事例ベース意思決定理論に基づく復号

出口 祥之
NTT コミュニケーション科学基礎研究所
hiroyuki.deguchi@ntt.com

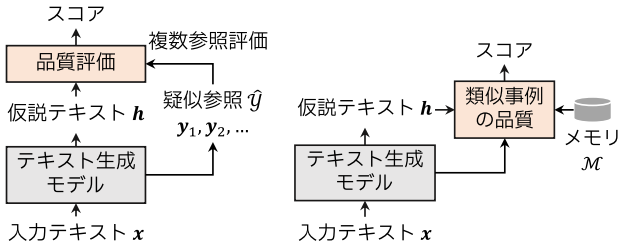
概要

最小ベイズリスク（minimum Bayes risk; MBR）復号は、出力候補の中から品質の期待値を最大化する仮説を選択する復号法であり、従来の最大事後確率復号よりも高品質なテキストを出力する。しかし、MBR 復号の出力は、テキスト生成モデルが生成するサンプルに依存するため、生成モデルの学習が不十分なドメインにおいては、ドメインの知識や情報を反映したテキストを出力することは難しい。この課題に対処するため、本研究では、ドメインデータを利用した事例ベース意思決定理論に基づく復号を提案する。独英ドメイン翻訳実験より、提案法は、最大事後確率復号よりも高品質なテキストを出力でき、また、MBR 復号と組み合わせることで、MBR 復号よりもドメインに特化した高品質なテキストを出力できることを確認した。

1 はじめに

最小ベイズリスク（minimum Bayes risk; MBR）復号は、仮説を出力したときに得られる効用の期待値（期待効用）を最大化する復号法であり、従来の最大事後確率（maximum a posteriori; MAP）復号と比較して、高品質で誤りの少ない頑健なテキストを生成する [1, 2, 3, 4, 5, 6]。MBR 復号の中心となる期待効用の最大化は、不確実性下での最善の意思決定を導くために提唱された期待効用理論（Expected Utility Theory; EUT）に基づいており [7]、自然言語処理のみならず、ミクロ経済や音声認識などに広く応用されている [8, 9, 10]。しかし、EUT は、直面する問題に対しての知識が不足していると、期待効用を正確に推定できず、選好を反映した仮説が選択できないという課題がある [11]。そのため、EUT に基づく MBR 復号では、ドメインの知識や情報を反映したテキストを出力することが難しい。

このような EUT の課題に対し、意思決定理論の分野では、過去に経験した事例をもとに帰納的に最善の行動を導く事例ベース意思決定理論（Case-Based Decision Theory; CBDT）が提唱されている [11]。CBDT は、現在直面する問題に対して、過去に経験した類似事例に着目し、その際に選択した行動と、その行動によって得られた報酬から、現在の行動の価値を類推的に評価する。



(a) MBR 復号（従来法）。 (b) CBDT 復号（提案法）。

図 1: MBR 復号と CBDT 復号の比較。復号時は複数の仮説から、スコアを最大化する仮説を選択する。

本研究では、CBDT に基づく復号（CBDT 復号）を提案し、ドメインデータを利用することで、ドメインの情報を反映した高品質なテキストの生成を目指す。CBDT 復号は、あらかじめ事前に用意したデータに対して複数の仮説を生成し、各仮説の品質を参照テキストにより評価し、メモリに記憶する。復号時は、メモリから類似事例を探索し、類似事例の仮説の品質に対して入力側・出力側それぞれの類似度で重みを付け、出力スコアを計算する。図 1 に MBR 復号（図 1(a)）と CBDT 復号（図 1(b)）の概要図を示す。MBR 復号は、疑似参照と呼ばれる複数のサンプルテキストを用いて複数参照評価を行うのに対し、CBDT 復号はメモリに記憶された類似事例の品質を利用する。さらに、本稿では、MBR 復号と CBDT 復号を組み合わせることで、より高品質なテキストを出力できることを示す。

