Transformer 言語モデルの予測ヘッド内バイアスによる頻度補正効果

小林悟郎 [1,2]　栗林樹生 [1,3]　横井祥 [1,2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所　[3]Langsmith 株式会社
goro.koba@dc.tohoku.ac.jp　{kuribayashi, yokoi, kentaro.inui}@tohoku.ac.jp

概要

近年 Transfrormer ネットワークを採用した言語モデルが大きな成功を収め，注意機構やフィードフォワードネットを中心に分析が盛んに行われてきた．本研究ではモデルの出口部分にあたり，出力に直接作用する予測ヘッドの働きを分析する．実験から予測ヘッド内のバイアスが単語予測を頻度補正していることを明らかにする．具体的には，このバイアスが高頻度語の予測確率を上げ，低頻度語の予測確率を下げることで，予測分布を実際の単語頻度分布に近づけていることが観察された．さらにこの知見を応用し，バイアスを制御することで言語モデルに多様かつ人間に近い文章生成を促せることを示す．

1　はじめに

Transformer ネットワーク [1] を採用した言語モデル（Transformer 言語モデル）[2, 3, 4] は今や自然言語処理分野全体を支える基盤技術であり，高品質な文章生成を通じて幅広い応用を支えている．さらに，その成功理由の解明や更なる性能改善を求め，内部機序の分析も盛んに行われている [5]．分析対象としてこれまで特に注目を浴びてきたのはアーキテクチャの中核をなす Transformer 層であり，注意機構の重みの観察からはじまり層正規化やフィードフォワードネットの挙動解明に至るまで，知見が順調に蓄積されてきた [6, 7, 8]．

本研究では，Transformer 言語モデルの出口部分，すなわちすべての Transformer 層を出たあとの**予測ヘッド**の働きを分析する（図 1）．予測ヘッドはモデルの単語予測に直接的に作用するため解釈もしやすく，また下層での処理をまとめて上書きできるという点でモデル改善に接続しやすいと期待できる．

実験の結果，**BERT および GPT-2 の予測ヘッド内のバイアスが，高頻度語の予測確率を上げ，低頻度語の予測確率を下げていることが分かった．**さらにこの知見を応用し，バイアスの効果を抑制しながら文章生成を行うことで，モデルが典型的な（高頻度な）表現ばかりを生成する問題 [9, 10] が軽減し，多様かつ人間に近い文章生成を促せることを示した．

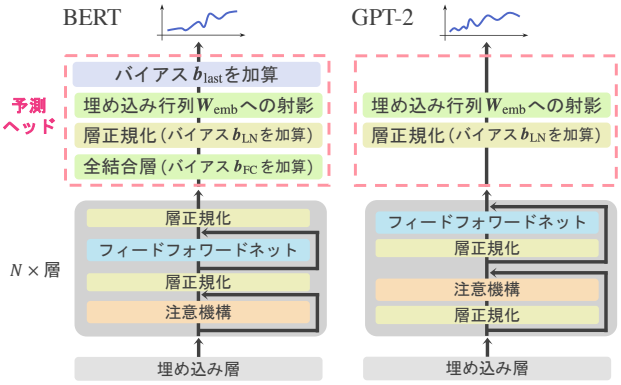

図 1　BERT および GPT-2 の構造概要．

2　準備：言語モデルの予測ヘッド

Transformer 言語モデルは埋め込み層に始まり，モデルの主要部分は「層」と呼ばれる同じ構造を積み重ねたネットワークである（図 1）．各層は注意機構などを通じて隠れ表現を更新していく．

