多様で高品質な非自己回帰テキスト生成に向けた Transformer の逆推論と文のアナロジー

野坂 瞭太　　松崎 拓也
東京理科大学大学院
1424519@ed.tus.ac.jp　　matuzaki@rs.tus.ac.jp

概要

自己回帰モデルはテキスト生成の代表的な手法として発展を続けているが，出力の速度と多様性には改善の余地がある．有力な代替案として拡散モデルがあるが，生成結果の質は十分とは言い難い．本研究は，多様性と品質の両立を目標に，Transformer の特徴量抽出過程を遡ることでテキストの潜在表現からテキストを出力する新たな生成モデルを提案し，その活用法を探求する．結果として，拡散モデルと同等以上に高速で多様，かつ自然な言い換え文の生成を実現した．

1　はじめに

テキスト生成技術は，自己回帰に基づく手法が目覚ましい発展を遂げており，様々なタスクへの応用が進められているが，生成結果の多様性には課題が残されている．例えば，あるテキストに対して意味を保ちつつ語の選択や統語構造が異なるテキストを作る **言い換え生成** は，入力と出力が同一言語で，かつ意味が近いほど良いという特殊なタスクで，一般的な系列変換モデルは入力をほぼそのまま出力するような単純な解に陥ることがある [1]．より多様な生成結果を得るために，単語の予測分布に温度を導入することや，入力に存在する n-gram の出力を禁止する方法 [1] が考えられるが，文法的な正確性を損なうことがある．多様性は，生成の質とのトレードオフとなるのではなく，場合により許容される一定程度の意味の揺らぎから生まれることが望ましい．

拡散モデルは，画像生成分野において生成品質と多様性がともに高い手法として注目を集めている生成モデルで [2, 3]，テキスト生成に適用する研究も行われている [4, 5, 6]．拡散言語モデルは，拡散モデルで単語埋め込みの列を生成して離散化することで単語列を得る．長所として，非自己回帰手法であることによる生成速度も挙げられる．多様なテキスト生成技術の主な用途にデータ拡張があることを踏まえると，高い多様性と速度はいずれも好ましい性質である．一方，生成の品質は長らく発展途上にある．

本稿では，多様かつ安定した非自己回帰テキスト生成の実現を目指し，**双方向 Transformer [7] を用いたテキストエンコーダの特徴量抽出過程を逆向きに辿る** という新たなパラダイムについて検討する．意味的類似度を測るためのテキストエンコーダとして訓練された理想的な Transformer は，言い換え関係にあるテキストに対して同じベクトルを出力する．これは統語構造などの情報を捨てて意味情報のみを残す多対一の変換と捉えられるので，確率的に一対多の逆推論を行うモデルを用意できれば，言い換えたいテキストをエンコードして逆推論することで，その意味を表す様々なテキストが得られるだろう．また，Transformer の隠れ状態という文脈依存の単語埋め込み列を中間状態とする点で，文脈非依存の単語埋め込み列の改善に基づく拡散言語モデルよりも合理的な非自己回帰生成過程と解釈できる．

ただし，実際には言い換え関係にあるテキストに対して全く同一のベクトルを出力する Transformer を用意することは困難であり [8, 9]，従って逆推論による生成結果もそのままでは多様になりにくい．そこで，言い換えたいテキストと類似度が一定のベクトルをサンプリングする方法や，テキスト埋め込み同士の加減算によって，生成の種となる様々なベクトルを作成し，意味のずれに起因するさらなる多様性の向上を追求する．

2　背景：拡散言語モデル

拡散モデル [2, 3] は，はじめに訓練データ $\boldsymbol{x}_0$ に少しずつガウシアンノイズを付加して $\boldsymbol{x}_1, ..., \boldsymbol{x}_T$ を作り，その除去を学習する．このノイズ付加（拡散過程）はマルコフ過程であり，任意の t におけるノイ