5 つのドメインでの独英翻訳実験 [12, 13] を行った結果、提案法の CBDT 復号は、MAP 復号よりも高品質なテキストを出力でき、また、MBR-CBDT 復号が MBR 復号よりもドメインに特化した高品質な出力が得られることを確認した。

2 背景・関連研究

最小ベイズリスク復号 テキスト生成における入力テキストを $\boldsymbol{x} \in \mathcal{X}$、出力テキストを $\boldsymbol{y} \in \mathcal{Y}$ とする。ただし、$\mathcal{X}, \mathcal{Y} \subseteq \mathcal{V}^*$ はそれぞれ入力空間と出力空間を表し、$\mathcal{V}^*$ は語彙 $\mathcal{V}$ のクリーネ閉包を示す。一般的なテキスト生成法である MAP 復号は、テキスト生成モデル θ を用い、入力テキストで条件付けされた生成確率を最大化する出力テキスト $\boldsymbol{y}^{\text{MAP}} = \operatorname{argmax}_{\boldsymbol{h} \in \mathcal{H}} p(\boldsymbol{h}|\boldsymbol{x};\theta)$ を選択する。ただし、出力空間 $\mathcal{Y}$ 全体を探索することはできないため、ビーム探索などによって得られた仮説集合 $\mathcal{H} \subset \mathcal{Y}$ から選択する。一方、MBR 復号は、効用関数 $u\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ の期待値（期待効用）を最大化する仮説 $\boldsymbol{y}^{\text{MBR}}$ を選択する：

$$\boldsymbol{y}^{\text{MBR}} = \operatorname*{argmax}_{\boldsymbol{h} \in \mathcal{H}} U^{\text{MBR}}(\boldsymbol{h};\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{h} \in \mathcal{H}} \mathop{\mathbb{E}}_{\boldsymbol{y} \sim \Pr(\cdot|\boldsymbol{x})} [u(\boldsymbol{h},\boldsymbol{y})]\,. \tag{1}$$

$\boldsymbol{y} \in \mathcal{Y}$ は真の出力確率 $\Pr(\cdot|\boldsymbol{x})$ に従って得られる参照テキストであり、効用関数 u は一般に出力テキストの品質を測る評価指標が用いられる。ただし、選好関係を $\succeq$ で表すとき、効用関数 u は $\boldsymbol{h} \succeq \boldsymbol{h}' \Leftrightarrow u(\boldsymbol{h},\boldsymbol{y}) \geq u(\boldsymbol{h}',\boldsymbol{y})$ を満たす。ここで、$\Pr(\cdot|\boldsymbol{x})$ は未知であるため、期待効用 U^{MBR} は典型的には Monte Carlo（MC）法により推定される [1, 2]：

$$U^{\text{McMBR}}(\boldsymbol{h};\hat{\mathcal{Y}}) = \frac{1}{|\hat{\mathcal{Y}}|} \sum_{\boldsymbol{y} \in \hat{\mathcal{Y}}} u(\boldsymbol{h},\boldsymbol{y}). \tag{2}$$

なお、参照テキストの入手は困難であるため、テキスト生成モデルからサンプリングした疑似的な参照テキスト（疑似参照）の多重集合 $\hat{\mathcal{Y}} \coloneqq \{\boldsymbol{y}_i\}_{i=1}^{|\hat{\mathcal{Y}}|} \sim p(\boldsymbol{y}|\boldsymbol{x};\theta)$ を用いる。

MBR 復号の出力は、疑似参照の分布に依存することが報告されている [14]。テキスト生成モデルの学習が不十分なドメインにおいては、疑似参照と真の参照テキストとの間の分布のずれにより、期待効用を正確に推定できないことがある。そのため、例えばテキスト生成モデルが学習していないドメインにおいては、ドメイン特有の知識や情報を反映した仮説を選択することが難しい。

