# 最小ベイズリスク復号におけるバイアスと多様性の分解

上垣外 英剛 [1]　出口 祥之 [1*]　坂井 優介 [1]　林 克彦 [2]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 東京大学
{kamigaito.h, deguchi.hiroyuki.db0,
sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp,
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

## 概要

一般的な自然言語生成は、探索空間を制限し出力品質を低下させる貪欲法やビーム探索に依存している。最小ベイズリスク (MBR) 復号は、自動評価尺度とモデルが生成した擬似参照を利用することでこの問題を緩和する。従来研究では、MBR 復号による生成性能の改善を明らかにするための経験的分析が行われ、様々な観察結果が報告されている。その一方で、それらの理論的な背景は不確かである。これに対処するために、本研究ではバイアス-多様性分解の観点から MBR 復号の新しい理論的解釈を提示する。この解釈では、MBR 復号による仮説の品質推定の誤差を、効用関数と人間の評価との近接性を考慮した**バイアス**と効用関数の品質推定のばらつきを表す**多様性**の二つの主要な要因に分解する。理論的分析により、バイアスと多様性の両方を同時に改善することの難しさが明らかになり、多様性を高めることによる MBR 復号の性能向上の妥当性が確認された。また、複数の NLP タスクにおける実験により、理論的特性と合致する結果が観測された。

## 1　はじめに

大規模言語モデル (LLM) の成功が示すように [1]、自然言語生成は自然言語処理における重要タスクの一つである。これらは一般に、探索空間を制限することで、生成テキストの品質低下を招く可能性がある貪欲法やビーム探索に依存している。

ベイズ最小リスク (MBR) 復号 [2] は、モデルが生成した擬似参照と共に、効用関数として自動評価尺度を使用することでこの問題を軽減できる。MBR 復号は当初、音声認識 [2] に適用され、その後、様々な自然言語生成タスクで利用されている。

このため、人手評価と高い相関を持つ評価尺度を効用関数として使用することを推奨する研究 [3, 4, 5, 6] や人手のものに近い、高品質な擬似参照と、その多様性の重要性を指摘する研究 [7, 8] 等、様々な経験的分析が行われている。これらは MBR 復号の様々な側面を扱っているが、理論的背景の欠如により、統一的な解釈が依然として困難である。

本研究ではこの問題に対処するために、MBR 復号の理論的解釈をバイアス-多様性分解 [9, 10] に基づき行う。この解釈は、MBR 復号における各仮説への品質推定の誤差に焦点を当てている。これらの誤差は、**バイアス** と **多様性** の二つの重要な要素に分解される。バイアスは効用関数の推定する品質と人間の評価の近さを表し、多様性は効用関数の品質推定のばらつきを反映する。この解釈に基づき、MBR 復号における多様性の向上が重要であることと、バイアスと多様性の両方を同時に改善することの困難さを理論的に示し、過去の研究から得られた経験的な知見との対応を確認した。さらに、機械翻訳、テキスト要約、画像キャプション生成を対象とした実験を実施した結果、我々の理論的な分析に沿う傾向が実際に観測された。

## 2　最小ベイズリスク復号化

MBR 復号 [11, 12] は、入力系列 $x$ に対するモデルの予測確率 $P(y|x)$ からサンプリングされた擬似参照 $y$ の集合 $|\mathcal{Y}|$ を用いて、候補集合 $\mathcal{H}$ 内の仮説 $h$ の品質を推定する。評価尺度を効用関数 $f_\theta(h, y)$ として扱い、$h$ と $y$ の類似度を計算し、MBR 復号は以下のように $\mathcal{H}$ 内の最良の仮説 $\hat{h}_{best}$ を選択する:

$$\hat{h}_{mbr} = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \sum_{y \in \mathcal{Y}} f_\theta(h, y), \quad y \sim P(y|x). \tag{1}$$