層を積み重ねた後には**予測ヘッド**があり，これが本研究の分析対象である．なお本研究では，事前学習で得られた，最終層から隠れ表現を受け取って各単語の予測確率を算出するヘッドのことを予測ヘッドと呼ぶ．具体的には，隠れ表現 $\boldsymbol{x} \in \mathbb{R}^d$ を受け取り，層正規化（LN）[11] を適用してから，埋め込み層でも参照する単語埋め込み行列 $\boldsymbol{W}_{\text{emb}} \in \mathbb{R}^{d \times |\mathcal{V}|}$ へ射影することで全語彙数 $|\mathcal{V}|$ 個分の予測確率を計算する．ただし，BERT では層正規化の前に全結合層（FC）を適用し，さらに埋め込み行列への射影後にはバイアスパラメータ $\boldsymbol{b}_{\text{last}} \in \mathbb{R}^{|\mathcal{V}|}$ を加算する．

以下，詳細を述べる．GPT-2 では入力最後尾の単語に対応する隠れ表現 $\boldsymbol{x}$ を受け取り，以下のように

次単語の予測確率分布 $\boldsymbol{p} \in \mathbb{R}^{|\mathcal{V}|}$ を計算する[1]：

$$\boldsymbol{p} = \mathrm{softmax}\Big(\mathrm{LN}(\boldsymbol{x})\boldsymbol{W}_{\mathrm{emb}}\Big) \tag{1}$$

$$\mathrm{LN}(\boldsymbol{x}) := \frac{\boldsymbol{x} - m(\boldsymbol{x})}{s(\boldsymbol{x})} \odot \boldsymbol{\gamma} + \boldsymbol{b}_{\mathrm{LN}} \in \mathbb{R}^d \tag{2}$$

BERT では [MASK] に対応する隠れ表現 $\boldsymbol{x}$ を受け取り，活性化関数 GELU [12] を含んだ全結合層も用いて穴埋め単語の予測確率分布 $\boldsymbol{p} \in \mathbb{R}^v$ を計算する：

$$\boldsymbol{p} = \mathrm{softmax}\Big(\mathrm{LN}(\boldsymbol{x}')\boldsymbol{W}_{\mathrm{emb}} + \boldsymbol{b}_{\mathrm{last}}\Big) \tag{3}$$

$$\boldsymbol{x}' = \mathrm{GELU}\big(\boldsymbol{x}\boldsymbol{W}_{\mathrm{FC}} + \boldsymbol{b}_{\mathrm{FC}}\big) \in \mathbb{R}^d \tag{4}$$

ここで，$m(\boldsymbol{x})$ および $s(\boldsymbol{x})$ はそれぞれ要素での平均と標準偏差を指し，$\odot$ は要素積を表す．また，$\boldsymbol{\gamma} \in \mathbb{R}^d$ および $\boldsymbol{W}_{\mathrm{FC}} \in \mathbb{R}^{d\times d}$ は学習可能な重みパラメータ，$\boldsymbol{b}_{\mathrm{LN}}, \boldsymbol{b}_{\mathrm{FC}} \in \mathbb{R}^d$ および $\boldsymbol{b}_{\mathrm{last}} \in \mathbb{R}^{|\mathcal{V}|}$ は学習可能なバイアスパラメータを表す．以上のように，BERT と GPT-2 の予測ヘッドは共通してバイアス $\boldsymbol{b}_{\mathrm{LN}}$ を持ち，BERT はさらに $\boldsymbol{b}_{\mathrm{FC}}$ と $\boldsymbol{b}_{\mathrm{last}}$ を持つ．

本研究では Transformer 言語モデルの予測ヘッドが単語予測へ与える影響を分析する．今回は分析の第一歩として，加算という単純な操作で作用するために分析が容易なバイアスパラメータに注目する．

3 実験

予測ヘッド内のバイアスが単語予測に与える影響を分析する．3.1 節ではバイアスがモデルの単語予測分布を実際の単語頻度分布に近づけるような頻度補正を行なっていることを明らかにする．3.2 節ではバイアスの頻度補正効果を抑制する簡単な方法で多様かつ人間に近い文章生成を促せることを示す．

モデル　事前学習済みの BERT-cased [2] と GPT-2 [3] を対象とした．BERT は base と large の 2 種類，GPT-2 は small, medium, large, xl の 4 種類を用いた．