ズ付きデータ $\boldsymbol{x}_t$ の分布は

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \bar{\beta}_t I\right)$$

と表される．$\bar{\alpha}_t = 1 - \bar{\beta}_t, 0 < \bar{\beta}_1 < \cdots < \bar{\beta}_T \approx 1$ はノイズの大きさを定めるハイパーパラメータである．その後，ランダムノイズ $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, I)$ をサンプリングし，学習した「ノイズ除去」を $t = T, T-1, ..., 1$ において順に行うことで，新たなデータ $\boldsymbol{x}_0$ を生成する（逆拡散過程）．拡散過程において画像は細部から不明瞭になるので，その逆を行う逆拡散過程は，サンプルを大まかな輪郭から徐々に生成する．

拡散言語モデルは，拡散モデルにより単語埋め込みの列 $\boldsymbol{x}_0$ を生成し，続く丸め機構で単語列を出力する[1] [4, 5, 6]．そのようなモデルは多様性や生成速度に優れるが，生成品質には課題を抱えている．原因の一つに，単語埋め込みをノイズ除去とともに学習することの難しさが指摘されている [10, 11]．また，根源的な疑問として，単語埋め込みにガウシアンノイズを付加することの妥当性も明らかではない．例えば，画像とは異なり，言語における「細部」や「輪郭」をモデリングしているとは考えにくい．

3 提案手法

まず，テキストエンコーダを非自己回帰生成に利用できるよう設定する．そして，新たな逆推論生成モデルを導入し，多様性向上のためのテキスト埋め込みの摂動について説明する．

3.1 テキストエンコーダ

単語列 $\boldsymbol{w} = [w_i]_{i=1}^N$ ($w_i \in \{1, 2, \ldots, V\}$) の埋め込み $\boldsymbol{c} \in \mathbb{R}^d$ ($\|\boldsymbol{c}\|^2 = 1$) を，双方向 Transformer の最後の隠れ状態を平均値プーリングして L_2 正規化を施すことで得る．単語列の長さ N は訓練・推論を通して固定し，それより短いものは右側にパディングを行う．パディングトークンも通常の単語と同様に扱い，さらに平均値プーリングを用いることで，全ての層・位置の隠れ状態がテキスト埋め込みに直接関与するため，出力長をパディングトークンの生成で調節する非自己回帰逆推論が可能になる．

SimCSE [12] に基づき，言い換えペアのテキストを正例，ミニバッチ内の他のテキスト対を負例とした対照学習を行う．この際，学習の強さを表すハイパーパラメータ（温度）τ を設定する [12]．

1) ここではいわゆる連続拡散言語モデルを取り上げる．離散拡散については 付録 A を参照されたい．

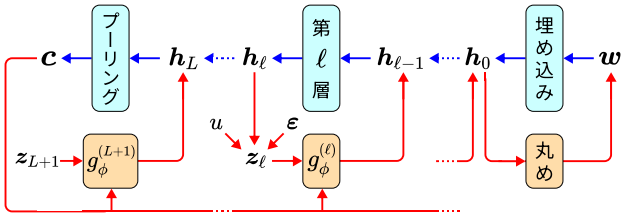


図 1 提案モデルの構造

3.2 逐次逆推論モデル

単語埋め込み列に位置埋め込みを加えたものを $\boldsymbol{h}_0$，テキストエンコーダの第 ℓ 層の隠れ状態を $\boldsymbol{h}_\ell$ とおき，$\boldsymbol{h}_\ell$ のうち i 番目のトークンに対応するベクトルを $\boldsymbol{h}_{\ell i}$ と書く．また，$\boldsymbol{h}_{L+1} = \boldsymbol{0}$ と定義する．逆推論モデル $g_\phi^{(\ell)}$ はテキスト埋め込み $\boldsymbol{c}$ とノイズ付き隠れ状態 z_ℓ から一層前の隠れ状態 $\boldsymbol{h}_{\ell-1}$ を予測する：

$$\hat{\boldsymbol{h}}_{\ell-1} = g_\phi^{(\ell)}(\boldsymbol{c}, z_\ell), \qquad z_{\ell i} = \boldsymbol{h}_{\ell i} + \|\boldsymbol{h}_{\ell i}\| u\boldsymbol{\varepsilon},$$
$$u \sim \mathrm{Uniform}(0, 1), \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, I).$$