$\theta$ は効用関数 $f_\theta(h, y)$ として使用される評価尺度のパラメータを表す。ここで、効用関数 $f_\theta(h, y)$ の代わりに、人間が推定する品質 [13, 14, 7, 15] を $\hat{f}_{\hat{\theta}}(h)$ のように仮定できる。この仮定下で、人間が推定す

* 現在、NTT コミュニケーション科学基礎研究所。

る理想的な復号は以下で表される:

$$\hat{h}_{human} = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \hat{f}_{\hat{\theta}}(h). \tag{2}$$

本稿では、MBR 復号により推定される品質と人間が推定する品質との違いを分析し、MBR 復号の特性をより深く理解することに焦点を当てる。

# 3 理論的解析

人間が推定した品質 $\hat{f}_{\hat{\theta}}(h)$ と MBR 復号が推定した品質 $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_\theta(h,y)$ との間の不一致を測定するために、$j$ 番目の疑似参照に基づいて各仮説の推定品質を表す $|\mathcal{H}|$ 次元のベクトル $\mathbf{u}^j$ を定義し、全ての $\mathbf{u}^j$ の平均ベクトル $\bar{\mathbf{u}}$ を以下のように定義する:

$$\mathbf{u}^j = \begin{bmatrix} u_1^j \\ \cdots \\ u_{|\mathcal{H}|}^j \end{bmatrix}, \quad u_i^j = f_\theta(h_i, y_j), \quad \bar{\mathbf{u}} = \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} \mathbf{u}^j. \tag{3}$$

同様に、各仮説に対する人間が推定した品質を表す $|\mathcal{H}|$ 次元のベクトル $\hat{\mathbf{u}}$ を以下のように定義する:

$$\hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 \\ \cdots \\ \hat{u}_{|\mathcal{H}|} \end{bmatrix}, \quad \hat{u}_i = \hat{f}_{\hat{\theta}}(h_i). \tag{4}$$

ここで、式 (3) および式 (4) を用いて、式 (1) の MBR 復号と式 (2) の理想的な復号を再定式化する:

$$(1) \equiv \hat{h}_{mbr} = \underset{h_i}{\operatorname{argmax}} \bar{u}_i, \ (2) \equiv \hat{h}_{human} = \underset{h_i}{\operatorname{argmax}} \hat{u}_i. \tag{5}$$

式 (5) に基づき、$\bar{\mathbf{u}}$ と $\hat{\mathbf{u}}$ を比較することで、MBR 復号と人間の間での品質推定の不一致である予測誤差を調査できる。本研究では、平均二乗誤差 (MSE) を用いて $\bar{\mathbf{u}}$ と $\hat{\mathbf{u}}$ の予測誤差を以下のように扱う:

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) = \frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}(\hat{u}_i - \bar{u}_i)^2. \tag{6}$$

式 (6) から以下の定理が導かれる。

**定理 1** MBR 復号における仮説の品質推定誤差 $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$ は、次のようにバイアスと多様性（曖昧さ）の項に分解できる [9][1):

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) = \underbrace{\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2}_{\text{バイアス}} - \underbrace{\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2}_{\text{多様性}}. \tag{7}$$

1) 証明は付録 A を参照。

式 (7) の二つの項はそれぞれバイアスと多様性を表している。有名なバイアス-分散分解 [16] が単一の推定器を対象とするのとは異なり、今回のバイアスと多様性への分解は複数の推定結果に基づく $\mathbf{u}$ を対象としている。また多様性を表す第二項は負となっており、多様性の増加が誤差の低減に寄与することが分かる。これが第二項が分散ではなく多様性と呼ばれる理由である [10]。バイアス項は、仮説の品質に対する効用関数の推定が人間の評価にどれだけ近いかを、多様性項は、効用関数の各推定品質が互いにどれだけ異なるかを示す。この分解は、式 (6) の仮説候補全てに対する品質推定誤差を改善するためには、各仮説に対する品質推定誤差 $(\hat{u}_i - \bar{u}_i)^2$ を改善する必要があり、各仮説に対するバイアス項を減少させつつ多様性項を増加させることの重要性を強調している。