事例ベース意思決定理論 意思決定理論の分野では、実際に過去に経験した事例をもとに最善の行動を導く事例ベース意思決定理論（Case-Based Decision Theory; CBDT）が提唱されている [11]。CBDT に従う意思決定者は、現在直面する問題に対し、過去に類似した問題下で選択した行動とその際に得られた報酬から、最善の行動選択を予測する。問題の集合を $\mathcal{Q}$、行動の集合を $\mathcal{A}$、報酬空間を $\mathcal{R}$ とすると、事例集合は $\mathcal{C} \coloneqq \mathcal{Q} \times \mathcal{A} \times \mathcal{R}$ と定義される。ここで、現在直面している問題を $q \in \mathcal{Q}$ とすると、CBDT に従う意思決定者はメモリ $\mathcal{M} \subseteq \mathcal{C}$ に基づいて、次式に従い行動 $a^\star \in \mathcal{A}$ を選択する：

$$a^\star = \operatorname*{argmax}_{a \in \mathcal{A}} \sum_{(q',a',r') \in \mathcal{M}} s(q,q') \mathbb{1}_{a=a'} r'. \tag{3}$$

なお、$s\colon \mathcal{Q} \times \mathcal{Q} \to [0,1]$ は問題の近さを表す類似度関数である。すなわち、CBDT に基づく意思決定では、行動 a を選択したときの事例に着目し、その際に得られた報酬に問題の類似度を重み付けした荷重和を最大化する行動を選択する。

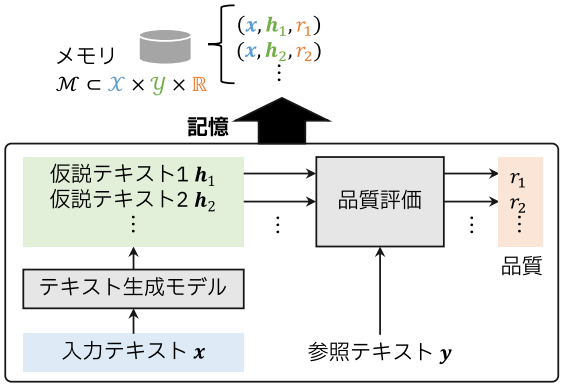


図 2: CBDT 復号の記憶時の概要図。

3 提案法：事例ベース意思決定理論に基づく復号

本研究では、**事例ベース意思決定理論に基づく復号（CBDT 復号）**を提案し、ドメインに応じたデータを活用することでテキスト生成の品質改善を狙う。提案法は、事前に、事例集合であるメモリにデータを記憶し（図 2）、復号時はメモリを参照しながら出力仮説を選択する（図 3）。なお、本論文ではテキスト生成を扱うため、これ以降、意思決定における問題 $\mathcal{Q}$ を入力テキスト $\mathcal{X}$ に、行動 $\mathcal{A}$ を出力テキスト $\mathcal{Y}$ に、報酬 $\mathcal{R}$ を出力テキストの品質評価スコア $\mathbb{R}$ に、それぞれ読み替える。

記憶 あらかじめ、入力テキストと正解の出力テキスト（参照テキスト）の対からなるデータ $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathcal{D}|}$ から、メモリ $\mathcal{M}$ を構築する。はじめに、データ $\mathcal{D}$ 内の各入力テキスト $\boldsymbol{x}$ についてそれぞれ H 個の仮説 $\mathcal{H}_{\boldsymbol{x}} \subset \mathcal{Y}$ を生成する。

$$\mathcal{H}_{\boldsymbol{x}} \coloneqq \{\boldsymbol{h}_\ell\}_{\ell=1}^{H} \sim p(\cdot|\boldsymbol{x};\theta). \tag{4}$$

