意味とスタイルの分離に基づくスタイル表現学習

近藤 里咲 [1]
[1] 愛媛大学大学院理工学研究科
kondo@ai.cs.ehime-u.ac.jp

梶原 智之 [1,2]
[2] 大阪大学 D3 センター
kajiwara@cs.ehime-u.ac.jp

概要

本研究では，文埋め込みから意味情報とスタイル情報を分離することで，スタイル表現を獲得する．既存研究では，対話コーパス上での対照学習によってスタイル表現を得ているが，これにはスタイル以外の情報も含まれており，改善の余地がある．本研究では，文埋め込みを意味表現とスタイル表現に分離することによって，高品質なスタイル表現を獲得する．ご当地キャラクターと一般話者の発話を用いた実験の結果，本手法の有効性を確認できた．

1 はじめに

大規模言語モデルの登場によって，人間と自然に会話できる対話システムが実現しつつある．最近では，より自然な会話を実現するために，対話システムにスタイルを付与する研究が盛んである [1–5]．一貫したスタイルの応答生成を実現するためには，モデルがスタイルの特徴を捉える必要があるため，高品質なスタイル表現の獲得が望まれている．

スタイル表現を獲得する試みとして，深層学習が主流となった現在では，対照学習に基づく手法が採用されている [6,7]．これらの手法では，同一話者の発話を正例，別の話者の発話を負例として対照学習を実施することで，スタイル表現を得ている．しかし，この手法では発話文を符号化した文埋め込みをそのまま使用するため，得られるスタイル表現には意味などのスタイル以外の情報も含まれており，スタイル表現の品質に改善の余地がある．

多言語文符号化器から得た文埋め込みを言語固有の表現と言語非依存の表現に分離する手法 [8] から着想を得て，本研究では発話文の文埋め込みをスタイル情報と意味情報に分離することで，スタイルに特化した埋め込みを得る．本手法は，埋め込みモデル自体の更新を必要としないため，低コストにスタイル表現を得られる利点を持つ．

ご当地キャラクターと一般話者の発話文を用いた評価実験の結果，高品質なスタイル表現を獲得できた．また，スタイル表現を用いて話者識別器を構築し評価した結果，有用なスタイル表現が得られたことを確認できた．

2 関連研究

2.1 スタイル表現獲得

スタイルの類似度を捉えるために，スタイル表現獲得の手法が提案されている [6,7,9–11]．深層学習が主流の近年では，より高品質なスタイル表現を獲得するために，対照学習に基づく手法が提案されている．稲葉 [6] は，ある発話に対して関連度の高い同一キャラクターの発話部分集合を正例，別キャラクターの発話部分集合を負例とした対照学習を実施した．銭本ら [7] は，同一キャラクターの発話文を正例，別キャラクターの発話文を負例として，対照学習を実施した．

本研究でも，埋め込みベースの手法によってスタイル表現獲得に取り組む．既存の対照学習に基づく手法では，発話文の埋め込みをそのまま近づけたり遠ざけたりしており，文埋め込みに含まれる意味の情報を排除していない．そのため，スタイルの類似度が意味的な類似度に左右される可能性があり，獲得できるスタイル表現の品質に改善の余地がある．本研究では，意味とスタイルを分離することによって，高品質なスタイル表現の取得を試みる．

2.2 言語情報と非言語情報の分離

意味的類似度推定の性能改善を目的として，文埋め込みから言語表現と意味表現を分離する手法が提案されている [8,12,13]．本研究では，これらの研究における言語情報をスタイル表現に置き換え，文埋め込みからのスタイル表現と意味表現の分離に取り組む．ただし，一部の手法 [12,13] では，対訳関係にある言語しか区別できず，登場するすべての言語を区別することはできない．本研究はスタイルの区