## 3.1 解釈

定理 1 で提示された分解は、以下の小節で説明するよう、MBR 復号について従来研究で経験的に分析された結果に対し理論的な解釈を提供する。

### 3.1.1 人間による推定との相関

バイアス項は、MBR 復号の性能向上のために効用関数と人間の推定が近接する必要性を示している。これは効用関数 $f_\theta(h_i, y_j)$ が疑似参照 $y_j$ に影響されるため、効用関数とサンプリング手法の両方が重要であることを意味する。従来研究 [15, 7] は、適切な擬似参照を選択する必要性を経験的に示しており、我々の発見はこれらを理論的に支持する。

### 3.1.2 評価尺度の多様性

分解の多様性項より、多様性を増加させることは、MBR 復号における推定誤差を減少させ、性能向上に寄与する。ここでの重要な洞察は、$(\bar{u}_i - f_\theta(h_i, y_j))^2$ によって表される多様性が、各効用関数 $f_\theta(h_i, y_j)$ によって生成される異なる品質推定に起因することである。この多様性は疑似参照 $y_j$ や評価尺度のモデルパラメータ $\theta$ による影響を受ける。この発見は、サンプリング手法の多様性が MBR 復号の性能向上に不可欠であると結論付けた先行研究 [17, 7, 8] を支持する。これは、疑似参照の多様性が各 $y_j$ によって $f_\theta(h_i, y_j)$ の多様性を間接的に高めることに寄与するためである。

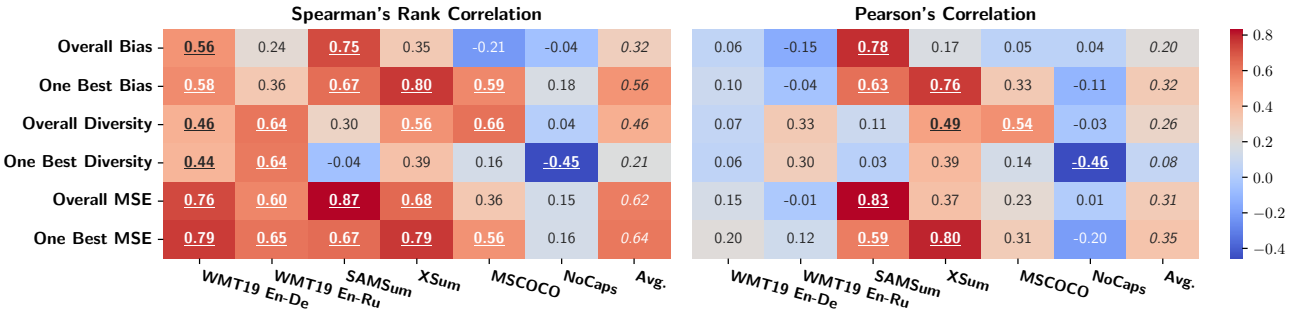


図 1: 各データセットにおけるバイアスと多様性に関する尺度とタスク性能の相関。下線付きのスコアは統計的に有意な結果を示す ($p < 0.05$)[2]。平均値 *Avg.* の斜体のスコアは有意差検定の対象ではないことを示す。

### 3.1.3 バイアスと多様性のトレードオフ

一見すると、§3.1.1 および §3.1.2 での解釈に基づき、MBR 復号における品質推定性能を向上させるための最善の戦略は、バイアスを減少させつつ多様性を高めることであるように思える。この戦略の妥当性は従来研究 [7] が経験的な調査に基づく提示している。しかし、この戦略の有効性を理解するためには、バイアスと多様性のトレードオフ [9] に注目する必要がある。このトレードオフは、バイアスを減少させつつ多様性を高めることの難しさを浮き彫りにする。式 (7) において、バイアス項がゼロに近づく場合、多様性の項もゼロに近づく。この理論的事実は、たとえ人間とよく相関する評価尺度や高品質な擬似参照を用意できても、多様性の低下により MBR 復号の性能が向上しない可能性があることを示している。一方で、評価尺度および擬似参照の品質が低い場合、バイアスの増加を犠牲にして多様性を高めることで性能の向上が期待できる。