データ　モデルへ入力するテキストとして GPT-2 の事前学習コーパスである OpenWebText のテストセットの一部 5000 系列を用いた[2]．BERT では事前学習の設定に従い，12%[3] のトークンを [MASK] に置換して入力した．GPT-2 で文章生成を行う 3.2 節では計算コストの観点から，5000 系列からランダムにサンプリングした 100 系列のみを用いた．また分析時に用いる単語コーパス頻度は，BERT および GPT-2 それぞれの学習コーパスで算出した[4]．

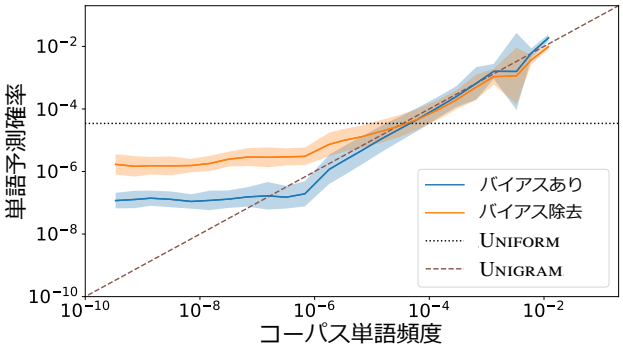

図 2　BERT（base）におけるバイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去による単語予測分布の変化．

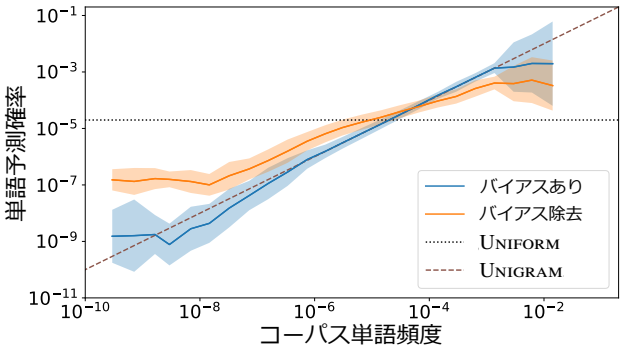

図 3　GPT-2（small）におけるバイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去による単語予測分布の変化．

3.1 バイアスが予測分布へ与える影響

BERT および GPT-2 の予測ヘッド内の各バイアスが単語予測に与える影響を調べる．各バイアスを除去して出力される単語予測分布を通常のモデルが出力した単語予測分布と比較する．

結果　BERT (base) でバイアス $\boldsymbol{b}_{\mathrm{LN}}$ を除去した際の単語予測分布の変化を図 2 に示す．図の横軸はコーパスで算出した実際の単語頻度，縦軸はモデルの単語予測確率である．コーパス単語頻度側で一定間隔にビンを分け，各ビンでの単語予測確率の幾何平均と幾何標準偏差をプロットした．バイアスを除去したところ，高頻度語（図右側）の予測確率が下がり，低頻度語（図左側）の予測確率が底上げされ，結果的に単語予測分布は平坦（図 2 中 Uniform 直線）に近づいた．言い換えれば，バイアス $\boldsymbol{b}_{\mathrm{LN}}$ は，高頻度語を予測しやすく低頻度語を予測しにくくすることで，単語予測分布を実際の単語頻度分布（図 2 中 Unigram 直線）に近づけていることがわ

1）本稿ではベクトルは横ベクトルとしている．
2）https://github.com/openai/gpt-2-output-dataset で公開されている webtext.test を用いた．
3）BERT の事前学習では入力系列の 15%が選ばれ，そのうち 80%（全体の 12%）が [MASK] トークンに置換される．
4）BERT の学習コーパスは Wikipedia と BooksCorpus [13]，GPT-2 の学習コーパスは OpenWebText である．それぞれ https://github.com/huggingface/datasets で一般公開されているデータセットを用いて再現し，単語頻度を算出した．