一層の逆推論を $\ell = L+1, L, \ldots, 1$ について逐次行うことで $\boldsymbol{c}$ から $\boldsymbol{h}_0$ を求める（図 1）．予測の損失としては平均二乗誤差を用いる．

テキストエンコーダはテキストの表層的な情報を捨てていく過程であるので，その逆推論は，全体の意味から細かい言葉選びにかけて少しずつテキストを作る非自己回帰的な生成過程と解釈できる．

ノイズ $\|\boldsymbol{h}_{\ell i}\| u\boldsymbol{\varepsilon}$ は逆推論を確率過程にするための確率変数として働く．$\|\boldsymbol{h}_\ell\|$ は隠れ状態とノイズのスケールを合わせるための係数である．また，拡散モデルのように様々な分散のガウシアンノイズをデータに付加することは，学習の効率と生成の多様性に有効であると期待し，一様乱数 u を導入している．具体的には，$u\boldsymbol{\varepsilon}$ は Transformer の隠れ状態のような高い次元の空間における決定境界の平滑化に効果があると考えられる[2]．

実験では，逆推論モデル $g_\phi^{(\ell)}$ として双方向 Transformer を 1 つ用い，「ステップ埋め込み」を入力に追加した．また，テキスト埋め込み $\boldsymbol{c}$ から $\boldsymbol{w}$ を直接予測する形の逆推論モデルと比較を行った．

3.3 丸め機構

最初の隠れ状態 $\boldsymbol{h}_{0i}$ が単語 w に丸められる確率を全結合 $R_\phi : \mathbb{R}^d \to \mathbb{R}^V$ と Softmax で定義する：

$$p_\phi(w \mid \boldsymbol{h}_{0i}) = \mathrm{Softmax}_w R_\phi(\boldsymbol{h}_{0i}).$$

訓練では負の対数尤度 $-\log p_\phi$ を損失に加える．

2) 高次元ガウシアンノイズは主に球面上に分布し，球内の総確率は小さい．ランダム分散はその対策．付録 B 参照．

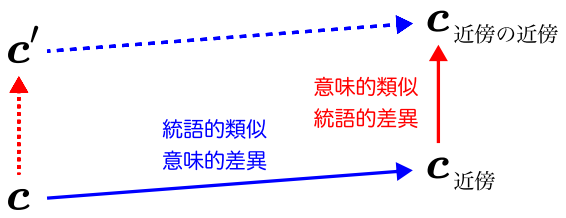


図 2　文のアナロジーの概念図

3.4　テキスト埋め込みの摂動

意味の揺らぎに由来する生成多様性を得るため，言い換えたいテキストの埋め込み $\boldsymbol{c}$ の代わりに，以下の方法で作成したベクトルをテキスト埋め込みとして逆推論モデルに入力することも試みる．

類似度指定サンプリング　テキスト埋め込み $\boldsymbol{c}$ と cos 類似度が ρ である単位ベクトルの集合 $\{\boldsymbol{c}' \in \mathbb{R}^d \mid \|\boldsymbol{c}'\|^2 = 1, \cos(\boldsymbol{c}, \boldsymbol{c}') = \rho\}$ からランダムにサンプリングする（方法は 付録 C）．