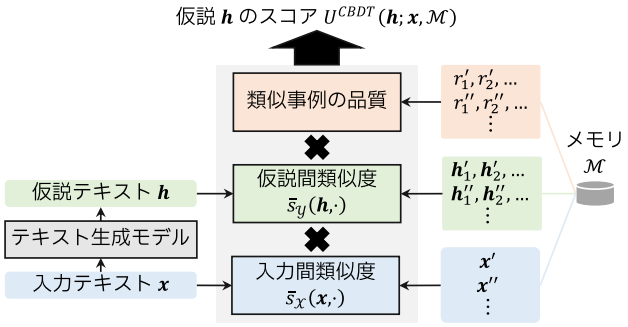


図 3: CBDT 復号の仮説 $\boldsymbol{h}$ に対するスコア計算の概要図。復号時は複数の仮説を生成し、スコア $U^{\textsc{Cbdt}}$ を最大化する仮説を選択する。

次に、生成した仮説に対し、参照テキスト $\boldsymbol{y}$ を用いて品質評価し、メモリ $\mathcal{M} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ に記憶する。

$$\mathcal{M} \coloneqq \{(\boldsymbol{x}, \boldsymbol{h}_\ell, u(\boldsymbol{h}_\ell, \boldsymbol{y})) \mid \boldsymbol{h}_\ell \in \mathcal{H}_{\boldsymbol{x}}, (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}\}. \quad (5)$$

復号　復号時は、事前構築したメモリを参照しながら、CBDT に基づき出力仮説を選択する。式 (3) に直接従った、CBDT に素朴に基づく復号 CbdtNaïve は次式のスコアを最大化する仮説を選択する：

$$U^{\textsc{CbdtNaïve}}(\boldsymbol{h};\boldsymbol{x},\mathcal{M}) = \sum_{(\boldsymbol{x}',\boldsymbol{h}',r')\in\mathcal{M}} s(\boldsymbol{x},\boldsymbol{x}')\mathbb{1}_{\boldsymbol{h}=\boldsymbol{h}'}r'. \quad (6)$$

ここで、CbdtNaïve は、次の 2 つの課題を抱えている。一つは、コサイン類似度やユークリッド距離などのテキスト間の類似度を表すさまざまな尺度は、必ずしも $[0,1]$ の値域をとるとは限らないため、s に使える尺度が限定される。もう一つは、指示関数 $\mathbb{1}_{\boldsymbol{h}=\boldsymbol{h}'}$ により、現在出力しようとしている仮説 $\boldsymbol{h}$ と全く同じテキストがメモリ $\mathcal{M}$ 中に含まれない限り、スコアを計算できない。これらの問題に対処するため、テキスト間の類似度を正規化した正規化類似度を用い、さらに、指示関数の代わりに仮説間 $\boldsymbol{h},\boldsymbol{h}'$ の類似度を利用するスコア関数 $U^{\textsc{Cbdt}}$ を提案する：

$$U^{\textsc{Cbdt}}(\boldsymbol{h};\boldsymbol{x},\mathcal{M}) = \sum_{(\boldsymbol{x}',\boldsymbol{h}',r')\in\mathcal{M}} \bar{s}_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}';\mathcal{M})\bar{s}_{\mathcal{Y}}(\boldsymbol{h},\boldsymbol{h}';\mathcal{H}_{\boldsymbol{x}'})r', \quad (7)$$

$$\bar{s}_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}';\mathcal{M}) = \frac{\exp\left(s_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}')/\tau_{\mathcal{X}}\right)}{\sum_{(\boldsymbol{x}'',\boldsymbol{h}'',r'')\in\mathcal{M}}\exp\left(s_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}'')/\tau_{\mathcal{X}}\right)}, \quad (8)$$

$$\bar{s}_{\mathcal{Y}}(\boldsymbol{h},\boldsymbol{h}';\mathcal{H}_{\boldsymbol{x}'}) = \frac{\exp\left(s_{\mathcal{Y}}(\boldsymbol{h},\boldsymbol{h}')/\tau_{\mathcal{Y}}\right)}{\sum_{\boldsymbol{h}''\in\mathcal{H}_{\boldsymbol{x}'}}\exp\left(s_{\mathcal{Y}}(\boldsymbol{h},\boldsymbol{h}'')/\tau_{\mathcal{Y}}\right)}. \quad (9)$$

