernandes et al., 2022; Freitag et al., 2022a; Kovacs et al., 2024). One of the problems of MBR decoding is its computational complexity. Several methods have been proposed to reduce the computational complexity of MBR decoding (Cheng and Vlachos, 2023; Jinnai and Ariu, 2024; Trabelsi et al., 2024), many of which are also applicable to MBR-OT.

The sampling strategy is an important aspect of MBR decoding. Epsilon sampling (Hewitt et al., 2022) is known to be an effective choice for MBR (Freitag et al., 2023a; Jinnai et al., 2024). We use epsilon sampling in this paper following their work.

Document-level text generation tasks. Improvement to the decoding algorithm is not the only solution to solve document-level text generation tasks. Hendy et al. (2023) show that few-shot learning is an effective strategy to generate high quality texts on document-level translation tasks using GPT. Briakou et al. (2024) propose a method to translate a document by decomposing the translation process into steps consisting of pre-translation research, drafting, refining, and proofreading. Jiang et al. (2022) proposes a metric specific to the evaluation of document-level machine translation tasks.

Blinova et al. (2023) proposes to use both the explicit summarization model and the simplification model to generate the final output. Cripwell et al. (2023) proposes a method to plan a simplification as a sequence of four simplification operations (copy, rephrase, split, or delete).

Because the objective of dense image captioning is to extract as much information from the given image, many of the studies have proposed methods to extract information from the image effectively. The importance of dense image captioning is recognized by the study by Krishna et al. (2017) is one of the first studies to show the use of rich annotation data for images. Hao et al. (2024) combines a detection model, a vision-language model, an OCR model, and other CV tools to generate dense image captions.

While these methods achieve domain-specific optimization to improve the performance of the given task, our contribution is to present a general-purpose decoding algorithm that is independent of the domain, except for its length.

6 Conclusions

We evaluate the performance of MBR-OT, a variant of MBR decoding using optimal transport distance to compute the document-level utility using a segment-level utility function. Compared to the prior work on document-level metric for machine translation (Vernikos et al., 2022), the advantage of MBR-OT is that it is naturally applicable to tasks where there are many options for disclosure structures in the output text. WD can adapt to the reordering, merging, and separation of disclosures, making it applicable to many tasks without engineering task-dependent optimization.

The empirical result shows that MBR-OT outperforms MBR decoding in document-level text generation tasks including machine translation, text simplification, and dense image captioning.

7 Limitations

MBR-OT requires additional computational overhead on MBR decoding which is already known to be computationally intensive. In our experiments using NVIDIA A100 GPUs with 80 GB VRAM, MBR-OT is roughly more than four times slower than MBR with the same number of samples. We foresee it to be used combined with efficient MBR methods (Cheng and Vlachos, 2023; Jinnai and Ariu, 2024) in practice. Also note that approximation algorithms to compute optimal transport distance faster have been proposed (Altschuler et al., 2017; Dvurechensky et al., 2018; Sommerfeld et al., 2019), which may enable us to compute MBR-OT faster.

Our study is limited to simple formulations of OT where the ordering of the segments are ignored. This may result in cases where the ordering of the segments matters. For example, the meaning of the pronoun is dependent on the ordering:

(1) Bob likes cats. Charles likes dogs. *He* always takes pictures of them.

(2) Charles likes dogs. Bob likes cats. *He* always takes pictures of them.

In the first text, “He” refers to Charles and “them” refers to dogs, but in the second text, they refers to Bob and cats. Therefore, the third sentence has semantically different meanings in the context. However, the formulations we present in the paper cannot distinguish such a difference. This suggests that the use of additional context proposed by Vernikos et al. (2022) may be also beneficial to the OT utility functions. Alternative approach is to use a more sophisticted OT formulation such as Fused Gromov-Wasserstein (Titouan et al., 2019) to take into account the structures of the segments.

We consider the weight of the segments to be uniform or propotional to its length (Eq. 9 and 10). Evaluation of other approaches such as using probability mass or entropy would be future work.

We focus on directed text generation tasks. Open-ended text generation tasks such as story generation is an interesting future work.

The evaluation is limited to moderately long documents with a couple of paragraphs. The evaluation of the method for much longer text generation tasks (Liang et al., 2023; Zhang et al., 2024; Hsieh et al., 2024) is future work. For generating much longer texts, we require the segmentation of paragraphs instead of sentences to align larger semantic blocks. How to build a hierarchy of segmented texts from a long text is not a trivial problem that we need to investigate for applying MBR-OT to these tasks.