## 4 経験的な分析

MBR 復号の包括的な理解を行うために、実験を通じて理論分析に対応する検証を行う。

### 4.1 実験設定

自然言語生成タスクとして、機械翻訳、テキスト要約、画像キャプション生成を対象とする。全タスクで、サンプリングの設定は文献 [19] に従った。仮説の生成は epsilon sampling [20] を用いた。擬似参照の生成にはビーム探索とサンプリングとして nucleus [21] ($p = 0.9$), ancestral, top-$k$ [22] ($k = 10$), epsilon ($\epsilon = 0.02$) を使用した。仮説のサイズは 64 に設定し、擬似参照のサイズは{4, 8, 16, 32, 64}から選択した。各タスクに対して[3]、以下のデータセット、モデル[4]、および評価尺度を使用した:

**機械翻訳** WMT19 の英語からドイツ語（En-De）および英語からロシア語（En-Ru）のデータセット [24] を使用した。En-De には `facebook/wmt19-en-de`、En-Ru には `facebook/wmt19-en-ru` を使用した。効用関数および評価指標として、`Unbabel/wmt22-comet-da` モデルを用いた COMET を使用した。

**要約** SAMSum [25] と XSum [26] を使用し、生成には `philschmid/bart-large-cnn-samsum` を SAMSum で、`facebook/bart-large-xsum` を XSum で利用した。効用関数および評価指標として、BERTScore [27] を `microsoft/deberta-xlarge-mnli` で使用した。

**画像キャプション生成** MSCOCO [28] の文献 [29] に基づく分割および NoCaps [30] を使用した。MSCOCO には `Salesforce/blip2-flan-t5-xl-coco`、NoCaps には `Salesforce/blip2-flan-t5-xl` を使用した。効用関数および評価指標として、BERTScore を `microsoft/deberta-xlarge-mnli` で使用した。

評価対象は、式 (6) として Overall MSE を、式 (7) の第一項として Overall Bias を、式 (7) の第二項として Overall Diversity を、MBR 復号で最良と推定された候補のみを考慮した際の式 (7) の第一項として One Best Bias を、その際の式 (7) の第二項として One Best Diversity を、その際の式 (6) として One Best MSE を使用した。バイアス項の計算については自動評価尺度を用いて近似的に概算した[5]。

### 4.2 実験結果: タスク性能との相関

バイアスと多様性のタスク性能への相関を調査した。比較のために、各データセットでの異なる 5 つのサンプリング方法と 5 つのサンプルサイズにおけるタスク性能に対し、評価対象のスピアマンの順位相関およびピアソン相関を計算した。性能向上には

2) 両相関係数に Student の t 検定 [18] を用いた。
3) 詳細な設定は付録 B を参照
4) モデルは `https://huggingface.co/models` [23] から使用。
5) 詳細は付録 C を参照。