加法的文アナロジー　まず，あるデータセット中の各テキストを埋め込んだベクトル集合を用意する．このとき，あるテキスト埋め込みの近傍には，意味が近いものに次いで統語構造が類似するテキストが存在すると期待される．すなわち，言い換えたいテキストと意味が近いものがデータセットに存在しない場合，埋め込み空間で近傍探索を行うと，統語構造が似たテキストの埋め込み $\boldsymbol{c}_{\text{近傍}}$ が見つかる．もしも，$\boldsymbol{c}_{\text{近傍}}$ の近傍には意味が類似する「近傍の近傍」$\boldsymbol{c}_{\text{近傍の近傍}}$ があるならば，入力から近傍の成分を差し引いて「近傍の近傍」の成分を加えることで，入力の意味を保ちつつ「近傍の近傍」の統語構造に変換するという方針が考えられる（図 2）．具体的には $\boldsymbol{c}' := (\boldsymbol{c} - \boldsymbol{c}_{\text{近傍}} + \boldsymbol{c}_{\text{近傍の近傍}}) / \|\boldsymbol{c} - \boldsymbol{c}_{\text{近傍}} + \boldsymbol{c}_{\text{近傍の近傍}}\|$ を逆推論モデルに入力する．データセットの大きさの都合から，探索は NMSLIB [13] で近似的に行う．

4　実験

本節では，まず逆推論を介さない典型的な系列変換モデルで言い換え生成を行う場合（以降，単に「系列変換」と呼ぶ）の課題を実験結果に基づき整理する．続いて，提案モデルが実際にテキストエンコーダの逆推論を精度よく行えること，及び生成速度の優位性を示し，さらにテキスト埋め込みを揺らすことによる多様な言い換え生成の可能性を検証する．4.3 項では逆推論の入力となる埋め込みとして (1) テキストエンコーダの出力そのもの ($\rho = 1.00$)，(2) ρ が一定の点からサンプリングした点の両者を利用した．4.4 項ではアナロジーの方法を用いた．

4.1　設定

データセット　Quora Question Pairs (QQP) [14] と Microsoft COCO (MSCOCO) [15] を利用する．

評価指標　言い換え及び逆推論の精度を 入力テキストとの BERTScore [16], BLEU [17]，ROUGE [18] で測る．BERTScore は意味的な類似度を表し，大きいほど性能が高い．BLEU と ROUGE は単語列としての類似度であり，言い換えタスクでは BERTScore が同等のサンプル間ならば小さいほど，逆推論タスクとしては大きいほど良い．また，GPT-2 [19] と Qwen3-0.6B [20] の Perplexity を文法的な正確性の指標と見做して報告する．さらに，ある入力に対する生成結果の多様性を Self-BLEU[3] [21] で，生成可能な語句の多様性を Distinct-n[4] [22] で計測する．

テキストエンコーダ　12 層 BERT [23] を予めファインチューニングしたものを用いる．

ベースライン　系列変換と逆推論を行う標準的な自己回帰モデルと拡散モデルをそれぞれ用意する．

4.2　系列変換

表 1 に系列変換及び逆推論の評価結果を示す．まず系列変換の結果をまとめる．自己回帰モデルは，QQP ではデータセットの言い換え例（ゴールド）に近い BERTScore や BLEU を示したが，Self-BLEU が 92.87 と非常に高く多様性に欠けた．サンプリング温度 $\tilde{\tau}$ を上げると多様性が改善したが，品質の悪化が顕著で，好ましい多様性の制御とは言えない．具体的には，$\tilde{\tau} = 2.52$ のとき，Perplexity が拡散モデルと同等以上まで増加したが，Self-BLEU は拡散モデルに劣った．MSCOCO ではゴールドと比べ BLEU や ROUGE が高く，「言い換えモデルは入力をコピーしがち」という課題が確認できる．

拡散モデルは，自己回帰モデルよりも生成が高速で多様だった．しかし，BERTScore はゴールドを大きく下回った．また，Perplexity が大きく，文法性にも難があることがわかる．

4.3　逆推論

次に逆推論の結果についてまとめる．表 1 の通り，提案手法は概して他手法よりも高い BLEU と ROUGE を示し，逆推論を正確に行うことができた．テキストエンコーダの温度 τ を上げると，精度が低下し，多様性が向上した．温度を上げると言い換え

3)　同一の入力に対する複数サンプル間の BLEU.
4)　全 n-gram のうちユニークなものの割合.