表 1　各バイアス除去時のモデル単語予測分布とコーパス単語頻度分布の KL ダイバージェンス．値が大きいほど，単語予測分布がコーパス単語頻度分布と離れていることを表す．$\boldsymbol{b}_{\mathrm{FC}}$ と $\boldsymbol{b}_{\mathrm{bias}}$ は BERT にのみ存在する．

モデル	除去なし	$\boldsymbol{b}_{\mathrm{LN}}$ 除去	$\boldsymbol{b}_{\mathrm{FC}}$ 除去	$\boldsymbol{b}_{\mathrm{last}}$ 除去
BERT				
base	0.20	**0.39**	0.22	0.23
large	0.21	**0.39**	0.23	0.23
GPT-2				
small	0.14	**0.83**	-	-
medium	0.14	**0.34**	-	-
large	0.14	**0.17**	-	-
xl	0.14	**0.17**	-	-

かった．この結果は単方向言語モデル（GPT-2）でも一貫した．GPT-2 (small) の結果を図 3，その他の結果を付録 A に示す．

各バイアスによる頻度補正を定量的に観察するため，検証した全てのモデルについて，モデルの単語予測分布とコーパスから算出した実際の単語頻度分布（Unigram）の KL ダイバージェンスを測った（表 1）．どのモデルでもバイアス $\boldsymbol{b}_{\mathrm{LN}}$ を除去すると値が大きくなることから，$\boldsymbol{b}_{\mathrm{LN}}$ が単語予測分布を実際の単語頻度分布に近づける頻度補正はモデルによらず行われていることがわかった．なお，BERT における $\boldsymbol{b}_{\mathrm{FC}}$ および $\boldsymbol{b}_{\mathrm{last}}$ も同様の頻度補正を行なっていたが $\boldsymbol{b}_{\mathrm{LN}}$ に比べて影響が小さかったため，これ以後は $\boldsymbol{b}_{\mathrm{LN}}$ に絞って分析・検証を行う．

考察：層の外で頻度補正する戦略　単語頻度分布の情報が予測ヘッドにある程度外出しされていることから，各 Transformer 層は頻度情報に表現力を割かずに頻度以外の意味について処理をおこなう傾向にあると考えられる．実際，Transformer 言語モデルの層は入力側から順に表層・構造・意味を処理していると報告されている [14]．例えば高頻度な名詞である “*valley*” と低頻度な類義語 “*dale*” は，品詞理解の上でも構文解析の上でも大雑把な意味理解の上でも見分ける必要がなく，頻度情報を一旦考えないのはむしろ効果的だと考えられる．また，表 1 ではモデルサイズが大きくなるほどバイアス除去前後の KL ダイバージェンスの変化は小さく，つまり，層の表現力および次元数が低いモデルほど頻度情報を予測ヘッドに任せており，これは限られた表現力を別の種類の言語処理に割いているのだろうという考察と一貫する．

考察：単語埋込空間における「頻度方向」の存在　Geva ら [15] に倣い，バイアスパラメータ $\boldsymbol{b}_{\mathrm{LN}}$ を埋め込み行列 $\boldsymbol{W}_{\mathrm{emb}}$ へ直接射影し語彙数 $|\mathcal{V}|$ 個のスコアを算出することで（$\boldsymbol{b}_{\mathrm{LN}}\boldsymbol{W}_{\mathrm{emb}} \in \mathbb{R}^{|\mathcal{V}|}$ を観察することで），バイアスパラメータが単語予測分布に与える影響を近似的に確認した．結果，単語頻度と「バイアスと単語埋込の内積」との間に比例関係が観察された（図 4）．バイアスパラメータが高頻度語の予測を強め低頻度語の予測を弱める役割を持つという本稿の主張を補強する結果と言える．この結果を幾何的に解釈すると，単語埋込空間（$\boldsymbol{W}_{\mathrm{emb}}$ をなす各単語ベクトルが入っている空間）の特定のわずか 1 次元の方向（$\boldsymbol{b}_{\mathrm{LN}}$ 方向）に頻度情報がエンコードされていると言える．これまでも，静的な埋込や系列変換器などさまざまなモデルで，頻度の近い単語群が埋込空間上で偏在することが報告されてきた [16, 17, 18]．Transformer 言語モデルでも同様の現象が起きていると考えられる．