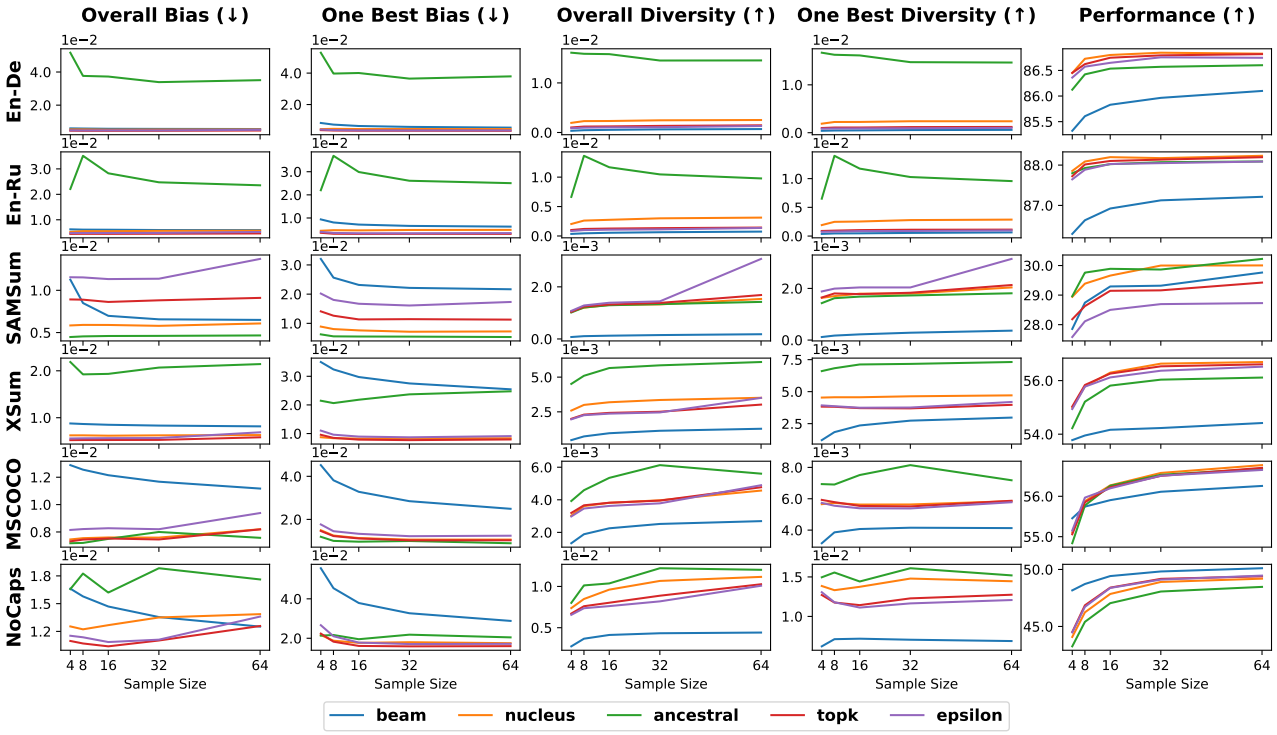


図 2: MBR 復号におけるバイアス、多様性、およびタスク性能の関係。x 軸は使用された擬似参照の数を示す。(↑) はスコアが高いほど良く、(↓) はスコアが低いほど良いことを示す。

バイアスと MSE が低い方が良いため、相関計算ではこれらの負の値を利用した。また、フィッシャーの z 変換 [31] により各データセットの相関を平均化して報告した。

図 1 は各データセットにおける評価対象とタスク性能の相関を示している。これらの結果より、One Best Bias、Overall Diversity、One Best MSE、Overall MSE の各指標がタスク性能に対し良好な順位相関を示すが、ピアソンの相関係数より、値の僅かな差異を正確に捉えることは難しい。

### 4.3 実験結果: バイアスと多様性

バイアスと多様性のトレードオフを検証するために、タスク評価性能に対するバイアスと多様性の関係を調査した。図 2 に各データセットにおける異なるサンプリング方法を用いた結果をプロットした。SAMSum データセットを除き、ancestral は最悪のバイアスを示す一方、多様性が高いために他のサンプリング方法を上回ることがあることを示している。バイアスが最も低い topk に注目すると、SAMSum データセットを除いて、バイアスの改善が多様性の増加を制限する傾向が見られる。この発見は、MBR デコーディングにおけるバイアスと多様性のトレードオフを支持する。しかし、多様性が最も低いビーム探索の性能から明らかなように、バイアスと多様性の重要性は対象データセットによって異なる。したがって、理論的分析は MBR 復号における性能の傾向を効果的に説明するが、性能向上のためにはタスク固有の特徴を適切に考慮することも依然として重要であることを示唆する。