表 1　主な結果．自己回帰モデルの $\tilde{\tau}$ は単語予測分布の温度，拡散モデルの S は逆拡散過程のステップ数．

			BERT Score	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity GPT-2	Perplexity Qwen3	Self-BLEU	Distinct-2	Distinct-4	生成時間
QQP	データセット	ゴールド	84.81	29.20	63.25	38.52	60.08	110.83	116.86	–	57.97	87.69	–
	系列変換	自己回帰 ($\tilde{\tau}=1.00$)	83.13	26.28	60.57	35.40	57.66	127.76	166.11	92.87	38.10	62.44	5 分 9 秒
		自己回帰 ($\tilde{\tau}=2.52$)	76.85	21.94	53.63	29.63	50.76	872.20	1271.16	36.55	52.42	82.17	10 分 2 秒
		拡散 ($S=13$)	79.88	28.45	62.80	37.58	60.06	709.56	1227.23	25.64	51.63	83.41	2 分 10 秒
	逆推論 ($\tau=0.05$)	自己回帰 ($\tilde{\tau}=1.00$)	92.75	68.41	86.08	74.79	84.08	112.30	123.88	92.39	42.52	75.69	5 分 22 秒
		拡散 ($S=13$)	90.91	68.33	85.22	73.72	83.92	319.94	957.64	73.41	47.52	80.42	1 分 15 秒
		提案手法 ($\rho=1.00$)	93.31	72.86	87.96	78.49	86.79	214.85	352.94	85.35	46.41	79.89	1 分 12 秒
	逆推論 ($\tau=0.20$)	提案手法 ($\rho=1.00$)	88.06	55.54	74.19	59.01	72.92	238.78	340.22	68.64	44.81	77.53	1 分 13 秒
		提案手法 ($\rho=0.98$)	84.35	43.62	66.45	47.76	64.68	265.93	347.89	40.34	46.18	79.90	1 分 14 秒
		提案手法 ($\rho=0.95$)	78.85	28.91	56.09	34.14	53.84	348.61	480.32	20.67	48.73	83.94	1 分 14 秒
MSCOCO	データセット	ゴールド	76.49	6.92	39.41	13.12	34.42	132.33	123.99	–	44.72	89.39	–
	系列変換	自己回帰 ($\tilde{\tau}=1.00$)	81.13	16.31	50.02	23.53	45.25	54.52	53.34	56.18	13.47	35.68	3 分 21 秒
		拡散 ($S=13$)	72.79	6.36	39.10	12.22	34.45	349.19	452.33	2.36	36.74	87.19	2 分 10 秒
	逆推論 ($\tau=0.20$)	自己回帰 ($\tilde{\tau}=1.00$)	83.14	22.19	56.06	30.23	51.18	58.62	55.90	66.72	16.16	46.46	3 分 29 秒
		提案手法 ($\rho=1.00$)	83.91	27.16	62.88	36.83	58.33	165.32	162.82	34.56	27.94	73.81	1 分 13 秒
		提案手法 ($\rho=0.99$)	76.55	13.64	49.20	21.95	44.30	354.16	296.99	9.25	37.65	85.69	1 分 14 秒

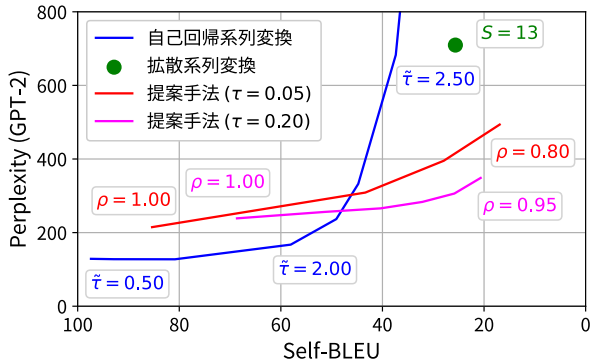


図 3　言い換えの度合いと文法性・多様性の関係 (QQP)

表 2　アナロジーによる言い換えの例 (QQP, $\tau=0.05$)

入力	How can I improve my Malay?
近傍	How could I improve my English?
近傍の近傍	How could I improve my English pronunciation?
生成結果	How can I improve my Malay pronunciation?
近傍	How can I improve English skills?
近傍の近傍	What is best way to learn English speaking?
生成結果	What is the best way to learn Malay?
近傍	Can I improve my English?
近傍の近傍	How we can learn good English speaking?
生成結果	How can I learn good Malay Malay?