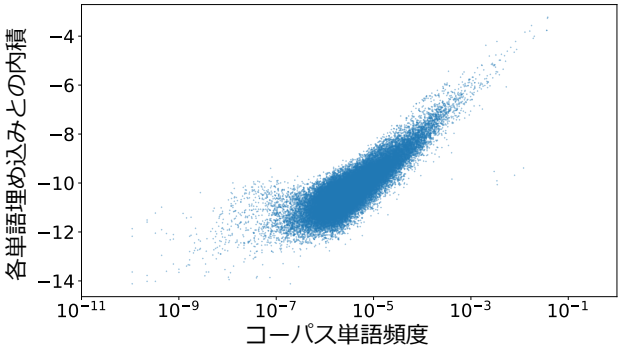

図 4　GPT-2 (small) においてバイアス $\boldsymbol{b}_{\mathrm{LN}}$ を埋め込み行列 $\boldsymbol{W}_{\mathrm{emb}}$ に射影した結果．

3.2　バイアスが文章生成に与える影響

前節では予測ヘッド内のバイアスが単語予測を実際の単語頻度に近づけていることが分かった．本節では推論時にこのバイアスを制御することで生成される文章の質を改善できる可能性を示す．

手順　バイアス $\boldsymbol{b}_{\mathrm{LN}}$ を制御した GPT-2 で文章生成し，その品質を評価する．具体的には，新たに係数 $\lambda \in [0, 1]$ を導入し，予測ヘッドの $\boldsymbol{b}_{\mathrm{LN}}$ を $\lambda\boldsymbol{b}_{\mathrm{LN}}$ に置き換える．λ を 0.1 刻みで変化させながら文章を生成させた．デコーディングには 5 種類の方法を用いた[5)]が，スペースの都合上，サンプリングに基づくデコーディングでの結果について述べる．ただし，本稿での主張は 5 種類全てで支持されている．詳細設定は付録 B に示す．

5)　貪欲探索，ビーム探索，サンプリング，Top-k サンプリング，Top-p サンプリングの 5 種類を用いた．

表 2　GPT-2 のバイアスを制御した際の予測および生成されたテキストの評価．$\lambda = 0, 1$ と顕著だった λ での結果．

モデル	λ	多様性↑			人間らしさ	
		D	D_1	D_2	MAUVE ↑	PPL ↓
small	1	0.77	0.29	0.82	**0.76**	**19.4**
	0.5	**0.87**	**0.53**	**0.96**	0.07	27.0
	0	0.73	0.44	0.81	0.03	65.9
med.	1	0.78	0.31	0.85	**0.77**	**14.6**
	0.3	**0.88**	0.56	**0.96**	0.25	17.8
	0	0.87	**0.58**	0.95	0.08	21.3
large	1	0.74	0.26	0.77	0.79	**12.7**
	0.7	0.76	0.29	0.80	**0.93**	12.8
	0	**0.81**	**0.41**	**0.89**	0.59	13.6
xl	1	0.74	0.25	0.76	0.90	**11.4**
	0.5	0.78	0.32	0.84	**0.94**	11.6
	0	**0.82**	**0.41**	**0.90**	0.68	12.1

評価　生成された文章は多様性と人間らしさの 2 つの観点で評価する．多様性の評価には Distinct-n (D_n) [19] と N-gram diversity (D) [20] を用いた．これらは生成された文章における N-gram 重複度を測定する評価指標であり，以下のように計算される．

$$D_n = \frac{\#\ \text{ユニークな n-gram}}{\#\ \text{生成された全ての n-gram}},\ D = \frac{1}{4}\sum_{n=1}^{4} D_n \quad (5)$$