## 5 結論

本研究では MBR 復号において経験的に得られた知見を統一的に理解するための理論的解釈を提供した。具体的には MBR 復号における生成されたテキストに対する品質推定を行う際に、人間の推定結果との差異、すなわち誤差をバイアスと多様性に分解した。そして、これら二つの要素がどのように誤差と関連するかを示し、MBR 復号における誤差の改善におけるバイアスと多様性のトレードオフの関係を明らかにし、既存研究で報告されている経験的に得られた知見と対応付けながら説明した。その上で特に多様性の向上の利点を強調した。これらにより、我々の理論的洞察が従来の経験的結果と一致しており、また、複数タスクにおける実験結果も我々の理論的発見が妥当であることを示した。

# 謝辞

本研究は JSPS 科研費 JP23H03458 の助成を受けたものです.

# 参考文献


[1] OpenAI. Gpt-4 technical report, 2024.

[2] Vaibhava Goel and William J Byrne. Minimum bayes-risk automatic speech recognition. **Computer Speech & Language**, Vol. 14, No. 2, pp. 115–135, 2000.

[3] Mathias Müller and Rico Sennrich. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **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)**, pp. 259–272, Online, August 2021. Association for Computational Linguistics.

[4] Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 811–825, 07 2022.

[5] Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics.

[6] Chantal Amrhein and Rico Sennrich. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1125–1141, Online only, November 2022. Association for Computational Linguistics.

[7] Yuu Jinnai, Ukyo Honda, Tetsuro Morimura, and Peinan Zhang. Generating diverse and high-quality texts by minimum Bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics ACL 2024**, pp. 8494–8525, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[8] David Heineman, Yao Dou, and Wei Xu. Improving minimum bayes risk decoding with multi-prompt, 2024.

[9] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, and T. Leen, editors, **Advances in Neural Information Processing Systems**, Vol. 7. MIT Press, 1994.

[10] Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. **J. Mach. Learn. Res.**, Vol. 24, No. 1, mar 2024.

[11] Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[12] Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Subhajit Naskar, Daniel Deutsch, and Markus Freitag. Quality estimation using minimum Bayes risk. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 806–811, Singapore, December 2023. Association for Computational Linguistics.

[14] Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 4265–4293, Toronto, Canada, July 2023. Association for Computational Linguistics.

[15] Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. On the true distribution approximation of minimum Bayes-risk decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 459–468, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[16] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. **Neural computation**, Vol. 4, No. 1, pp. 1–58, 1992.

[17] Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 9198–9209, Singapore, December 2023. Association for Computational Linguistics.

[18] Student. Probable error of a correlation coefficient. **Biometrika**, pp. 302–310, 1908.

[19] Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. Model-based minimum Bayes risk decoding for text generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, **Proceedings of the 41st International Conference on Machine Learning**, Vol. 235 of **Proceedings of Machine Learning Research**, pp. 22326–22347. PMLR, 21–27 Jul 2024.

[20] John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

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

[22] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[24] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, **Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)**, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics.

[25] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, **Proceedings of the 2nd Workshop on New Frontiers in Summarization**, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics.

[26] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

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

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

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

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

[31] David M Corey, William P Dunlap, and Michael J Burke. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. **The Journal of general psychology**, Vol. 125, No. 3, pp. 245–261, 1998.

[32] Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mbrs: A library for minimum bayes risk decoding, 2024.

[33] 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. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 578–628, Singapore, December 2023. Association for Computational Linguistics.

[34] Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. Results of the WMT16 metrics shared task. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi, editors, **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**, pp. 199–231, Berlin, Germany, August 2016. Association for Computational Linguistics.