関係にあるテキスト埋め込み同士が近づくため，自然な傾向である．生成時間は拡散モデルと同等以上に短かった．文法性は自己回帰モデルには及ばなかったものの，拡散モデルを上回った．また，自己回帰モデルが MSCOCO でゴールドを超える文法性を示した点には注意が必要である．Distinct-n が低く，出力が限られた定型句に偏っていた．

QQP において，逆推論の入力 $\boldsymbol{c}'$ の類似度 ρ を低下させると，Self-BLEU と Distinct-n が改善し，多様な言い換えを得ることができた．しかし，Perplexity の増加も見られ，完全に意味のずれのみに由来する多様性とはならなかったことが伺える．この結果は，有効な逆推論結果が得られない「種」ベクトルの存在，すなわち，テキスト埋め込みの非等方性を示唆している．ただし，図 3 が示すように，ρ の低下に伴う Perplexity の悪化は自己回帰モデルの温度を上昇させるよりはなだらかで，拡散モデルに勝る Self-BLEU と Perplexity の両立を実現した．

MSCOCO では，QQP よりも逆推論の精度が低く，多様性が高い傾向があった．MSCOCO はゴールドの BLEU 等が小さく，言い換えとして粗いデータセットであるため，テキストエンコーダがテキスト埋め込みに残した情報が少なかったと考えられる．

4.4　文のアナロジー

表 2 は提案モデル (QQP, $\tau=0.05$) と加法的文アナロジーの方法で “How can I improve my Malay?” を言い換えた結果である．これに近い意味のテキストはデータセットになく，近傍には代わりに “English” に関連するテキストが多く存在した．結果として，特に近傍の統語構造が入力と近い場合，アナロジーを実現することができた．ただし，一部を簡略化したり，近傍と入力の構造が似ていない場合に非文法的なテキストを生成したりしてしまうことがあった．

5　おわりに

本研究は，多様で高品質な非自己回帰テキスト生成を目標に，Transformer の逆推論に基づく手法を提案した．展望として，より妥当なテキスト埋め込みの摂動や，他のタスクへの応用がある．

謝辞

本研究の一部は，キオクシア株式会社の支援をうけて実施したものです．

参考文献

[1] Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. Unsupervised Paraphrasing with Pretrained Language Models. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5136–5150. Association for Computational Linguistics, 2021.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In **The Ninth International Conference on Learning Representations**, 2021.

[4] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM Improves Controllable Text Generation. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 4328–4343. Curran Associates, Inc., 2022.

[5] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In **The Eleventh International Conference on Learning Representations**, 2023.

[6] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Text Diffusion Model with Encoder-Decoder Transformers for Sequence-to-Sequence Generation. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 22–39. Association for Computational Linguistics, 2024.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[8] John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. Text Embeddings Reveal (Almost) As Much As Text. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12448–12460. Association for Computational Linguistics, 2023.

[9] Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language Models are Injective and Hence Invertible, 2025.

[10] Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Empowering Diffusion Models on the Embedding Space for Text Generation. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 4664–4683. Association for Computational Linguistics, 2024.

[11] Ryota Nosaka and Takuya Matsuzaki. Timestep Embeddings Trigger Collapse in Diffusion Text Generation. In **Proceedings of the 29th Conference on Computational Natural Language Learning**, pp. 397–406. Association for Computational Linguistics, 2025.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910. Association for Computational Linguistics, 2021.