人間らしさの評価には MAUVE [21] を用いた．MAUVE は人間が生成した文集合とモデルが生成した文集合を文埋込空間上の点群として比較することで，モデルが生成する文の分布が人間と近いかを評価する．加えて，モデルの単語レベルの予測傾向が人間と近いかを評価するため Perplexity (PPL) でも評価した．

結果　表 2 に各モデルが生成した文章の評価結果を示す．どのサイズの GPT-2 でもバイアス $\boldsymbol{b}_{\mathrm{LN}}$ を弱める（$\lambda < 1$）ことで，生成される文章の多様性は上がったが，同時に PPL が悪化し，両者はトレードオフの関係にあった．一方で，サイズの大きい 2 モデルでは PPL の悪化が微小なのにも関わらず，多様性および MAUVE が改善する λ があった．すなわち，バイアス $\boldsymbol{b}_{\mathrm{LN}}$ の補正効果を係数 λ で抑制する簡単な操作が，言語モデルに多様かつ人間に近い文章生成を促しうることがわかった．ただし，これらの指標のみから多様な側面を持つ文章品質の全体像を完全に把握するのは不可能であり，たとえば人手による流暢性の評価などが今後の展望として残されている．

考察：ロジット補正法との関係　本研究では，Transformer 言語モデルの出口付近のバイアス $\boldsymbol{b}_{\mathrm{LN}}$ が単語予測分布を実際の単語頻度分布に近づけていることを明らかにし，また $\boldsymbol{b}_{\mathrm{LN}}$ の効果を抑制する操作が文章生成の制御・改善につながる可能性を示した．この一連の知見は Menon ら [22] が提案したロジット補正法と非常に類似している．ロジット補正とは，クラス不均衡な分類タスクにおいてモデルが高頻度クラスばかりを予測するよう学習されてしまう問題をロジットの調整を通じて緩和する技術である．彼らが提案した手法の 1 つは，訓練時にはロジットにクラス頻度分布を加算して予測させ，推論時には加算せず予測させることで，均衡誤差[6] の最小化を目指すものである．本研究の知見と照らし合わせると，予測ヘッドでの頻度補正バイアス $\boldsymbol{b}_{\mathrm{LN}}$ の加算はロジットにクラス頻度分布を加算する操作と一貫しており，推論時にこれを除去することが生成分布の多様性につながる点も同様である．頻度分布を外から与えるという工夫を加えていないにも関わらず，Transformer 言語モデルが暗に均衡誤差の最小化に近い形で学習されていることは興味深い．

4　関連研究

Transformer 言語モデルの分析のスコープとしては，1 節でも述べたように Transformer 層が注目を集めてきた．また，特に位置埋め込みなどモデルの入り口部分である埋込層にも関心が寄せられている [23, 24]．我々は分析領域を予測ヘッドに広げ，その機序について新たな発見を提供した．予測ヘッドは層正規化を含むなど表現力が高く複雑な処理が可能であり（式 2），さらに単語予測に直接作用できるため積み重ねた Transformer 層の処理の上書きすら可能な魅力的な分析対象と言える．

5　おわりに

本稿では，Transformer 言語モデルの予測ヘッドを構成するバイアスが予測の頻度補正を行い，予測分布をコーパス単語頻度分布に近づけていることを明らかにした．さらに，このバイアスの効果を抑制しながら文章生成させることで，生成される文章の多様性や自然さを改善できる可能性を示した．一連の知見は，単語埋込空間における幾何やロジット補正法との関連など，Transformer 言語モデルが持つ性質について興味深い示唆を提供する．

今後は，OPT [25] などのより巨大な言語モデルでも分析を行いたい．また，今回は対象外とした予測ヘッド内の重みパラメータの分析も興味深い．

6)　クラス頻度によらず，各クラスを平等に捉えた誤差．

謝辞

言語モデルが生成した文章の評価方法についてアドバイスをくださった東北大学の佐藤志貴さんに感謝申し上げます．本研究は JSPS 科研費 JP22J21492, JP22H05106, JP22H03654 の助成を受けたものです．また本研究は JST, ACT-X, JPMJAX200S および JST, CREST, JPMJCR20D2 の支援を受けたものです．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems 30 (NIPS)**, pp. 5998–6008, 2017.

