Case-Based Decision-Theoretic Decoding with Quality Memories

Hiroyuki Deguchi and **Masaaki Nagata**
NTT, Inc.
{hiroyuki.deguchi,masaaki.nagata}@ntt.com

Abstract

Minimum Bayes risk (MBR) decoding is a decision rule of text generation, which selects the hypothesis that maximizes the expected utility and robustly generates higher-quality texts than maximum a posteriori (MAP) decoding. However, it depends on sample texts drawn from the text generation model; thus, it is difficult to find a hypothesis that correctly captures the knowledge or information of out-of-domain. To tackle this issue, we propose case-based decision-theoretic (CBDT) decoding, another method to estimate the expected utility using examples of domain data. CBDT decoding not only generates higher-quality texts than MAP decoding, but also the combination of MBR and CBDT decoding outperformed MBR decoding in seven domain De–En and Ja↔En translation tasks and image captioning tasks on MSCOCO and nocaps datasets.

1 Introduction

Minimum Bayes risk (MBR) decoding robustly generates high-quality texts compared with maximum a posteriori (MAP) decoding, i.e., one-best decoding using beam search (Kumar and Byrne, 2004; Eikema and Aziz, 2020, 2022; Freitag et al., 2022; Fernandes et al., 2022; Cheng and Vlachos, 2023; Deguchi et al., 2024b; Heineman et al., 2024; Lyu et al., 2025). The key concept of MBR decoding is maximizing the expected utility (EU) of choice from multiple output hypotheses, which is based on EU theory (EUT) in decision-making under uncertainty (von Neumann and Morgenstern, 1944). EUT is widely applied beyond natural language processing (NLP) to include microeconomics and speech recognition (Conte et al., 2011; Goel and Byrne, 2000; Raina and Gales, 2023). However, due to a limitation of EUT, if there is a lack of knowledge about the problem or task, it is difficult to accurately estimate the EU, and a hypothesis that reflects preferences cannot be chosen (Gilboa and Schmeidler, 1995). Thus, MBR decoding, based on EUT, makes it difficult to generate texts that reflect domain knowledge.

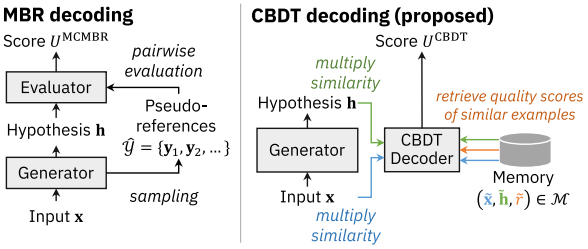

Figure 1: MBR decoding (left) and our proposed CBDT decoding (right). Both methods select the best hypothesis that maximizes the EU from a hypothesis set. "Generator" denotes a text generation model, "Evaluator" denotes a utility function, i.e., evaluation metric, and "CBDT Decoder" computes CBDT scores.

In response to such issues with EUT, in the field of decision theory, case-based decision theory (CBDT) has been proposed, which inductively derives the best action from past experiences (Gilboa and Schmeidler, 1995). CBDT evaluates the value of an action on the basis of analogies with similar cases experienced in the past.

To address the limitation of MBR decoding based on EUT, we propose case-based decision-theoretic decoding (CBDT decoding), a novel decision rule for high-quality text generation, which reflects domain-specific preferences by using domain data. As the preprocessing step, we pre-evaluate the quality scores of multiple generated texts and store them in a "memory". During generation, CBDT decoding calculates output scores from the memorized quality scores of similar examples by multiplying the similarity weights between the current problem and memorized examples. Figure 1 shows an overview of MBR decoding and CBDT decoding. MBR decoding evaluates the quality of a hypothesis using sampled texts called "pseudo-references" in decoding time, whereas CBDT decoding retrieves the precomputed quality scores of

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33691–33706

similar examples. Notably, CBDT decoding does not depend on "pseudo-references"; instead, it uses "true references" of similar examples. Moreover, we also propose MBR-CBDT decoding, a combination of MBR decoding and CBDT decoding with score normalization. Both MBR decoding and CBDT decoding estimate the expected quality but use different information orthogonally, i.e., possibilities and past experiences; thus, combining them further leads to even better quality.

Our CBDT decoding outperformed MAP decoding, and MBR-CBDT decoding generated higher-quality texts compared with MBR decoding in seven domain German–English and Japanese↔English translation tasks (Koehn and Knowles, 2017; Aharoni and Goldberg, 2020; Nakazawa et al., 2016; Neubig, 2011) and image captioning tasks on MSCOCO (Lin et al., 2014; Karpathy and Fei-Fei, 2015) and nocaps (Agrawal et al., 2019) datasets.

2 Background

MBR decoding Text generation is a fundamental NLP task that returns the best text given a context. This paper focuses on decision rules regarding the choice of outputs in conditional text generation.

Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ be an input and output text, respectively[1], where $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{V}^*$ denote the input and output spaces, respectively, and $\mathcal{V}^*$ denotes the Kleene closure of the vocabulary $\mathcal{V}$. The most widely used text generation method, MAP decoding, finds the most probable text:

$$\mathbf{y}^{\text{MAP}} := \operatorname*{argmax}_{\mathbf{h} \in \mathcal{Y}} p\left(\mathbf{h}|\mathbf{x}; \theta\right), \tag{1}$$

where θ is a text generation model. Because $\mathcal{Y}$ is an infinite set, the output is selected from a set of hypotheses $\mathcal{H} \subset \mathcal{Y}$ instead of $\mathcal{Y}$.

As a more quality-aware decision strategy[2], MBR decoding selects the hypothesis $\mathbf{y}^{\text{MBR}}$ that maximizes the expected utility (EU):

$$\mathbf{y}^{\text{MBR}} := \operatorname*{argmax}_{\mathbf{h} \in \mathcal{H}} U^{\text{MBR}}(\mathbf{h}; \mathbf{x}) \tag{2}$$

$$= \operatorname*{argmax}_{\mathbf{h} \in \mathcal{H}} \mathop{\mathbb{E}}_{\mathbf{y} \sim \Pr(\cdot|\mathbf{x})} \left[u(\mathbf{h}, \mathbf{y})\right]. \tag{3}$$

A reference text $\mathbf{y} \in \mathcal{Y}$ occurs according to the true output distribution $\Pr(\cdot|\mathbf{x})$, and $u\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denotes the utility function that satisfies $\mathbf{h} \succeq \mathbf{h}' \iff u(\mathbf{h}, \mathbf{r}) \geq u(\mathbf{h}', \mathbf{r})$, where $\succeq$ denotes the preference relation. For a utility function, an evaluation metric of output quality is often employed. Here, $\Pr(\cdot|\mathbf{x})$ is unknown; thus, the EU U^{MBR} is typically estimated using the Monte Carlo (MC) method (Eikema and Aziz, 2020, 2022):

$$U^{\text{MCMBR}}(\mathbf{h}; \hat{\mathcal{Y}}) := \frac{1}{|\hat{\mathcal{Y}}|} \sum_{\mathbf{y} \in \hat{\mathcal{Y}}} u(\mathbf{h}, \mathbf{y}), \tag{4}$$

where $\hat{\mathcal{Y}} := \{\mathbf{y}_i\}_{i=1}^{|\hat{\mathcal{Y}}|} \sim p(\mathbf{y}|\mathbf{x}; \theta)$ are pseudo-references, a multiset (a.k.a. bag) of sampled texts that are drawn from the text generation model θ.