[13] Leonid Boytsov and Bilegsaikhan Naidan. Engineering Efficient and Effective Non-metric Space Library. In **Similarity Search and Applications – 6th International Conference, SISAP 2013**, Vol. 8199 of **Lecture Notes in Computer Science**, pp. 280–293. Springer, 2013.

[14] DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. Quora Question Pairs, 2017. Kaggle.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In **Computer Vision – ECCV 2014, Part V**, pp. 740–755. Springer International Publishing, 2014.

[16] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **The Eighth International Conference on Learning Representations**, 2020.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318. Association for Computational Linguistics, 2002.

[18] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81. Association for Computational Linguistics, 2004.

[19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[20] Qwen Team. Qwen3 Technical Report, 2025.

[21] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A Benchmarking Platform for Text Generation Models. In **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**, pp. 1097–1100. Association for Computing Machinery, 2018.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119. Association for Computational Linguistics, 2016.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[24] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. In **Advances in Neural Information Processing Systems**, Vol. 34, pp. 17981–17993. Curran Associates, Inc., 2021.

[25] Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4521–4534. Association for Computational Linguistics, 2023.

[26] Inception Labs, Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, Aditya Grover, and Volodymyr Kuleshov. Mercury: Ultra-Fast Language Models Based on Diffusion, 2025.

[27] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion Large Language Models, 2025.

[28] Google DeepMind. GeminiDiffusion, 2025.

A　関連研究：離散拡散言語モデル

単語埋め込みとガウシアンノイズを用いる連続拡散言語モデルに対し，単語を他の単語やマスクトークンに直接置換することを「ノイズ」と定義する離散拡散言語モデルも提案されている [24, 25]．連続拡散と同様に生成速度の高さが長所である．また，離散拡散は生成品質が比較的安定しており，近年は大規模に訓練されたモデルが登場したことで注目を集めている [26, 27, 28]．一方で，連続拡散とは異なり，生成多様性に関する報告は乏しい．

B　高次元ノイズ

異なるテキスト A, B のある隠れ状態 $\boldsymbol{h}^{\mathrm{A}}_{\ell i}, \boldsymbol{h}^{\mathrm{B}}_{\ell j}$ について考える．逆推論モデルはこれらにノイズを付加した $\boldsymbol{z}^{\mathrm{A}}_{\ell i}, \boldsymbol{z}^{\mathrm{B}}_{\ell j}$ からそれぞれ $\boldsymbol{h}^{\mathrm{A}}_{\ell-1,i}, \boldsymbol{h}^{\mathrm{B}}_{\ell-1,j}$ を予測する．

ガウシアンノイズはモデルの頑健性向上を目的として訓練データに付加するノイズによく用いられる．しかし，言語モデルの隠れ状態のような高次元空間においては，超球面上に薄く分布するために（図 4），学習する領域を複雑化することになる（図 5a）．対して提案ノイズは，超球面の内部にほぼ均一に分布するので（図 4），高次元空間においても決定境界の学習を自然に助けることができると考えられる（図 5b）．

提案手法では生成の過程においてもノイズを付加するが，これは平滑化した決定境界とともに生成多様性に寄与すると見込んでいる．図 5b のように，隠れ状態 $\boldsymbol{h}^{\mathrm{A}}_{\ell i}, \boldsymbol{h}^{\mathrm{B}}_{\ell j}$ があり，$\boldsymbol{z}^{\mathrm{A}}_{\ell i}, \boldsymbol{z}^{\mathrm{B}}_{\ell j}$ がサンプリングされる範囲に重なる部分があるとする．このとき，前ステップの予測値が $\boldsymbol{h}^{\mathrm{A}}_{\ell i}$ だった場合，モデルは $\boldsymbol{h}^{\mathrm{A}}_{\ell i}$ を $\boldsymbol{h}^{\mathrm{B}}_{\ell j}$ と捉えて $\boldsymbol{h}^{\mathrm{B}}_{\ell-1 j}$ を出力することが可能になる．