$s_{\mathcal{X}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$、$s_{\mathcal{Y}}\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ は、それぞれ入力・出力テキスト間の類似度を表す関数、$\bar{s}_{\mathcal{X}}\colon \mathcal{X}\times\mathcal{X}\times 2^{\mathcal{M}}\to\mathbb{R}$、$\bar{s}_{\mathcal{Y}}\colon \mathcal{Y}\times\mathcal{Y}\times 2^{\mathcal{Y}}\to\mathbb{R}$ はそれらの正規化類似度であり、$\tau_{\mathcal{X}}$、$\tau_{\mathcal{Y}}$ は類似度に対する温度パラメータである。

本稿で提案する **CBDT 復号**は、入力テキストの類似度 $s_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}')$ が高い上位 k 件の事例 $\hat{\mathcal{M}} \subseteq \mathcal{M}$ に絞り込んだ後、前述のスコア関数 $U^{\textsc{Cbdt}}$ を最大化する仮説 $\boldsymbol{y}^{\textsc{Cbdt}}$ を選択する。なお、メモリは各入力テキスト $\boldsymbol{x}'$ につき H 件の仮説 $\mathcal{H}_{\boldsymbol{x}'}$ を持つため、$|\hat{\mathcal{M}}| = Hk$ であることに注意されたい。

$$\boldsymbol{y}^{\textsc{Cbdt}} = \operatorname*{argmax}_{\boldsymbol{h}\in\mathcal{H}} U^{\textsc{Cbdt}}(\boldsymbol{h};\boldsymbol{x},\hat{\mathcal{M}}). \quad (10)$$

MBR 復号と CBDT 復号　MBR 復号と CBDT 復号は、仮説集合の中から得られる効用を最大化する仮説の選択を目指す、という点において、目的が共通している。一方で、手法は直交しており、両者の組み合わせによってさらなる品質の改善が期待できる。ここで、両者のスコアを線形結合したスコアを最大化する、MBR-CBDT 復号を提案する：

$$\boldsymbol{y}^{\textsc{McMbr-Cbdt}} = \operatorname*{argmax}_{\boldsymbol{h}\in\mathcal{H}} \bar{U}^{\textsc{McMbr}}(\boldsymbol{h};\hat{\mathcal{Y}}) + \bar{U}^{\textsc{Cbdt}}(\boldsymbol{h};\boldsymbol{x},\hat{\mathcal{M}}). \quad (11)$$

ただし、$\bar{U}^{\textsc{McMbr}}$ と $\bar{U}^{\textsc{Cbdt}}$ は、それぞれ、仮説の $U^{\textsc{McMbr}}$、$U^{\textsc{Cbdt}}$ スコアを求めたあと、仮説集合 $\mathcal{H}$ 内の最小・最大スコアを用いて min-max 正規化する。

4　実験

提案法の有効性を確かめるため、5 つのドメイン（IT、コーラン、法律、医療、字幕）の独英翻訳データセット [12, 13] を用いて翻訳実験を行った。翻訳仮説を与えたときの、MAP 復号（Map）、品質推定（Quality Estimation; QE）モデルを用いたリランキング（Qe）、MBR 復号（McMbr）、CBDT 復号（Cbdt）および MBR-CBDT 復号（McMbr-Cbdt）による出力テキストの品質を評価した。