The decision of MBR decoding highly depends on the distribution of pseudo-references (Ohashi et al., 2024; Kamigaito et al., 2025). Hence, in domains where the text generation model lacks knowledge, the EU estimation could be unreliable due to the discrepancy in distribution between the pseudo-references and true references, making it difficult to select the hypothesis that reflects domain-specific knowledge and information.

Case-based decision theory In decision theory, case-based decision theory (CBDT), which derives the best action from past experiences, has been proposed (Gilboa and Schmeidler, 1995). Decision-makers following CBDT choose actions on the basis of the rewards of similar experienced examples. In CBDT, the set of examples is defined by the triplet: a set of problems $\mathcal{Q}$, a set of actions $\mathcal{A}$, and the reward space $\mathcal{R}$. Let $q \in \mathcal{Q}$ denote the problem currently being faced. Decision-makers following CBDT choose an action $a^\star \in \mathcal{A}$ on the basis of the memory $\mathcal{M} \subseteq \mathcal{Q} \times \mathcal{A} \times \mathcal{R}$, a set of examples they have experienced in the past, as follows:

$$a^\star := \operatorname*{argmax}_{a \in \mathcal{A}} \sum_{(\tilde{q}, \tilde{a}, \tilde{r}) \in \mathcal{M}} s(q, \tilde{q}) \mathbb{1}_{a=\tilde{a}} \tilde{r}, \tag{5}$$

where $s\colon \mathcal{Q} \times \mathcal{Q} \to [0, 1]$ is the similarity between problems. From Equation (5), CBDT selects the action that maximizes the sum of the rewards weighted by the similarity between the current facing problem and experienced problems.

3 Proposed Method: CBDT Decoding

We propose *case-based decision-theoretic (CBDT) decoding* for high-quality text generation utilizing domain data. It pre-evaluates and stores the rewards of hypothesis selection (Figure 2) and decides the

[1]For simplicity, we formulate an input $\mathbf{x}$ as a text, but other modalities such as images can also be used for inputs.

[2]MAP decoding and MBR decoding are equivalent when u is an indicator function $\mathbb{1}_{\mathbf{h}=\mathbf{y}}$; thus, MAP decoding can be regarded as a special case of MBR decoding.

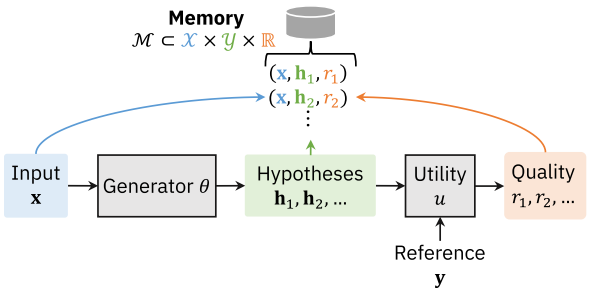

Figure 2: Memorization of CBDT decoding.

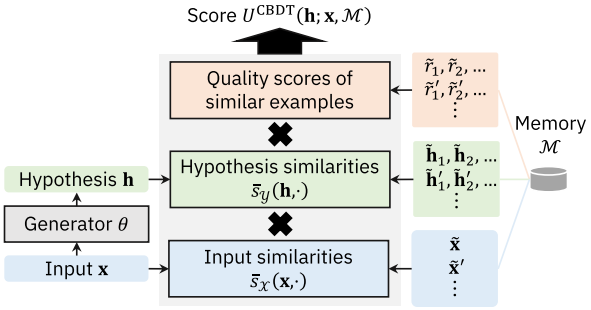

Figure 3: CBDT score calculation for a hypothesis $\mathbf{h} \in \mathcal{H}$. CBDT decoding selects the hypothesis that maximizes the score U^{CBDT} from hypotheses $\mathcal{H}$.

output referring to the memorized information (Figure 3). For the purpose of text generation, we hereafter redefine the problem set $\mathcal{Q}$ as the input space $\mathcal{X}$, action set $\mathcal{A}$ as the output space $\mathcal{Y}$, and reward space $\mathcal{R}$ as the quality scores of output texts calculated with the utility function u.

3.1 CBDT decoding

Memorization We first construct a *memory* from parallel data consisting of pairs of input and its reference output texts $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$. We generate sets of hypotheses $\mathcal{H}_{\mathbf{x}} \subset \mathcal{Y}$ with $H \in \mathbb{N}$ hypotheses for each input $\mathbf{x}$ in the parallel data $\mathcal{D}$.

$$\mathcal{H}_{\mathbf{x}} := \{\mathbf{h}_\ell\}_{\ell=1}^{H} \sim p(\cdot|\mathbf{x}; \theta). \tag{6}$$

Since $\mathcal{H}_{\mathbf{x}}$ is a set, $|\mathcal{H}_{\mathbf{x}}| \leq H$, i.e., the elements are deduplicated. Then, we evaluate each generated hypothesis using the reference text $\mathbf{y}$ and store the triplets in the memory $\mathcal{M} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ as follows:

$$\mathcal{M} := \{(\mathbf{x}, \mathbf{h}_i, r_i) \mid \mathbf{h}_i \in \mathcal{H}_{\mathbf{x}}, (\mathbf{x}, \mathbf{y}) \in \mathcal{D}\}, \tag{7}$$

$$\text{where } r_i = u(\mathbf{h}_i, \mathbf{y}). \tag{8}$$

Decoding CBDT decoding decides the output referring to the preconstructed memory. The naïve CBDT decoding based on Equation (5) selects the hypothesis that maximizes the following score:

$$U^{\text{CBDT}}_{\text{NAÏVE}}(\mathbf{h}; \mathbf{x}, \mathcal{M}) := \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{h}}, \tilde{r}) \in \mathcal{M}} s(\mathbf{x}, \tilde{\mathbf{x}}) \mathbb{1}_{\mathbf{h}=\tilde{\mathbf{h}}} \tilde{r}. \tag{9}$$

There are two problems with naïve CBDT decoding. One is that the similarity function s must return values within the range $[0, 1]$; thus, arbitrary similarity functions cannot be used. The other is due to the sparsity of natural language data. If the hypothesis $\mathbf{h}$ is not contained in $\mathcal{M}$, i.e., if the model does not generate exactly the same text as $\mathbf{h}$ in memorization, $U^{\text{CBDT}}_{\text{NAÏVE}}$ always returns 0 because of the indicator function $\mathbb{1}_{\mathbf{h}=\tilde{\mathbf{h}}}$. To solve these problems, we instead use the normalized similarity, and also introduce the similarity between hypotheses $\mathbf{h}$ and $\tilde{\mathbf{h}}$ instead of the indicator function to soften the equivalence checking.

$$U^{\text{CBDT}}(\mathbf{h}; \mathbf{x}, \mathcal{M}) := \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{h}}, \tilde{r}) \in \mathcal{M}} \bar{s}_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}; \mathcal{M}) \bar{s}_{\mathcal{Y}}(\mathbf{h}, \tilde{\mathbf{h}}; \mathcal{H}_{\tilde{\mathbf{x}}}) \tilde{r}. \tag{10}$$

$\bar{s}_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to [0, 1]$ and $\bar{s}_{\mathcal{Y}} \colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ are the normalized similarities for input and output sides, respectively, and we formulate them as:

$$\bar{s}_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}; \mathcal{M}) := \frac{\exp \frac{s_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}})}{\tau_{\mathcal{X}}}}{\sum_{(\tilde{\mathbf{x}}', \tilde{\mathbf{h}}', \tilde{r}') \in \mathcal{M}} \exp \frac{s_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}')}{\tau_{\mathcal{X}}}}, \tag{11}$$

$$\bar{s}_{\mathcal{Y}}(\mathbf{h}, \tilde{\mathbf{h}}; \mathcal{H}_{\tilde{\mathbf{x}}}) := \frac{\exp \frac{s_{\mathcal{Y}}(\mathbf{h}, \tilde{\mathbf{h}})}{\tau_{\mathcal{Y}}}}{\sum_{\tilde{\mathbf{h}}' \in \mathcal{H}_{\tilde{\mathbf{x}}}} \exp \frac{s_{\mathcal{Y}}(\mathbf{h}, \tilde{\mathbf{h}}')}{\tau_{\mathcal{Y}}}}, \tag{12}$$

where $s_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $s_{\mathcal{Y}} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are arbitrary similarity functions for the input space and output space, respectively, and $\tau_{\mathcal{X}}$ and $\tau_{\mathcal{Y}}$ are the temperatures for the similarities.