[2] 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 (NAACL-HLT)**, pp. 4171–4186, 2019.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Advances in Neural Information Processing Systems 33 (NeurIPS)**, Vol. 33, pp. 1877–1901, 2020.

[5] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. **Transactions of the Association for Computational Linguistics (TACL)**, Vol. 8, pp. 842–866, 2021.

[6] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What Does BERT Look At? An Analysis of BERT's Attention. In **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, 2019.

[7] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4547–4568, 2021.

[8] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 8493–8502, 2022.

[9] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In **8th International Conference on Learning Representations (ICLR)**, 2020.

[10] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural Text Generation With Unlikelihood Training. In **8th International Conference on Learning Representations (ICLR)**, 2020.

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. **arXiv preprint arXiv:1607.06450**, 2016.

[12] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). **arXiv preprint arXiv:1606.08415**, 2016.

[13] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In **The IEEE International Conference on Computer Vision (ICCV)**, 2015.

[14] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 4593–4601, 2019.

[15] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5484–5495, 2021.

[16] Jiaqi Mu and Pramod Viswanath. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In **6th International Conference on Learning Representations (ICLR)**, 2018.

[17] Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Frage: Frequency-agnostic word representation. In **Advances in Neural Information Processing Systems 31 (NeurIPS)**, pp. 1341–1352, 2018.

[18] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 1882–1892, 2020.

[19] 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 (NAACL-HLT)**, pp. 110–119, 2016.

[20] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally Typical Sampling. **arXiv preprint arXiv:2202.00666v4**, 2022.

[21] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In **Advances in Neural Information Processing Systems 34 (NeurIPS)**, 2021.

[22] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In **9th International Conference on Learning Representations (ICLR)**, 2021.

[23] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. On position embeddings in BERT. In **9th International Conference on Learning Representations (ICLR)**, 2021.

[24] Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. SHAPE: Shifted absolute position embedding for transformers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 3309–3321, 2021.

[25] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models. **arXiv preprint arXiv:2205.01068v4**, 2022.

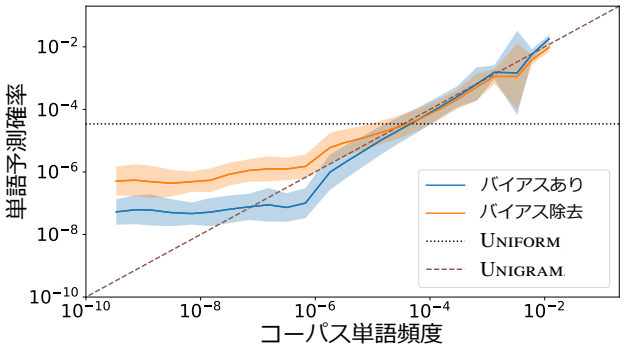

図 5　BERT (large) におけるバイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去による単語予測分布の変化.

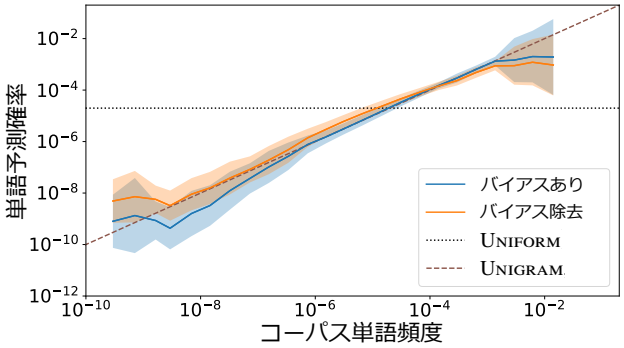

図 6　GPT-2 (medium) におけるバイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去による単語予測分布の変化.