設定　復号の実装には mbrs [15] を用いた。仮説を生成する翻訳モデルには 418M パラメータの M2M100 [16] を使用した。復号時の翻訳仮説は epsilon サンプリング（$\varepsilon = 0.02$）[17] により、1 入力あたり 1024 候補生成した。MBR 復号のための疑似参照には仮説自身を用いた。CBDT 復号のメモリは、それぞれのドメインの訓練用対訳データから、1 件あたり $H = 256$ 個の仮説を生成して構築した。入力テキスト間、出力テキスト間の類似度には、どちらも LaBSE [18] の文埋め込みのコサイン類似度を用いた。スコア $U^{\textsc{Cbdt}}$ は、入力テキストの類似度が高い上位 $k = 256$ 件を用いて計算した。類似度の温度パラメータは、IT ドメインの開発データを用いて $\{1.0, 0.1, 0.01\}$ からチューニングし、$\tau_{\mathcal{X}} = 0.1$、$\tau_{\mathcal{Y}} = 0.01$ に設定した。効用関数 u には、

表 1: 独英ドメイン翻訳の実験結果。緑色の背景色の行は McMbr-Cbdt の行を表し、太字は各列の中で Oracle を除く最良のスコアを示す。[0, 1] の範囲で定義されるスコアはすべてパーセントで示す。

		IT			コーラン			法律			医療			字幕		
復号	効用	chrF	comet	kiwi	chrF	comet	kiwi	chrF	comet	kiwi	chrF	comet	kiwi	chrF	comet	kiwi
Map		45.5	76.1	73.4	23.7	57.9	62.3	48.5	74.6	72.6	50.7	78.0	76.7	39.8	73.5	76.0
Qe	kiwi	51.0	79.1	**81.2**	36.2	71.9	**80.5**	58.3	84.0	**83.1**	55.0	81.6	**83.1**	40.6	77.1	**83.3**
McMbr	chrF	52.7	77.8	75.1	**37.3**	68.1	73.1	60.1	82.3	79.5	56.7	80.3	79.0	42.4	74.4	76.9
Cbdt	chrF	51.7	77.3	71.5	35.0	64.0	65.8	58.0	80.2	74.9	55.6	78.5	73.9	38.9	70.3	72.0
McMbr-Cbdt	chrF	**54.0**	78.5	74.5	**37.3**	67.4	71.2	**60.8**	82.3	78.6	**57.9**	80.1	77.1	**43.1**	73.8	75.8
McMbr	comet	51.6	81.1	76.1	35.9	**73.5**	76.0	58.4	84.6	80.8	56.0	82.9	80.1	41.6	**77.9**	78.9
Cbdt	comet	49.8	79.7	74.0	33.8	67.6	70.0	56.3	82.2	77.8	53.1	80.9	77.3	36.9	73.5	75.3
McMbr-Cbdt	comet	52.1	**81.8**	76.2	35.3	72.5	74.9	58.4	**84.8**	80.5	55.6	**83.0**	79.7	40.3	77.6	78.9
Oracle	chrF	63.9	82.4	73.0	48.0	69.8	69.3	69.5	84.1	78.4	67.0	82.4	77.0	57.1	77.3	74.2
Oracle	comet	58.5	86.5	75.5	39.8	77.8	75.2	63.7	87.2	80.9	61.6	86.0	79.4	50.3	83.1	78.6

表 2: IT ドメイン翻訳の仮説選択の計算時間（秒）。

	効用		
復号	chrF	comet	kiwi
Qe	–	–	751.8
McMbr	11,379.7	1,281.4	–
Cbdt	366.2	353.3	–
McMbr-Cbdt	13,379.7	1,634.1	–

表 3: 医療ドメインの独英翻訳の出力例。効用関数に chrF を用いたときの結果を示している。

入力	Sehr häufig • Übelkeit • Erbrechen
参照	Very common • Vomiting • Nausea
McMbr	Very frequent • illness • vomiting
McMbr-Cbdt	Very frequent • nausea • vomiting

chrF [19] と comet [20] をそれぞれ用い、QE モデルには CometKiwi [21]（kiwi）を用いた。comet、kiwi によるスコア算出と LaBSE の文埋め込みは、すべてバッチサイズを 256 文に設定して計算した。翻訳品質は、chrF、comet、kiwi を用いて評価した。詳細な設定は付録 A に示す。