Our proposed *CBDT decoding* first retrieves the k-nearest neighbor examples $\hat{\mathcal{M}} \subseteq \mathcal{M}$ based on the input side similarity $s_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}})$ to reduce the space complexity[3], then selects the hypothesis that maximizes the score U^{CBDT}:

$$\mathbf{y}^{\text{CBDT}} := \operatorname*{argmax}_{\mathbf{h} \in \mathcal{H}} U^{\text{CBDT}}(\mathbf{h}; \mathbf{x}, \hat{\mathcal{M}}). \tag{13}$$

Note that $|\hat{\mathcal{M}}| \leq Hk$ since the memory has at most H hypotheses $\mathcal{H}_{\tilde{\mathbf{x}}}$ for each input text $\tilde{\mathbf{x}}$.

[3]This is because the similarity matrix between the current hypotheses $\mathcal{H}$ and memorized hypotheses $\{\tilde{\mathbf{h}} \mid (\tilde{\mathbf{x}}, \tilde{\mathbf{h}}, \tilde{r}) \in \mathcal{M}\}$ is often space-consuming. If all the memorized examples are used, the space complexity will be $\mathcal{O}(|\mathcal{H}||\mathcal{M}|) = \mathcal{O}(|\mathcal{H}||\mathcal{D}|H)$. For instance, if $|\mathcal{H}| = 1,024$, $|\mathcal{D}| = 100,000$, and $H = 256$, the size of the similarity matrix will be $1,024 \times 100,000 \times 256 \times 32$ bit $\simeq 97.7$ GiB.

3.2 Interpolation of MBR decoding and CBDT decoding

Both MBR and CBDT decoding aim to find the hypothesis that maximizes the utility, but they use orthogonal approaches to estimate the EU. Intuitively, MBR decoding mainly focuses on possibilities and CBDT decoding utilizes past experiences. Now, we propose *MBR-CBDT decoding*, which combines them, to further improve the quality:

$$\mathbf{y}^{\text{MCMBR}}_{\text{-CBDT}} := \operatorname*{argmax}_{\mathbf{h}\in\mathcal{H}} (1-\lambda)\bar{U}^{\text{MCMBR}}(\mathbf{h};\hat{\mathcal{Y}}) + \lambda\bar{U}^{\text{CBDT}}(\mathbf{h};\mathbf{x},\hat{\mathcal{M}}), \quad (14)$$

where $\lambda \in [0,1]$ balances two terms. To adjust the range of scores, $\bar{U}^{\text{MCMBR}}$ and $\bar{U}^{\text{CBDT}}$ normalize the scores of U^{MCMBR} and U^{CBDT}, respectively, by min-max normalization over the hypothesis set $\mathcal{H}$.

3.3 Fast similarity calculation

CBDT decoding calculates similarities between the current input/hypothesis and memorized inputs/hypotheses during decoding, respectively, which consumes time linearly proportional to the number of examples. Fortunately, many similarity functions, including BM25 (Jones et al., 2000) and contextualized text embeddings, can prevent slowdown in decoding speed by precomputing the statistical information or embeddings on the example side in advance. Note that CBDT decoding only uses neighboring examples as described in Equation (13); thus, it does not need to load the cache of all examples into the RAM.

4 Experiments

We evaluated our methods in translation and image captioning tasks. To compare the pure decoding performance, we evaluated the output quality of each decoding method without additional model training. Specifically, we compared MAP decoding (MAP), N-best reranking using a quality estimation model (QE), MBR decoding with MC estimation (MBR), CBDT decoding (CBDT), and MBR-CBDT decoding (MBR-CBDT) with $\lambda = 0.5$. We also compared them with oracle (ORACLE), which selects the hypotheses using the reference texts. For the translation tasks, we evaluated k-nearest neighbor machine translation (kNN-MT) (Khandelwal et al., 2021), a baseline of example-based decoding. We implemented our methods using `mbrs` (Deguchi et al., 2024a), and used `knn-transformers` (Alon et al., 2022) for kNN-MT.

4.1 Machine translation

Setup We conducted German–English (De–En) translation experiments in five domains: IT, Koran, law, medical, and subtitles (Koehn and Knowles, 2017; Aharoni and Goldberg, 2020), and Japanese↔English (Ja↔En) translation experiments in two domains: scientific paper (ASPEC) (Nakazawa et al., 2016) and Wikipedia's Kyoto articles (KFTT) (Neubig, 2011). We generated 1,024 translation hypotheses for each input using M2M100[4] (Fan et al., 2021) via epsilon sampling with $\varepsilon = 0.02$ (Freitag et al., 2023). We used the hypothesis set for the pseudo-references. In kNN-MT, we stored the input representations of the final feed-forward layer in the decoder, retrieved the top-64 nearest neighbor tokens using `Faiss` (Johnson et al., 2019; Douze et al., 2024), and interpolated the output probability with the temperature of 100.0 and $\lambda = 0.5$ (Khandelwal et al., 2021). For CBDT and MBR-CBDT, the memories were constructed from parallel data of each domain, and we generated $H = 256$ hypotheses for each example in De–En and $H = 64$ in Ja↔En. In ASPEC, we only used the top 1 million translation pairs from the training set to avoid noisy data, following Nakazawa et al. (2016). For the similarity functions of input and output texts, we used the cosine similarity of sentence embeddings of `intfloat/multingual-e5-large-instruct`[5] ($\text{mE5}^{\text{instruct}}_{\text{large}}$) (Wang et al., 2024). U^{CBDT} was calculated using examples that have the $k = 256$ nearest neighbor input texts, i.e., using $Hk = 65{,}536$ triplets at most in De–En. We set the temperature parameters to $\tau_{\mathcal{X}} = 0.01$ and $\tau_{\mathcal{Y}} = 0.01$ for De–En, and $\tau_{\mathcal{X}} = 0.1$ and $\tau_{\mathcal{Y}} = 0.01$ for Ja↔En, respectively[6]. We used CHRF (Popović, 2015) and COMET[7] (Rei et al., 2022a) for the utility function, and CometKiwi[8] (KIWI) (Rei et al., 2022b) for the QE model. We evaluated the translation quality on CHRF, COMET, and BLEURT (BLRT) (Sellam et al., 2020). We also measured the execution time on a 32-core Intel® Xeon® Gold 6426Y and a single NVIDIA RTX™ 6000 Ada. We calculated utility scores and similarity with a batch size of 256 sentences.