As mentioned in Section 5, there are many methods proposed for each specific document-level text generation task. For example, several methods have been proposed for document-level machine translation (Voita et al., 2019; Kudo et al., 2024). Evaluation of MBR-OT combined with task-dependent optimizations is future work.

The study depends on automated metrics. Although the metrics we used in the study are known to have a high correlation with human evaluation, they are not flawless. Human evaluation is desirable for evaluation.

8 Impact Statements

This work was conducted using existing, publicly available, including WMT datasets, JADOS, PixelProse, and CNNDM. The datasets are constructed as benchmarks for research use in the research community (Table 14).

As for MBR, MBR-OT is not designed to prevent the system from generating toxic texts. Thus, it requires a countermeasure besides MBR-OT, such as language model alignment (Ouyang et al., 2022; Eisenstein et al., 2023; Rafailov et al., 2023).

Acknowledgments

We thank the anonymous reviewers for their insightful comments on the manuscript, which enabled us to conduct a more fine-grained analysis of the experiments. We would like to thank our colleagues at CyberAgent AI Lab, Kenshi Abe, Kaito Ariu, Tetsuro Morimura, Mitsuki Sakamoto, Yuma Fujimoto, and Ryota Mitsuhashi for their encouragement and for the helpful discussions during the early stages of this work.

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. 2017. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. 2023. Regularization for wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:33.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023a. CLAIR: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.

David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. 2023b. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Singapore. Association for Computational Linguistics.

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.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Petros Drineas and Michael W. Mahoney. 2005. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(72):2153–2175.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. 2018. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376. PMLR.

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.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*.

António Farinhas, José G. C. de Souza, and André F. T. Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. *arXiv*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José 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. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023a. 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. 2022a. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023b. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kamil Guttmann, Mikołaj Pokrywka, Adrian Charkiewicz, and Artur Nowakowski. 2024. Chasing COMET: Leveraging minimum Bayes risk decoding for self-improving machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 80–99, Sheffield, UK. European Association for Machine Translation (EAMT).

Jing Hao, Yuxiang Zhao, Song Chen, Yanpeng Sun, Qiang Chen, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. 2024. Fullanno: A data engine for enhancing image comprehension of mllms. *arXiv preprint arXiv:2409.13540*.

Yoichiro Hasebe and Jae-Ho Lee. 2015. Introducing a readability evaluation system for japanese language education. In *Proceedings of the 6th international conference on computer assisted systems for teaching & learning Japanese*, pages 19–22.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, volume 28.

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.

John T. Holodnak and Ilse C. F. Ipsen. 2015. Randomized approximation of the gram matrix: Exact computation and probabilistic bounds. *SIAM Journal on Matrix Analysis and Applications*, 36(1):110–137.

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*.

Eduard Hovy and Laurie Gerber. 1997. MT at the paragraph level: improving English synthesis in SYSTRAN. In *Proceedings of the 7th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, St John's College, Santa Fe. ?? Not mentionned on TOC.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Yuu Jinnai and Kaito Ariu. 2024. Hyperparameter-free approach for faster minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8547–8566, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. 2024. Model-based minimum Bayes risk decoding for text generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22326–22347. PMLR.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

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.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Xiaobo Liang, Zecheng Tang, Juntao Li, and Min Zhang. 2023. Open-ended long text generation via masked language modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 223–241, Toronto, Canada. Association for Computational Linguistics.

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

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476, Torino, Italia. ELRA and ICCL.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. Docci: Descriptions of connected and contrasting images. In *Computer Vision – ECCV 2024*, pages 291–309, Cham. Springer Nature Switzerland.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Gabriel Peyré and Marco Cuturi. 2020. Computational optimal transport. *arXiv preprint arXiv:1803.00567*.

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.

Libo Qin, Weiyun Wang, Qiguang Chen, and Wanxiang Che. 2023. CLIPText: A new paradigm for zero-shot text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1077–1088, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

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.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*.

Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. 2019. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024a. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024b. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. 2019. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. 2024. Efficient minimum bayes risk decoding using low-rank matrix completion algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26700–26709.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level mt metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cédric Villani. 2021. *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. 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 877–886, Hong Kong, China. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

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

Ian Wu, Patrick Fernandes, Amanda Bertsch, Seungone Kim, Sina Pakazad, and Graham Neubig. 2024. Better instruction-following through minimum bayes risk. *arXiv preprint arXiv:2410.02902*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

A Additional Evaluation using COMET-22

Using the same evaluation metric as the utility function that MBR decoding uses is known to cause a bias in the evaluation. Prior work has shown that the metric overestimates the performance of MBR decoding using the evaluation metric (Kovacs et al., 2024).

Thus, the evaluation on Section 4.2 may exhibit bias. To this end, we additionally evaluate the outputs using COMET-22 (`Unbabel/wmt22-comet-da`; Rei et al. 2022) as a verification of the evaluation.

Tables 9 and 10 show the comparison of MBR-OT methods using varying OT formulations. Overall, we observe that MBR-WD outperforms the baselines.

Figure 5 shows the performance of MBR-WD with varying numbers of hypotheses. Overall, the same qualitative results are observed as in Figure 2.

B Analysis of MBR-OT on WMT24 En-Ja

To further investigate the behavior of MBR-OT, we computed the average scores for each domain in the WMT24 En-Ja task (Table 11). Note that the number of documents is uneven (news = 17, speech = 111, literary = 8, and social = 34). The performance of MBR-WD (MetricX) is better than MBR (MetricX) in all four categories. Interestingly, in the social domain, where MBR drops the score compared to the other domains, MBR-WD achieves a score on par with the other domains. We speculate that this is because documents in social domains are sourced from SNS (Mastodon) (Kocmi et al., 2024) and are less structured than those in news domains sourced from online news sites. OT may not offer a strong advantage in the news domain as the document structure is relatively fixed.

Note that the analysis is based on a small number of samples, so further investigation is needed for more fine-grained analysis.

COMET-22	En-Ja	En-De
Beam	58.14	67.19
MBR (SFR2)	53.98	61.72
MBR (MetricX)	67.76	70.77
MBR-LA (MetricX)	64.12	72.56
MBR-LA$_L$ (MetricX)	64.56	73.17
MBR-WD (MetricX)	66.82	72.70
MBR-WD$_L$ (MetricX)	**70.46**	73.13
MBR-WD$^\epsilon$ (MetricX)	66.82	73.62
MBR-WD$^\epsilon_L$ (MetricX)	70.23	**73.86**

Table 9: Comparison of MBR-OT methods with LA, WD, and EWD using COMET-22. Llama-3.1 is used as the text generation model.

COMET-22	En-Ja CALM2-DPO	En-De EuroLLM
MBR	83.26	75.44
MBR-WD$^\epsilon$	83.06	76.27
MBR-WD$^\epsilon_L$	**83.55**	**76.35**

Table 10: COMET-22 scores of the MBR-OT and MBR on WMT24 with 32 samples on LLMs specifically trained for the languages.

C Generation Examples on WMT24 En-Ja

Below are the generation examples of the Llama-3.1 model by random sampling. The source English document is the following.

Source document (Abbreviated)

> Siso's depictions of land, water center new gallery exhibition "People Swimming in the Swimming Pool" from 2022 is one Vicente Siso artwork that will display at Tierra del Sol Gallery beginning Jan. 13. (photo courtesy of Vicente Siso) Tierra del Sol is pleased to present "Vicente Siso: Memories of the Land and Water" at the new gallery location in West Hollywood. Siso has been an artist in the Studio Arts Program since 2012, this marks his debut solo exhibi-

Category	Overall	News	Speech	Literary	Social
Beam	61.57	63.22	60.55	56.01	65.37
MBR (SentBERT)	56.26	72.46	54.38	52.51	60.35
MBR (COMET-22)	60.55	**77.06**	58.15	56.17	61.16
MBR (SFR-2)	57.38	70.40	54.35	61.23	59.86
MBR (MetricX)	68.81	73.32	71.14	68.48	59.03
MBR-LA (MetricX)	70.01	77.34	72.64	**73.62**	66.93
MBR-WD$^{\epsilon}$ (MetricX)	68.07	74.64	70.41	58.77	59.31
MBR-WD (MetricX)	**75.29**	77.00	**75.94**	74.96	**72.40**

Table 11: Evaluation of the MetricX scores on the WMT'2024 En-Ja. The number of documents is news=17, speech=111, literary=8, and social=34.

tion. ... "Vicente Siso: Memories of the Land and Water" opens on Saturday, Jan. 13, with a reception from 6-8 p.m. The exhibition is on view through Sunday, March 3. The Tierra del Sol Gallery is located at 7414 Santa Monica Blvd. For information, visit tierradelsolgallery.org.