品質評価　ドメイン翻訳の実験結果を表 1 に示す。まず、Cbdt は Map と比べて、chrF において最大+11.3%、comet において最大+9.7%改善していることがわかる。また、McMbr-Cbdt と McMbr を比較すると、効用として与えた指標をより改善しており、chrF において最大+1.3%、comet において最大+0.7%改善していることが確認できた。

速度評価　IT ドメイン翻訳の 2,000 件の評価セットに対する仮説選択に要した計算時間を表 2 に示す。Qe と Cbdt は効用に依らず常に 1,000 秒を切っているが、McMbr は効用に comet を用いた際に 1,000 秒を超え、chrF を用いた際に 10,000 秒を超えた。McMbr は各仮説ごとに複数参照評価を行うため二乗のオーダーを要するが、Qe は仮説数に対して線形に評価、Cbdt は復号時に品質を評価しないため、このような速度差が生まれたと考えられる。

事例分析　医療ドメインにおいて効用関数に chrF を用いたときの実際の出力例を表 3 に示す。表に示すとおり、McMbr-Cbdt では “*nausea*” を正しく出力できていることがわかる。ここで、メモリ内の事例を検索したところ、入力テキストに “*Übelkeit*” を含み、かつ、その参照テキストに “*nausea*” または “*Nausea*” を含む事例が 1061 件存在していた。一方で、入力テキストに “*Übelkeit*” を含むとき、その参照テキストに “*illness*” または “*Illness*” を含む事例は 9 件のみであった。したがって、“*Übelkeit*” が入力された際に “*nausea*” または “*Nausea*” を出力すると U^{Cbdt} スコアが高くなりやすいといえる。このことから、McMbr-Cbdt はメモリに含まれる類似事例の情報を利用して出力を決定していることがわかった。

5　おわりに

本稿では、事例に基づいて仮説を選択する CBDT 復号を提案し、ドメインにおけるテキスト生成の品質を改善した。5 つのドメインの翻訳実験より、提案法は従来の MAP 復号より高品質なテキストを生成できること、また、MBR 復号と組み合わせることで、MBR 復号を上回る品質となることを確認した。

今後は、他のテキスト生成タスクにおいても、提案法の有効性を検証していきたい。

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] Julius Cheng and Andreas Vlachos. Faster minimum Bayes risk decoding with confidence-based pruning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12473–12480, Singapore, December 2023. Association for Computational Linguistics.

[5] Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. Centroid-based efficient minimum Bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 11009–11018, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] David Heineman, Yao Dou, and Wei Xu. Improving minimum Bayes risk decoding with multi-prompt. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 22525–22545, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

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

[8] Anna Conte, John D Hey, and Peter G Moffatt. Mixture models of choice under risk. **Journal of Econometrics**, Vol. 162, No. 1, pp. 79–88, 2011.

[9] Shankar Kumar and William Byrne. 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**, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[10] Vyas Raina and Mark Gales. Minimum Bayes' risk decoding for system combination of grammatical error correction systems. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, **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)**, pp. 105–112, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[11] Itzhak Gilboa and David Schmeidler. Case-based decision theory. **The Quarterly Journal of Economics**, Vol. 110, No. 3, pp. 605–639, 1995.

[12] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, **Proceedings of the First Workshop on Neural Machine Translation**, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics.

[13] Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7747–7763, Online, July 2020. Association for Computational Linguistics.

[14] 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.

[15] Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mbrs: A library for minimum Bayes risk decoding. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 351–362, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[16] 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. Beyond english-centric multilingual machine translation. **J. Mach. Learn. Res.**, Vol. 22, No. 1, January 2021.