[4] `facebook/m2m100_418M`

[5] We used the instruction for the semantic textual similarity (STS) task as follows: "`Instruct: Retrieve semantically similar text.\nQuery: `"

[6] The details of tuning are described in Appendix A.

[7] `Unbabel/wmt22-comet-da`

[8] `Unbabel/wmt22-cometkiwi-da`

Decoding	IT			Koran			Law			Medical			Subtitles		
	CHRF	COMET	BLRT	CHRF	COMET	BLRT	CHRF	COMET	BLRT	CHRF	COMET	BLRT	CHRF	COMET	BLRT
MAP	45.5	76.1	58.3	23.7	57.9	46.1	48.5	74.6	61.8	50.7	78.0	61.8	39.8	73.5	55.5
QE	51.0	79.1	58.9	36.2	71.9	**50.4**	58.3	84.0	66.4	55.0	81.6	62.5	40.6	77.1	56.4
kNN-MT	50.2	79.8	**61.4**	28.0	63.7	46.2	58.6	82.5	66.7	57.3	81.7	**64.9**	42.1	74.7	56.1
Utility u: CHRF															
MBR	52.7	77.8	58.7	37.3	68.1	49.6	60.1	82.3	65.8	56.7	80.3	62.6	42.4	74.4	56.1
CBDT	51.7	77.1	57.7	34.0	63.0	46.3	58.9	80.4	63.9	56.2	78.4	60.1	39.3	71.7	53.0
MBR-CBDT	**54.6**	78.7	59.8	**37.5**	67.4	49.2	**61.6**	82.4	66.2	**58.4**	80.4	62.8	**43.3**	74.5	55.8
ORACLE	63.9	82.4	66.9	48.0	69.8	53.3	69.5	84.1	70.0	67.0	82.4	67.5	57.1	77.3	61.8
Utility u: COMET															
MBR	51.6	81.1	59.7	35.9	**73.5**	50.2	58.4	84.6	66.4	56.0	82.9	63.5	41.6	77.9	**56.6**
CBDT	50.6	79.7	59.3	33.6	68.1	48.0	57.0	82.7	64.7	54.0	81.1	60.9	38.3	74.9	54.0
MBR-CBDT	52.9	**82.2**	**61.4**	35.9	73.3	**50.4**	59.2	**85.0**	**67.2**	56.5	**83.3**	63.6	41.1	**78.2**	56.4
ORACLE	58.5	86.5	67.6	39.8	77.8	53.9	63.7	87.2	70.7	61.6	86.0	68.3	50.3	83.1	62.3

(a) The translation quality in the five domain De–En translation tasks.

Decoding	ASPEC						KFTT					
	Ja–En			En–Ja			Ja–En			En–Ja		
	CHRF	COMET	BLRT	CHRF	COMET	BLRT	CHRF	COMET	BLRT	CHRF	COMET	BLRT
MAP	34.9	65.8	49.5	14.0	69.6	44.9	13.6	40.0	31.6	8.4	57.7	31.9
QE	45.0	79.4	56.2	19.8	85.1	53.7	28.4	63.8	**40.1**	13.8	73.0	37.8
kNN-MT	42.8	75.0	53.2	19.9	81.8	52.4	22.0	56.1	36.3	12.5	69.5	37.0
Utility u: CHRF												
MBR	47.2	76.5	54.2	19.7	81.8	51.8	30.7	57.6	37.1	13.8	70.1	37.0
CBDT	46.3	76.4	53.9	20.2	83.0	52.9	30.3	59.3	37.4	14.0	70.0	36.9
MBR-CBDT	**48.3**	77.6	55.4	**20.8**	83.4	53.7	**31.9**	59.1	37.9	**14.5**	71.2	38.0
ORACLE	57.4	78.0	57.1	31.0	83.3	55.7	39.2	60.9	40.3	21.2	71.5	40.3
Utility u: COMET												
MBR	45.2	80.2	55.4	19.5	86.4	53.4	28.6	64.3	38.3	13.8	75.5	37.8
CBDT	44.6	78.2	55.0	19.1	84.2	53.4	28.3	63.2	38.3	13.6	73.0	38.5
MBR-CBDT	45.8	**80.4**	**56.3**	20.1	**86.5**	**54.7**	29.2	**66.0**	39.4	14.0	**76.0**	**39.2**
ORACLE	49.5	82.8	58.6	23.2	88.7	57.1	30.9	70.7	41.6	15.2	80.2	41.0

(b) The translation quality in the two domain Ja↔En translation tasks.

Table 1: Translation quality in the seven domain translation tasks. Green rows show the results of MBR-CBDT, and gray rows show the results of ORACLE. The **bold** and underlined scores indicate the best and second-best scores in each column except for ORACLE, respectively. All scores are shown as percentages (%).

Translation quality Table 1 demonstrates the results of the seven domain translation tasks in De–En and Ja↔En. In all translations, CBDT improved up to 16.7% +11.3% in CHRF and +23.2% in COMET compared with MAP. In addition, MBR-CBDT outperformed MBR in the given utility. Specifically, it improved up to +1.9% in CHRF and +1.7% in COMET. Moreover, it also achieved the best BLRT scores in six of the nine test sets, even though it was a non-target utility.

Decoding speed We measured the running time of hypothesis selection in the IT domain. Table 2 shows the statistics of running time per test case over 5 runs. CBDT took less than 0.2 second regardless of utilities, whereas MBR took 0.9 second when using COMET and more than 6.4 seconds when using CHRF. CBDT decoding is designed so that the decoding speed does not depend on the cost of the utility function. This is because MBR evaluates each hypothesis using multiple pseudo-references, i.e., it requires quadratic time, whereas CBDT does not call the utility function during decoding. One of the bottlenecks of CBDT is text encoding, but by leveraging the proposed method described in Section 3.3, only the query, consisting of a single input and $|\mathcal{H}|$ hypotheses, needs to be encoded during decoding. Thus, the computational cost remains relatively low. The running time of MBR-CBDT is the sum of that of MBR and a small utility-independent overhead incurred by CBDT.

Decoding	Avg	SD	Min	Max
QE	363.5	± 0.4	362.9	364.0
kNN-MT	2972.9	± 5.8	2965.4	2981.3
Utility u: CHRF				
MBR	6400.3	± 104.7	6192.1	6463.8
CBDT	120.2	± 0.8	119.0	121.1
MBR-CBDT	6521.6	± 105.3	6312.2	6585.2
Utility u: COMET				
MBR	899.8	± 1.2	898.8	902.0
CBDT	158.6	± 3.8	155.6	165.9
MBR-CBDT	1087.1	± 3.8	1084.2	1094.5

Table 2: Running times per test case (msec) over 5 runs. Columns "Avg", "SD", "Min", and "Max" indicate average, standard deviation, minimum, and maximum running times, respectively. CBDT and MBR-CBDT include times of encoding texts and calculating similarity.

	MSCOCO			nocaps		
Decoding	BLEU	CHRF	BS	BLEU	CHRF	BS
MAP	17.5	35.1	64.5	6.2	20.4	53.6
MBR	25.6	43.2	68.4	26.3	40.6	65.8
CBDT	19.5	41.4	66.3	20.1	37.3	63.5
MBR-CBDT	**26.0**	**44.4**	**68.6**	**26.8**	**41.6**	**66.3**
ORACLE	39.4	53.2	72.4	39.2	50.8	70.3

Table 3: Results of image captioning tasks on the MSCOCO and nocaps datasets.