## A　定理 1 の証明

まず、$(\hat{u}_i - \bar{u}_i)^2$ を次のように展開する:

$$(\hat{u}_i - \bar{u}_i)^2 \tag{8}$$

$$=(\hat{u}_i)^2 - 2\hat{u}_i\bar{u}_i + (\bar{u}_i)^2 \tag{9}$$

$$=(\hat{u}_i)^2 - 2\hat{u}_i\bar{u}_i + 2(\bar{u}_i)^2 - (\bar{u}_i)^2 \tag{10}$$

$$=(\hat{u}_i)^2 - 2\hat{u}_i\bar{u}_i + 2\bar{u}_i\bar{u}_i - (\bar{u}_i)^2 \tag{11}$$

$$=(\hat{u}_i)^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\hat{u}_i u_i^j + \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\bar{u}_i u_i^j - (\bar{u}_i)^2 \tag{12}$$

$$=\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i)^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\hat{u}_i u_i^j + \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\bar{u}_i u_i^j - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i)^2 \tag{13}$$

$$=\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i)^2 - 2\hat{u}_i u_i^j + 2\bar{u}_i u_i^j - (\bar{u}_i)^2) \tag{14}$$

$$=\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i)^2 - 2\hat{u}_i u_i^j + (u_i^j)^2 - (u_i^j)^2 + 2\bar{u}_i u_i^j - (\bar{u}_i)^2) \tag{15}$$

$$=\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i)^2 - 2\hat{u}_i u_i^j + (u_i^j)^2 - ((u_i^j)^2 - 2\bar{u}_i u_i^j + (\bar{u}_i)^2)) \tag{16}$$

$$=\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i - u_i^j)^2 - (\bar{u}_i - u_i^j)^2) \tag{17}$$

$$=\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2 \tag{18}$$

ここで、式 (18) の結果を用いることで、次のように $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$ を分解する:

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) \tag{19}$$

$$=\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}(\hat{u}_i - \bar{u}_i)^2 \tag{20}$$

$$=\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\Big(\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2\Big) \tag{21}$$

$$=\underbrace{\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2}_{\text{バイアス}} - \underbrace{\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2}_{\text{多様性}} \tag{22}$$

以上より、定理 1 が証明された。

## B　実験設定の詳細

生成部分の実装は [19] の公開コード[6]を基にしており、MBR デコーディング部分は [32] によるツールキット `mbrs`[7]を基にしている。サンプルの生成は NVIDIA GeForce RTX 3090 上で行い、MBR デコーディングは NVIDIA RTX A6000 上で実施した。

## C　バイアス項の近似計算

バイアス項の計算には人間による評価が必要であり、各設定ごとに人間による評価を実施することは現実的ではない。この問題に対処するために、本研究ではバイアス項の近似として擬似バイアスを導入する。人手で作成された参照 $\hat{y}$ の数 $|\hat{\mathcal{Y}}|$ を用いて、擬似バイアスは以下のように定義される：

$$\underbrace{\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}\left(\hat{u}_i - f_\theta(h_i, y_j)\right)^2}_{\text{バイアス}} \approx \underbrace{\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}\left(\widetilde{u}_i - f_\theta(h_i, y_j)\right)^2}_{\text{擬似バイアス}},$$

$$\widetilde{u}_i = \frac{1}{|\hat{\mathcal{Y}}|}\sum_{j=1}^{|\hat{\mathcal{Y}}|} f_\theta(h_i, \hat{y}_j). \tag{23}$$

この定式化は、自動評価尺度が人手で作成された参照を使用した際に人手評価と相関するという前提に基づく。擬似バイアスには、COMET (`Unbabel/wmt22-comet-da`) および `microsoft/deberta-xlarge-mnli` をエンコーダとして使用した BERTScore を用いた。これらはそれぞれ WMT23 の英語からドイツ語へのシステムレベル翻訳タスクでのピアソン相関係数が 0.990 [33]、および WMT16 での英語への翻訳タスクで 0.7781[8]と人手評価への高い相関を示している [34]。

6) https://github.com/CyberAgentAILab/model-based-mbr
7) https://github.com/naist-nlp/mbrs
8) https://github.com/Tiiiger/bert_score