A　各モデルサイズでの結果

3.1 節で行った，バイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去した際の予測確率分布の変化を各モデルで可視化した結果を図 5 から 8 に示す．3.1 節で行った，バイアス $\boldsymbol{b}_{\mathrm{LN}}$ を埋め込み行列 $\boldsymbol{W}_{\mathrm{emb}}$ への射影を BERT (base) で可視化した結果を図 9 に示す．スペースの都合でその他モデルでの可視化は省略する．

B　実験の詳細設定

言語モデルの単語予測分布を観察する実験（3.1 節および 3.2 節の PPL 計算）では，BERT には [MASK] 部分を単語予測させ，GPT-2 には各系列の 2 単語目以降を単語予測させた．入力系列の長さがモデルの最大入力長 k を超える場合は，先頭から k トークンのみを用いた．GPT-2 に文章を生成させる実験（3.2 節）では，先頭から 10 単語を文脈として与え，入力系列と同じ長さになるか終端トークンを生成するまで後続を生成させた．ビーム探索のビーム幅は 5，Top-k サンプリングの k は 50，Top-p サンプリングの p は 0.9 とした．また，モデルが生成した文章の多様性を評価する際には，NLTK の単語分割器を適用してから N-gram をカウントした．

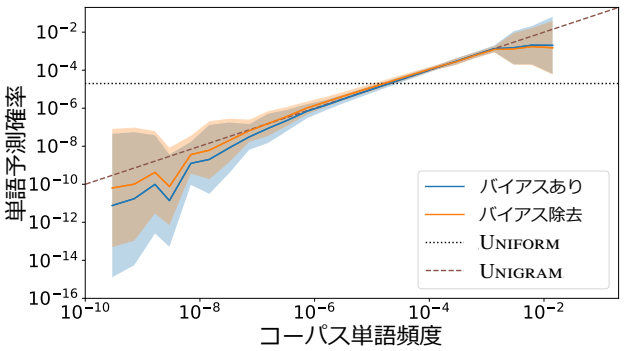

図 7　GPT-2 (large) におけるバイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去による単語予測分布の変化.

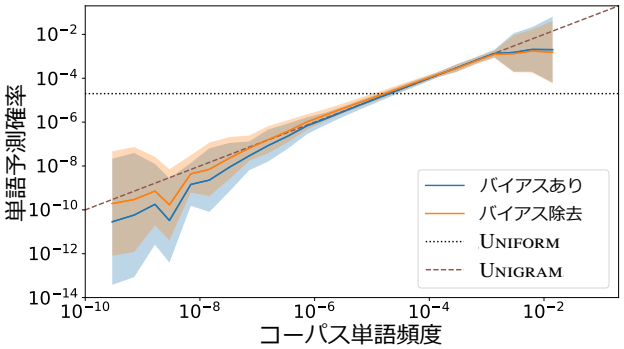

図 8　GPT-2 (xl) におけるバイアス $\boldsymbol{b}_{\mathrm{LN}}$ 除去による単語予測分布の変化.

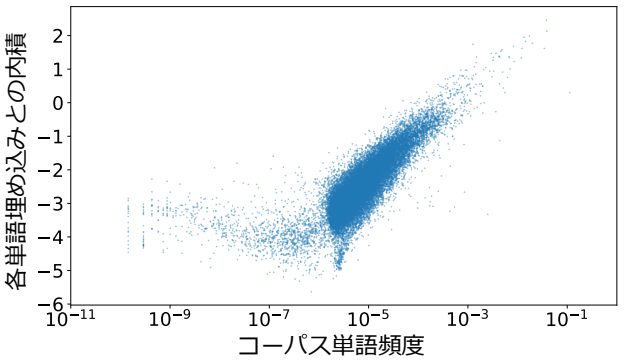

図 9　BERT (base) においてバイアス $\boldsymbol{b}_{\mathrm{LN}}$ を埋め込み行列 $\boldsymbol{W}_{\mathrm{emb}}$ に射影した結果.