4.2 Image captioning

To evaluate the effectiveness of our methods in multimodal tasks, we experimented in the image captioning task. In this task, the input is an image instead of text. Note that our methods do not require cross-modal embeddings for similarity calculation because the similarities of the input and output sides are calculated independently.

Setup We evaluated our methods on the MSCOCO (Lin et al., 2014; Karpathy and Fei-Fei, 2015) and nocaps (Agrawal et al., 2019) datasets. We generated 256 captions per image using BLIP-2[9] (Li et al., 2023) with epsilon sampling ($\varepsilon = 0.02$) for both memory construction and decoding. We used BERTScore[10] (BS) (Zhang* et al., 2020; He et al., 2023) for the utility function and evaluated captions on BLEU (Papineni et al., 2002), CHRF, and BS. In CBDT decoding, we employed DINOv2[11] (Oquab et al., 2024) for the image similarity and $\text{mE5}^{\text{instruct}}_{\text{large}}$ for the caption text similarity. We set $k = 256$, $\tau_{\mathcal{X}} = 0.1$, and $\tau_{\mathcal{Y}} = 1.0$, respectively. In memory construction, we used the training set of MSCOCO for its evaluation, and Localized Narratives (LN) (Pont-Tuset et al., 2020) for the nocaps evaluation[12].

[9] Salesforce/blip2-flan-t5-xl

[10] microsoft/deberta-v3-large

[11] facebook/dinov2-large

Input	Wie wirkt Intelence ?
Reference	How does Intelence work?
kNN-MT	How does intelligence work?
MBR	How does intelligence work?
MBR-CBDT	How does Intelence work?

Table 4: Translation examples in the medical domain. Both MBR and MBR-CBDT used CHRF for the utility function. Highlighted spans indicate the difference between translations. Other examples are shown in Appendix C.2.

Results Table 3 demonstrates the results of image captioning tasks. MBR-CBDT outperformed MBR in BLEU, CHRF, and BS on both MSCOCO and nocaps datasets, though we used BS for the utility function. These results suggest that, in addition to the choice of utility function, the domain data used for memory construction is also important. We further discuss the limitations and additional show experimental results when constructing the memory with out-of-domain data in Appendix C.1. From the experiments, we confirmed that our methods are also effective in multimodal generation.

5 Discussion

5.1 Case study: Medical translation

We list the examples of medical translation in Table 4. MBR-CBDT correctly retained the medication name "*Intelence*", but MBR mistranslated it to "*intelligence*". Accurate translation in the medical domain, like this example, is crucial for preventing serious incidents that may threaten patient safety.

We also investigated our used memory, and found that there were 37 examples that contain "*Intelence*" on both the input and output sides in the memory, respectively. In contrast, there were no examples in which "Intelence" is translated as "intelligence". Thus, when "*Intelence*" is given, the CBDT scores of hypotheses that retain "*Intelence*" are likely to be high. To summarize, we confirmed that MBR-CBDT determines the output by utilizing the information of similar examples in the memory.

[12] The nocaps dataset does not contain a training set, and it was created from the Open Images (Krasin et al., 2017). The LN (Pont-Tuset et al., 2020) was also created from it; thus, we used this dataset for memory construction.

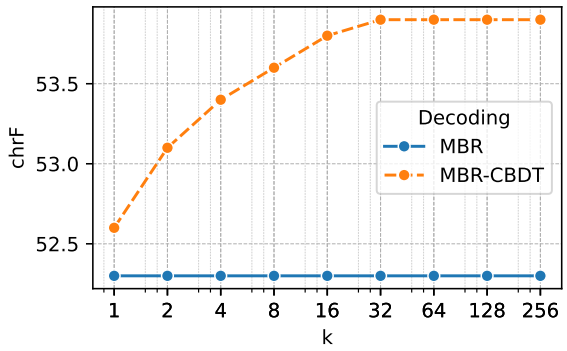

Figure 4: Translation quality (CHRF%) when the number of retrieved similar examples k was varied in the development set of the IT domain.

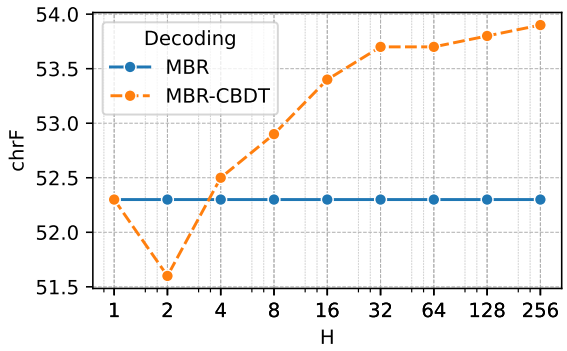

Figure 5: Translation quality (CHRF%) when the number of memorized hypotheses per input H was varied in the development set of the IT domain.

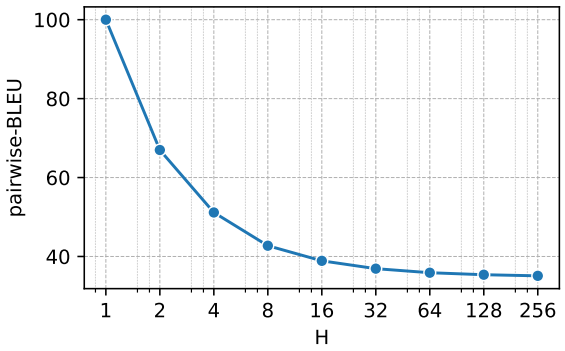

Figure 6: Relationship between the number of memorized hypotheses per input H and the diversity of $\mathcal{H}_{\mathbf{x}}$. A lower pairwise BLEU score means greater diversity.

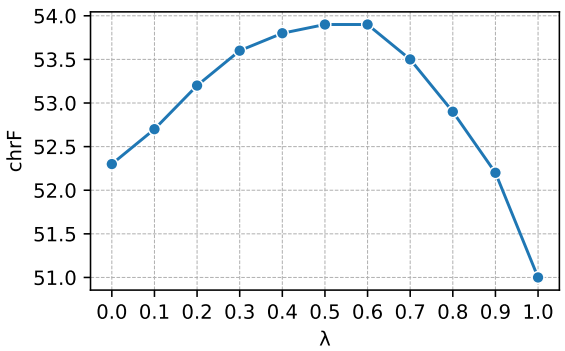

Figure 7: CHRF% scores of MBR-CBDT when varying the balancing weight $\lambda \in [0, 1]$ in the development set of the IT domain.

5.2 Relationship between number of examples and output quality

We investigated the relationship between the number of examples and output quality. Figure 4 and Figure 5 show the translation quality (CHRF%) when the number of retrieved similar examples k and number of memorized hypotheses per source H were varied, respectively. As k increased, memory usage increased in decoding time, but quality also improved because more examples were referenced. H also affected performance, but the quality improvement for $H \geq 16$ was less than $H < 16$.

We also analyzed the diversity in $\mathcal{H}_{\mathbf{h}}$ by evaluating the averaged pairwise BLEU (Shen et al., 2019) of $\mathcal{H}_{\mathbf{x}}$ on the memory in the IT domain. Figure 6 shows the averaged scores of 1,000 random samples. A lower score means a greater diversity of $\mathcal{H}_{\mathbf{x}}$. The results indicate that as H increased, $\mathcal{H}_{\mathbf{x}}$ became more diverse, but when $H \geq 16$, the pairwise BLEU converged. These results are similar to the relationship between the number of pseudo-references and output quality in MBR decoding, i.e., a larger pseudo-reference set estimates the EU stably and converges the quality (Eikema and Aziz, 2022; Kamigaito et al., 2025; Ichihara et al., 2025).