[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] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[19] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[20] 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. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[21] 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. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

A　実験設定の詳細

表 4: 独英ドメイン翻訳データの対訳文対数。訓練データは提案法のメモリの構築のみに使用。

ドメイン	訓練	評価
IT	222,927	2,000
コーラン	17,982	2,000
法律	467,309	2,000
医療	248,099	2,000
字幕	500,000	2,000

表 5: 実験設定の詳細。

項目	設定
CPU	Intel® Xeon® Gold 6426Y （8 コア）
GPU	NVIDIA RTX™ 6000 Ada （1 基）
翻訳モデル	`facebook/m2m_418M`
COMET モデル	`Unbabel/wmt22-comet-da`
KIWI モデル	`Unbabel/wmt22-cometkiwi-da`
入力間類似度	`sentence-transformers/LaBSE`
仮説間類似度	`sentence-transformers/LaBSE`

表 4 に実験に使用した独英ドメイン翻訳データの対訳文対数を、表 5 に実験設定の詳細を示す。

B　参考情報

表 6: IT ドメインの開発データにおいて、類似度の温度パラメータを変えたときの McMbr-Cbdt の翻訳品質。効用、評価指標には chrF を使用した。

	$\tau_{\mathcal{Y}}$		
$\tau_{\mathcal{X}}$	0.01	0.1	1.0
0.01	52.7	52.9	52.5
0.1	**53.9**	53.6	53.0
1.0	53.4	53.2	52.8

参考情報として、IT ドメインの開発データにおいてハイパーパラメータを変化させたときの実験結果を記す。表 6 に類似度の温度パラメータを変えたときの McMbr-Cbdt の翻訳品質を示す。また、図 4 に、復号時に用いる類似事例数 k を変えたときの、図 5 に、メモリ内の各事例が持つ仮説数 H を変えたときの McMbr-Cbdt の翻訳品質を、それぞれ示す。この結果より、$\tau_{\mathcal{X}} = 0.1$、$\tau_{\mathcal{Y}} = 0.01$、$k = 256$、$H = 256$ に設定した。

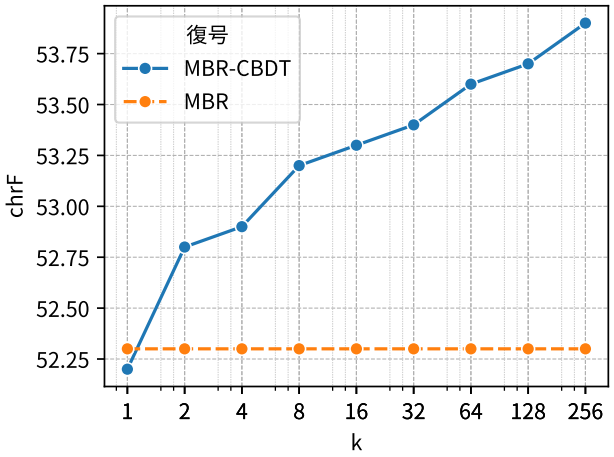


図 4: IT ドメインの開発データの翻訳において、用いる類似事例数 k を変化させたときの翻訳品質（chrF%）。

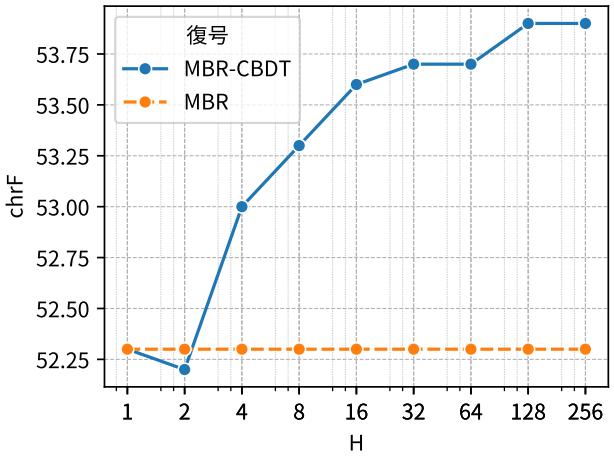


図 5: IT ドメインの開発データの翻訳において、メモリ内の各事例が持つ仮説数 H を変化させたときの翻訳品質（chrF%）。