C　類似度指定サンプリング

集合 $\{\boldsymbol{c}' \in \mathbb{R}^d \mid \|\boldsymbol{c}'\|^2 = 1, \cos(\boldsymbol{c}, \boldsymbol{c}') = \rho\}$ からのランダムサンプリングは，等方性を持つ標準ガウシアン $\boldsymbol{\varepsilon}$ の $\boldsymbol{c}$ と直交する成分 $\boldsymbol{\varepsilon} - \boldsymbol{c}^\top \boldsymbol{\varepsilon} \boldsymbol{c}$ を $\boldsymbol{c}$ と混合させることで行う．すなわち，

$$\boldsymbol{c}' = \rho \boldsymbol{c} + \sqrt{1-\rho^2} \frac{\boldsymbol{\varepsilon} - \boldsymbol{c}^\top \boldsymbol{\varepsilon} \boldsymbol{c}}{\|\boldsymbol{\varepsilon} - \boldsymbol{c}^\top \boldsymbol{\varepsilon} \boldsymbol{c}\|}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, I).$$

D　実験設定詳細

Quora Question Pairs (QQP) [14] は Quora 上に投稿された質問の重複を集めたデータである．DiffuSeq [5] の分割を使用し，訓練データは約 14 万ペアである．Microsoft COCO (MSCOCO) [15] は画像キャプション生成のためのデータセットで，1 画像あたり約 5 つのキャプションがある．我々は，これらのキャプションを言い換えテキストの集合と見做し，約 120 万ペアを収集して訓練データとした．テストデータはともに 2,500 ペアである．

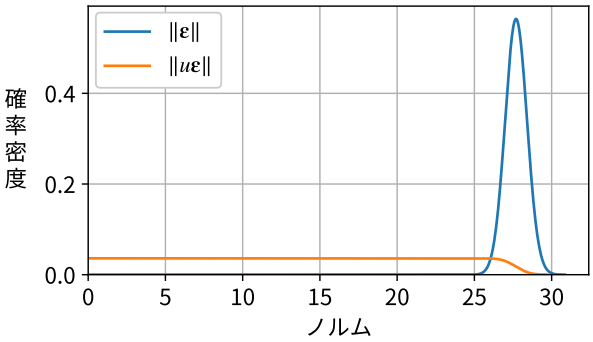
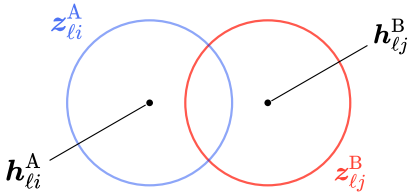
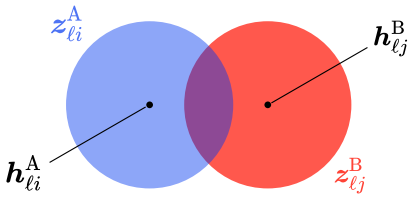


図 4　$\|\boldsymbol{\varepsilon}\|$ と $\|u\boldsymbol{\varepsilon}\|$ の確率密度 ($d = 768$)



(a) $\boldsymbol{\varepsilon}$ の場合



(b) $u\boldsymbol{\varepsilon}$ の場合

図 5　ノイズ付き隠れ状態のサンプリング範囲

系列長 N は 64 とする．逆推論モデル g_ϕ はテキストエンコーダと同等の双方向 Transformer を用いる．自己回帰系列変換モデルには一般的な Encoder-Decoder 型 Transformer を利用する．自己回帰逆推論はテキスト埋め込みを長さ 1 の系列として Encoder を使わずクロスアテンションする．拡散モデルは DiffuSeq [5] をベースとしつつ，ノイズ除去モデルを提案手法の逆推論モデルと同等の設定に調整する．ノイズスケジュールは sqrt [4]，拡散ステップ数は 2,000 とする．逆拡散過程のスキップは等間隔に行う．いずれのモデルも乱数で初期化する．

生成は，1 台の NVIDIA RTX A6000 を使い，各入力に対して 5 回行う．つまり，1 実験あたり $2,500 \times 2 \times 5 = 25,000$ テキストを生成し評価する．