5.3 Balancing MBR and CBDT decoding

We tuned the balancing weight $\lambda \in [0, 1]$. Figure 7 shows the CHRF% scores of MBR-CBDT with the various $\lambda \in \{0.0, 0.1, \ldots, 1.0\}$ in the development set of the IT domain translation. Note that we normalized the MBR and CBDT scores by the min-max normalization, as described in Equation (14). As shown in the figure, by adjusting the weight λ, the quality can be maximized, and we selected $\lambda = 0.5$ for all experiments.

5.4 Combination of CBDT and approximated MBR decoding

MBR-CBDT takes at least as long as MBR to execute in principle, and running time is dominated by the MBR term. Recently, approximated methods for faster MBR decoding have been proposed (DeNero et al., 2009; Vamvas and Sennrich, 2024; Deguchi et al., 2024b; Cheng and Vlachos, 2023; Jinnai and Ariu, 2024; Trabelsi et al., 2024),

Decoding	CHRF	COMET	BLRT	wall-time
MBR	52.7	77.8	58.7	6400.3
MBR-CBDT	**54.6**	78.7	59.8	6521.6
PMBR-CBDT	53.2	**79.0**	**60.2**	1926.5

Table 5: Results of the combination of CBDT and approximated MBR decoding. "wall-time" indicates wall-clock time of averaged running time per test case (msec) over 5 runs.

Similarity	CHRF	COMET	BLRT
$mE5^{instruct}_{large}$	**54.6**	**78.7**	**59.8**
LaBSE	54.0	78.5	59.1
BM25	51.2	76.5	57.4

Table 6: Comparisons of similarity functions in the IT domain translation.

and we investigate the effectiveness of combining CBDT with them. Table 5 shows the results of the combination of CBDT decoding and probabilistic MBR decoding (PMBR) (Trabelsi et al., 2024), which reduces the number of utility function calls. In PMBR, we reduced the number of utility function calls by a factor of 64 and decomposed the pairwise score matrix with 8-dimensional two low-rank matrices. The results indicate that PMBR-CBDT ran 3 times faster than MBR and MBR-CBDT, yet it still outperformed MBR in all evaluation metrics. Interestingly, PMBR-CBDT outperformed MBR-CBDT in COMET and BLRT with a cheaper computational cost. This may have been owing to the mitigation of overfitting with PMBR, but further analyses remain for future work.

5.5 Choice of similarity functions

To clarify how the choice of similarity function affects the quality of output texts, we compared similarity functions and the translation quality in the IT domain. Table 6 compares similarity functions: the cosine similarity of $mE5^{instruct}_{large}$ and LaBSE (Feng et al., 2022), and BM25 (Jones et al., 2000) implemented by `BM25S` (Lù, 2024). We observed that the cosine similarity of contextualized embeddings achieved higher-quality than the lexical similarity, BM25. Furthermore, $mE5^{instruct}_{large}$, which has strong correlations with human assessment on the semantic textual similarity (STS) task (Muennighoff et al., 2023), outperformed LaBSE. These experiments suggest that a better similarity function yields higher-quality texts.

6 Related Work

Reranking To enhance output quality, various reranking methods have been proposed. One involves using different probability distributions from the generation probability. Liu et al. (2016) and Imamura and Sumita (2017) used right-to-left generation models to rescore the hypotheses generated with left-to-right generation models. Another method uses backward probabilities, known as noisy channel decoding (NCD). NCD first appeared in statistical machine translation (Brown et al., 1990; Koehn et al., 2003), and its effectiveness has also been demonstrated in reranking of neural text generation (Yu et al., 2017; Yee et al., 2019; Ng et al., 2019). All of the above are "generative reranking" based on the likelihood calculated using text generation models. In contrast, discriminative reranking (Shen et al., 2004; Lee et al., 2021) directly distinguishes between good and bad texts and optimizes to rank the hypotheses. The advantage of the discriminative approach is that it directly maximizes the evaluation metrics. This means that developing metrics that accurately estimate quality will directly lead to improvements in text generation. Quality estimation (QE) (Kim and Lee, 2016; Kim et al., 2017; Zheng et al., 2021) evaluates the quality of generated texts without the reference texts. Recent QE models employ large pre-trained encoder models or large language models and achieved high correlation with human assessments (Rei et al., 2022b; Guerreiro et al., 2024; Juraska et al., 2024; Li et al., 2024).

MBR decoding While most reranking methods score multiple hypotheses independently, MBR decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004; Eikema and Aziz, 2020, 2022; Freitag et al., 2022; Fernandes et al., 2022; Lyu et al., 2025) calculates the expected utility using a set of candidates. Typically, pseudo-references sampled from the text generation model are used as a proxy for the reference texts, and their distribution affects the output quality (Ohashi et al., 2024; Kamigaito et al., 2025). Daheim et al. (2025) improved the robustness of MBR decoding by using multiple generator models. CBDT and MBR-CBDT decoding use true references in example data, not pseudo-references.

One of the major challenges of MBR decoding is that it takes quadratic time proportional to the number of candidates during inference. To tackle this issue, efficient variants of MBR decoding have been proposed; however, they still have limitations in

the evaluation metrics that can be applied (DeNero et al., 2009; Vamvas and Sennrich, 2024; Deguchi et al., 2024b), or cannot avoid on-the-fly evaluation even if the utility calling is reduced (Cheng and Vlachos, 2023; Jinnai and Ariu, 2024; Trabelsi et al., 2024). CBDT decoding does not calculate the utility during decoding, making it faster to select the hypothesis regardless of a utility function.

Example-based generation Example-based generation is particularly useful for domain adaptation, especially in scenarios where example databases are hot-swapped. Like other methods, CBDT decoding generates texts that reflect domain-specific information without additional training. The idea of example-based generation originated from analogy-based machine translation (Nagao, 1984). Gu et al. (2018); Zhang et al. (2018) incorporated the information of similar examples retrieved from bilingual translation memories into neural machine translation models. Non-parametric domain adaptation using monolingual translation memory has also been proposed (Cai et al., 2021). Neural fuzzy repair (Bulte and Tezcan, 2019; Xu et al., 2022; Nieminen et al., 2025) and retrieve-edit-rerank (Hossain et al., 2020) retrieve similar translations from translation memories. These methods augment the input sequence with similar examples, while CBDT decoding does not. Thus, CBDT decoding can be applied to tasks where similar examples cannot be directly concatenated with an input sequence, such as an image captioning task. kNN-MT (Khandelwal et al., 2021) retrieves translation examples at the token level, and interpolates the output probability based on the distance between the current hidden representation and its nearest neighbors. CBDT decoding mainly differs from kNN-MT in three key respects: the search unit, i.e., token level for kNN-MT and text level for CBDT decoding, the use of input side similarity, and the incorporation of utility functions.

Pointwise Hilbert–Schmidt independence criterion Yokoi et al. (2018) proposed pointwise Hilbert–Schmidt independence criterion (PHSIC), a kernel-based co-occurrence measure. PHSIC relaxes the indicator functions on the input and output sides in pointwise mutual information (PMI) by using kernel functions. It has been shown to be effective for selecting data from parallel corpora in translation tasks (Yokoi et al., 2018; Kiyono et al., 2020). Our method multiplies the input and output side similarities to soften the indicator function in Equation (9), and measures the value of each example in the memory.