Generation example 1

Sisoの作品、「2022年に製作された「Swimming Poolの人間たち」」が、2023年1月13日からTierra del Sol Galleryで展示されることになります。この展示は、西ハリウッドの新しいギャラリーで開催される「Vicente Siso:ランドと水の記憶」というタイトルのグループ展です。... 「Vicente Siso:ランドと水の記憶」は、2023年1月13日に開幕し、6-8 p.m.の受け入れを含み、3月3日まで展示します。ティエラデルソル・ギャラリーは、サンタモニカブラビュー7414号にあります。情報は、tierradelsolgallery.orgにアクセスしてください。

Generation example 2

2022年の作品「People Swimming in the Swimming Pool」が、西海岸の Tierra del Sol ギャラリーで展示されるVICENTE SISOの作品です。Tierra del Sol ギャラリーは、ロサンゼルス西部の新しいギャラリーで、VICENTE SISOの個展「Memories of the Land and Water」を開催します。

...

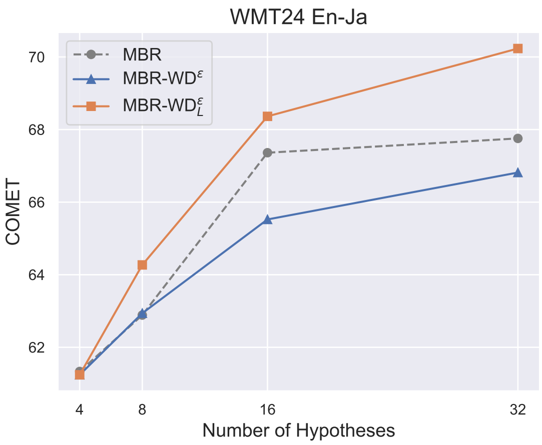
VICENTE SISOの個展「Memories of the Land and Water」は、1月13日の土曜日6-8pmに開催され、3月3日の日曜日まで展示されます。Tierra del Sol ギャラリーは、サンタモニカ・ブールバード7414番地に所在しており、tierradelsolgallery.orgで情報を閲覧できます。

Generation example 3

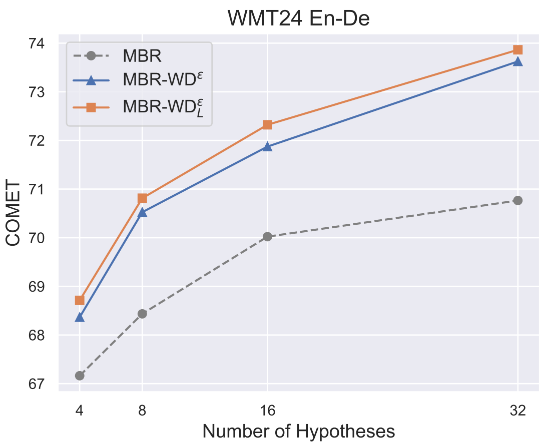
シソ・ソロ展「Vicente Siso：Land and Waterのメモリー」

新しいギャラリー、ティエラ・デル・ソル・ギャラリーにおいて、2022年の「People Swimming in the Swimming Pool」を始まり、1月13日から出展します。（写真：Vicente Siso） ... 「Vicente Siso：Land and Waterのメモリー」は、1月13日(土)6-8pmのオープン・リセプションに始まり、3月3日(日)まで開催されます。ティエラ・デル・ソル・ギャラリーは、サンタ・モニカ・ブールバード7414番地にあります。詳細はtierradelsolgallery.orgを参照してください。

The "。" character serves as the period in Japanese, so most texts are segmented based on this character. In the examples, the bolded text highlights segments where the structure differs from the rest of the generated outputs. In the second example, two sentences from the original English document are merged into a single sentence. In the



(a) WMT24 En-Ja



(b) WMT24 En-De

Figure 5: Evaluation of MBR-OT on document-level machine translation tasks using COMET-22. Llama-3.1 is used as the text generation model.

third example, the first sentence of the document appears to be interpreted as the title of the article.

Although the document-level MT is a directed text generation task, the generated texts from the LLM may have various structures.

D Prompts for Text Generation

We use the following prompt for the machine translation tasks in Section 4.2.