PHSIC and CBDT decoding are mathematically similar but differ in several key respects. Specifically, PHSIC is estimated as follows:

$$\begin{aligned}&\mathrm{PHSIC}(\mathbf{x}, \mathbf{y}; \mathcal{D}) \\ &\coloneqq \frac{1}{|\mathcal{D}|} \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in \mathcal{D}} \bar{K}_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}) \bar{K}_{\mathcal{Y}}(\mathbf{y}, \tilde{\mathbf{y}}), \end{aligned} \tag{15}$$

where $\bar{K}_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\bar{K}_{\mathcal{Y}} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are centered kernel functions for the input and output spaces, respectively. This closely resembles U^{CBDT} in Equation (5), but there are crucial differences. For each input $\tilde{\mathbf{x}}$, PHSIC estimates relying solely on observed data $\tilde{\mathbf{y}}$, whereas CBDT decoding uses multiple hypotheses $\tilde{\mathbf{h}} \in \mathcal{H}_{\tilde{\mathbf{x}}}$ along with their utilities $\tilde{r}$, i.e., our U^{CBDT} explicitly incorporates uncertainty regarding $\tilde{\mathbf{y}}$ through $\mathcal{H}_{\tilde{\mathbf{x}}}$ and the utility $\tilde{r} = u(\tilde{\mathbf{h}}, \tilde{\mathbf{y}})$. In fact, when $\mathcal{H}_{\tilde{\mathbf{x}}}$ contains the reference output $\tilde{\mathbf{y}}$ and the utility function u is defined as the exact match 0–1 loss $u(\mathbf{h}, \mathbf{y}) \coloneqq \mathbb{1}_{\mathbf{h}=\mathbf{y}}$, the two approaches essentially become equivalent.

7 Conclusion

We propose CBDT decoding, which selects a high-quality hypothesis based on rewards experienced in the past to improve the quality of text generation. The proposed method stores the rewards of hypothesis selection during memorization and uses them with similarity weights during decoding. We further improve output quality by combining MBR and CBDT decoding. CBDT decoding achieved better performance compared with MAP decoding, and MBR-CBDT decoding outperformed MBR decoding by up to 1.9% in CHRF and 1.7% in COMET in the seven domain De–En and Ja↔En translation tasks and the image captioning tasks on the MSCOCO and nocaps datasets. We plan to investigate the effectiveness of our method in generation tasks other than text generation.

Limitations

Biases from memorized data Our method also relies on utility functions, as does MBR decoding. This means that the generation texts are affected by biases of utility functions. Unlike MBR decoding, CBDT decoding is an example-based method; thus, it is susceptible to biases from memorized data. Note that CBDT decoding does not require any additional training, so we can switch on-the-fly to

use it or not. In addition, by constructing memories for each target domain, it can hot-swap the memory according to user requests. That is, by inserting and/or deleting examples online, such biases can be dynamically controlled.

Domain of memorized data CBDT decoding works with domain data, but when there is no domain data, it may degrade. Specifically, we observed that the output quality became lower when constructing memory from out-of-domain data, as mentioned in Appendix C.1.

Computational cost When the memorized examples are fixed, CBDT decoding works in constant time regardless of the choice of utility functions, since the utility scores are precomputed. This efficiency comes at the cost of storage. Specifically, the memory $\mathcal{M}$ and intermediate representation cache of similarity functions, described in Section 3.3, often consume large space. This paper focused on a new paradigm for decision rules of decoding and validating its effectiveness. Optimizing the memory and intermediate representation cache of the similarity is beyond the scope of this work, but we plan to address this limitation as future work.

Ethical Considerations

Potential risks As mentioned in the "Limitations" section, CBDT decoding is susceptible to biases from memorized data. If harmful examples are included in the memorized data and they receive high scores, harmful text is more likely to be generated. Conversely, if we use a utility function that evaluates safe and debiased examples with high scores, CBDT decoding may suppress the output of harmful cases. Therefore, to use CBDT decoding safely, we should develop non-toxic datasets used for the memories and utility functions that evaluate them with high scores.

Use of benchmark datasets In our experiments, we only used publicly available benchmark datasets as described in Section 4. All datasets we used were created on domains that do not contain personal information or offensive content. They are released under the licenses described in Appendix D, and we complied with them.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning*, pages 468–485. PMLR.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. 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 7307–7318, Online. 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.

Anna Conte, John D Hey, and Peter G Moffatt. 2011. Mixture models of choice under risk. *Journal of Econometrics*, 162(1):79–88.

Nico Daheim, Clara Meister, Thomas Möllenhoff, and Iryna Gurevych. 2025. Uncertainty-aware decoding with minimum bayes risk. In *The Thirteenth International Conference on Learning Representations*.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024a. mbrs: A library for minimum Bayes risk decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama.

2024b. Centroid-based efficient minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11009–11018, Bangkok, Thailand. Association for Computational Linguistics.

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 567–575, Suntec, Singapore. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint*.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. 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.

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

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

Itzhak Gilboa and David Schmeidler. 1995. Case-based decision theory. *The Quarterly Journal of Economics*, 110(3):605–639.

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

J Gu, Y Wang, K Cho, and V O K Li. 2018. Search engine guided neural machine translation. *AAAI*.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum Bayes risk decoding with multi-prompt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545, Miami, Florida, USA. Association for Computational Linguistics.

Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. 2020. Simple and effective retrieve-edit-rerank text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2532–2538, Online. Association for Computational Linguistics.

Yuki Ichihara, Yuu Jinnai, Kaito Ariu, Tetsuro Morimura, and Eiji Uchibe. 2025. Theoretical guarantees for minimum Bayes risk decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16262–16284, Vienna, Austria. Association for Computational Linguistics.

Kenji Imamura and Eiichiro Sumita. 2017. Ensemble and reranking: Using multiple models in the NICT-2 neural machine translation system at WAT2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 127–134, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, and Taro Watanabe. 2025. Diversity explains inference scaling laws: Through a case study of minimum Bayes risk decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29060–29094, Vienna, Austria. Association for Computational Linguistics.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–498, San Diego, California. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-AIP-NTT at WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 145–155, Online. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, and Victor Gomes. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification.

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.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for 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 7250–7264, Online. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 411–416, San Diego, California. Association for Computational Linguistics.

Boxuan Lyu, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2025. Unveiling the power of source: Source-based minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2976–2994, Vienna, Austria. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

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.

Tommi Nieminen, Jörg Tiedemann, and Sami Virpioja. 2025. Incorporating target fuzzy matches into neural fuzzy repair. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 408–418, Tallinn, Estonia. University of Tartu Library.

Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. 2024. On the true distribution approximation of 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 459–468, Mexico City, Mexico. Association for Computational Linguistics.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Featured Certification.

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

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Vyas Raina and Mark Gales. 2023. Minimum Bayes' risk decoding for system combination of grammatical error correction systems. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 105–112, Nusa Dua, Bali. Association for Computational Linguistics.

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for 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 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

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*.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

John von Neumann and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Jitao Xu, Josep Crego, and Jean Senellart. 2022. Boosting neural machine translation with similar translations. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 282–292, Orlando, USA. Association for Machine Translation in the Americas.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling 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 5696–5701, Hong Kong, China. Association for Computational Linguistics.