> Translate the following paragraph from English to [[LANGUAGE]].
>
> [[QUESTION]]
>
> Translate this paragraph from English to [[LANGUAGE]].

We use the following prompt for the JADOS dataset (Section 4.3).

> 以下のWikipediaの記事を、小学生でも分かるやさしい日本語で要約してください。
>
> 記事: [[QUESTION]]
> Q: この記事をやさしい日本語で要約してください。
> A:

The text below is the above prompt translated into English.

> Please summarize the following Wikipedia article in simple Japanese that even an elementary school student can understand.
>
> Article: [[QUESTION]]
> Q: Please summarize this article in simple Japanese.
> A:

For the dense image captioning (Section 4.5), we use the same prompt as Singla et al. 2024:

> Describe the image in detail. Please specify any objects within the image, backgrounds, scenery, interactions, and gestures or poses. If they are multiple of any object, please specify how many and where they are. If any text is present in the image, mention where it is, and the font. Describe the text in detail with quotation marks. For example, if the image has text, Merry Christmas, write it down as “Merry Christmas”. Describe the style of the image. If there are people or characters in the image, what emotions are they conveying? Identify the style of the image, and describe it as well. Please keep your descriptions factual and terse but complete. The description should be purely factual, with no subjective speculation. Make sure to include the style of the image, for example cartoon, photograph, 3d render etc.

E Hyperparameters

Table 12 shows the hyperparameters used for text generation. We use epsilon sampling (Hewitt et al., 2022) for all the experiments as it is shown to be effective for generating samples for MBR decoding (Freitag et al., 2023a; Jinnai et al., 2024).

Parameter	Value
Temperature	1.0
top_p	1.0
epsilon_cutoff	0.01
max_new_tokens	1024

Table 12: Hyperparmeters for text generation

Parameter	Value
Temperature	0.3
Version	2024-05-13

Table 13: Hyperparmeters for GPT-4

Table 13 shows the hyperparameters we use for GPT-4 evaluation and for CLAIR in Section 4.5.

F Reproducibility Statement

All datasets and models used in the experiments are publicly available. The code is implemented using Huggingface's Transformers library (Wolf et al., 2020). The computation of Wasserstein distance is implemented by POT: Python Optimal Transport library (Flamary et al., 2021). For Section 4.1, we use the codebase of Vernikos et al. (2022)[4] and use MT Metrics Eval V2[5] for the evaluation. Our code is available at `https://github.com/jinnaiyuu/mbr-optimal-transport`.

The experiments are conducted using NVIDIA A100 GPUs with 80 GB VRAM. The total amount of GPU time for the study is estimated to be 100-1000 GPU hours.

[4] `https://github.com/amazon-science/doc-mt-metrics`

[5] `https://github.com/google-research/mt-metrics-eval`

Table 14: List of datasets and models used in the experiments.

Name	Reference
WMT22 Metric Task	(Freitag et al., 2022b) https://www.statmt.org/wmt22/metrics/index.html
WMT23 Metric Task	(Freitag et al., 2023b) https://wmt-metrics-task.github.io/
WMT24 General Task	(Kocmi et al., 2024) https://www2.statmt.org/wmt24/translation-task.html
CNNDM	(Hermann et al., 2015) https://github.com/google-deepmind/rc-data
JADOS	(Nagai et al., 2024) https://github.com/tmu-nlp/JADOS
PixelProse	(Singla et al., 2024) https://huggingface.co/datasets/tomg-group-umd/pixelprose
Llama-3.1	(Dubey et al., 2024) https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
PaliGemma-2	(Steiner et al., 2024a) https://huggingface.co/google/paligemma2-10b-ft-docci-448
MetricX-23	(Juraska et al., 2023) https://huggingface.co/google/metricx-23-xxl-v2p0
MPNet	(Song et al., 2020) https://huggingface.co/sentence-transformers/all-mpnet-base-v2
CLIP	(Radford et al., 2021) https://huggingface.co/openai/clip-vit-large-patch14
COMET-22	(Rei et al., 2022) https://huggingface.co/Unbabel/wmt22-comet-da
D-SARI	(Sun et al., 2021) https://github.com/jinnaiyuu/mbr-optimal-transport Implemented by the authors.
JReadability	(Hasebe and Lee, 2015) https://github.com/joshdavham/jreadability
CLAIR	(Chan et al., 2023b) https://github.com/davidmchan/clair