Sho Yokoi, Sosuke Kobayashi, Kenji Fukumizu, Jun Suzuki, and Kentaro Inui. 2018. Pointwise HSIC: A linear-time kernelized co-occurrence norm for sparse linguistic expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1763–1775, Brussels, Belgium. Association for Computational Linguistics.

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. 2017. The neural noisy channel. In *International Conference on Learning Representations*.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. 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*.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Hyperparameters

		$\tau_{\mathcal{X}}$		
Development set	$\tau_{\mathcal{Y}}$	0.01	0.1	1.0
IT De–En	0.01	**53.93**	53.76	53.72
	0.1	53.69	53.62	53.60
	1.0	53.70	53.60	53.57
ASPEC Ja–En	0.01	48.03	**48.13**	48.09
	0.1	47.96	48.02	47.97
	1.0	47.94	47.95	47.93

Table 7: CHRF scores when temperatures varied on the development set in the De–En IT domain and Ja–En ASPEC translation tasks.

		$\tau_{\mathcal{X}}$		
Development set	$\tau_{\mathcal{Y}}$	0.01	0.1	1.0
MSCOCO	0.01	67.93	68.54	67.93
	0.1	68.38	68.49	68.31
	1.0	68.25	**68.55**	68.53

Table 8: BS scores when temperatures varied on the development set in the MSCOCO image captioning task.

CBDT decoding has four hyperparameters: number of memorized hypotheses per input H, number of neighboring examples k, and temperatures of similarities in the input side $\tau_{\mathcal{X}}$ and output side $\tau_{\mathcal{Y}}$. As described in Section 5.2, H and k are trade-off parameters between memory usage and quality. The other two, the temperatures of similarities $\tau_{\mathcal{X}}$ and $\tau_{\mathcal{Y}}$, need to be tuned.

In our experiments, we tuned them from $\{0.01, 0.1, 1.0\}$ on the development set. Intuitively, lower temperatures emphasize the similarity scores. In the De–En domain translation tasks, we tuned them on the development set of the IT domain so

that CHRF is maximized. Likewise, in the Ja↔En domain translation tasks, we maximized CHRF on the development set of the ASPEC Ja→En. In the image captioning tasks, we tuned them on the development set of the MSCOCO dataset by maximizing BERTScore (BS). The results of development sets when the temperatures varied in the translation and image captioning tasks are shown in Table 7 and Table 8, respectively.

B Dataset Statistics

Dataset	Train	Dev	Test
De–En domain translation tasks			
IT	222,927	2,000	2,000
Koran	17,982	2,000	2,000
Law	467,309	2,000	2,000
Medical	248,099	2,000	2,000
Subtitles	500,000	2,000	2,000
Ja↔En domain translation tasks			
ASPEC	1,000,000	1,790	1,812
KFTT	440,288	1,166	1,160
Image captioning tasks			
MSCOCO	113,287	5,000	5,000
nocaps	[13]504,413	4,500	[14](10,600)

Table 9: Number of examples for each dataset. Note that we used training sets for memory construction.

Table 9 shows the number of examples for each dataset.

C Further Analyses

C.1 Relationship between memory domain and output quality

Memorized dataset	BLEU	CHRF	BS
LN (target domain)	26.8	41.6	66.3
MSCOCO (non-target domain)	25.7	41.7	66.0

Table 10: Results of MBR-CBDT decoding with different domain memories in the nocaps image captioning task.

We investigated the effectiveness of using target domain data for memory construction. As mentioned in Section 4.2, we constructed the memory using the Localized Narratives (LN) dataset (Pont-Tuset et al., 2020) for the image captioning task on the nocaps dataset. We compared this with constructing the memory from the training set of the MSCOCO dataset, which differs from the target domain. Table 10 shows the results. When using non-target domain data for memory construction, MBR-CBDT degraded in BLEU and BS. These results indicate that MBR-CBDT decoding is effective when the memory contains similar examples for the current input.

C.2 Additional case studies

Input	Insulin Human Winthrop Comb 50 ist eine Flüssigkeit (Suspension) zum Spritzen unter die Haut .
Reference	Insulin Human Winthrop Comb 50 is a fluid (suspension) for injection under the skin .
*k*NN-MT	Insulin Human Winthrop Comb 50 is a liquid (suspension) for injection under the skin .
MBR	Insulin Human Winthrop Comb 50 is a liquid (suspension) to be sprayed under the skin .
MBR-CBDT	Insulin Human Winthrop Comb 50 is a liquid (suspension) for injection under the skin .

Table 11: Translation examples in the medical domain. Both MBR and MBR-CBDT used CHRF for the utility function. Highlighted spans indicate the difference between translations.

We also present examples of medical translation in Table 11. MBR-CBDT correctly translated "*zum Spritzen unter die*" to "*for injection under the skin*", while MBR translated it to "*to be sprayed under the skin*". *k*NN-MT generated a mistranslation in Table 4, but it succeeded in this case. In this example, there were 11 instances in the memory where the input text contained "*zum Spritzen unter die*" and the corresponding reference contained "*for injection under the skin*". Similar to Table 4, there were no memory instances where the reference text contained "*to be sprayed under the skin*" when the input text contained "*zum Spritzen unter die*".

D Licenses

Datasets The five domain De–En translation datasets can be used for research purposes as described in the original paper (Koehn and Knowles, 2017; Aharoni and Goldberg, 2020). ASPEC (Nakazawa et al., 2016) can be used for research purposes as described in https://jipsti.jst.go.jp/aspec/. KFTT (Neubig, 2011) is li-

[13]We used Localized Narratives (Pont-Tuset et al., 2020) as described in Section 4.2.

[14]We did not use the test set but used the development set for the nocaps evaluation because the reference captions of the test set are not publicly available. Thus, we tuned the hyperparameters for both MSCOCO and nocaps on the development set of MSCOCO.

Model	License	Reference
facebook/m2m100_418M	MIT	Fan et al. (2021)
intfloat/multilingual-e5-large-instruct	MIT	Wang et al. (2024)
Unbabel/wmt22-comet-da	Apache-2.0	Rei et al. (2022a)
Unbabel/wmt22-cometkiwi-da	CC BY-NC-SA 4.0	Rei et al. (2022b)
Salesforce/blip2-flan-t5-xl	MIT	Li et al. (2023)
facebook/dinov2-large	Apache-2.0	Oquab et al. (2024)
microsoft/deberta-v3-large	MIT	He et al. (2023)
sentence-transformers/LaBSE	Apache-2.0	Feng et al. (2022)

Table 12: Licenses of models we used.

censed by Creative Commons Attribution-Share-Alike License 3.0. In the MSCOCO dataset (Lin et al., 2014; Karpathy and Fei-Fei, 2015), the annotations belong to the COCO Consortium and are licensed under a Creative Commons Attribution 4.0 license. The COCO Consortium does not own the copyright of the images, and use of the images must abide by the Flickr Terms of Use. The nocaps dataset (Agrawal et al., 2019) is released under a CC-BY-2.0 License, and the Localized Narratives dataset (Pont-Tuset et al., 2020) is released under a CC-BY-4.0 License.

Models The licenses of models we used are listed in Table 12.

Tools We used the following MIT-licensed tools.

- `mbrs` (Deguchi et al., 2024a)
- `knn-transformers` (Alon et al., 2022)
- `Faiss` (Johnson et al., 2019; Douze et al., 2024)
- `BM25S` (Lù, 2024)