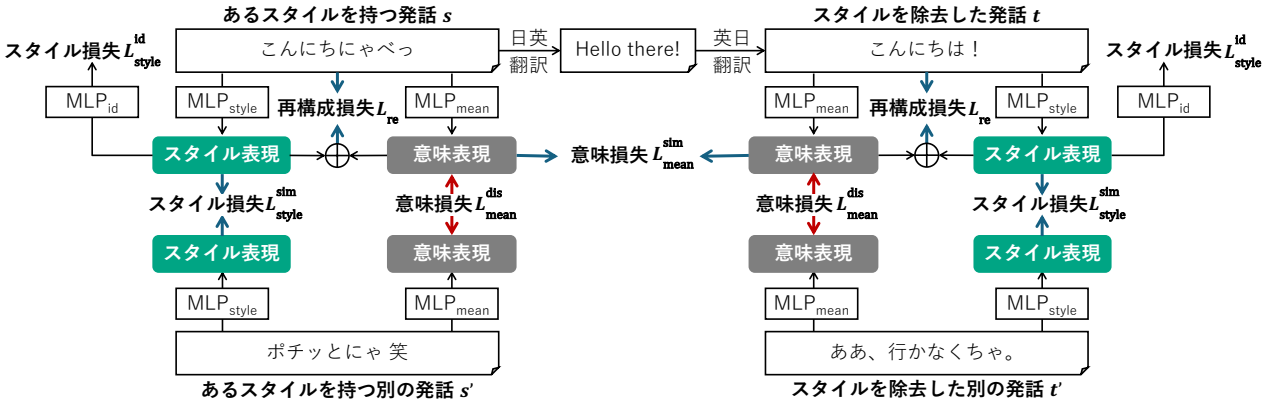

図 1　本研究の概要図

別が目的であるため，これらの手法は適さない．したがって，本研究では言語識別を行いつつ分離を学習する DREAM [8] をベースにスタイル表現を得る．

2.3　スタイルと折り返し翻訳

テキストからスタイルを除去するために，折り返し翻訳が用いられる．折り返し翻訳は，ある言語の文を中間言語に翻訳し，さらに元の言語に再翻訳する操作である．折り返し翻訳を用いることで，翻訳の過程でスタイルの情報が削ぎ落とされるため，スタイル変換に有用であることが報告されている [14, 15]．本研究でも，特定のスタイルを持つ発話に対して折り返し翻訳を適用することでスタイルを除去し，擬似的なパラレルコーパスを作成する．

3　スタイル表現の獲得

本研究では，スタイルの類似度を自動評価するために，発話文の文埋め込みをスタイル表現と意味表現に分離する．本研究の概要を図 1 に示す．まず，スタイルを持つ発話文に対して折り返し翻訳を適用し，スタイルを持つ発話とスタイルを除去した発話のペアを得る．その後，この文対を利用してスタイル抽出器と意味抽出器を共同で訓練することで，スタイルに特化した埋め込みを獲得する．

3.1　パラレルコーパスの作成

スタイル抽出器を訓練するためには，意味的に対応しつつスタイルの異なる文対が必要となる．しかし，このようなパラレルコーパスは非常に稀であるため，本研究では先行研究 [15] と同様に，スタイルを持つ発話に対して折り返し翻訳を適用することで擬似的なパラレルコーパスを構築する．

3.2　スタイル抽出器の訓練

本研究では，文埋め込みを意味表現と言語表現に分離する手法である DREAM [8] に基づき，言語表現をスタイル表現に置き換えることで，スタイル表現を獲得する．具体的には，文埋め込みから意味表現を抽出する多層パーセプトロン $\mathrm{MLP}_{\mathrm{mean}}$ とスタイル表現を抽出する多層パーセプトロン $\mathrm{MLP}_{\mathrm{style}}$ を訓練する．文埋め込みを $\mathbf{e}$ とすると，意味表現 $\hat{\mathbf{e}}_{\mathrm{mean}}$ とスタイル表現 $\hat{\mathbf{e}}_{\mathrm{style}}$ は下式のように表される．

$$\hat{\mathbf{e}}_{mean} = \mathrm{MLP}_{\mathrm{mean}}(\mathbf{e}) \tag{1}$$

$$\hat{\mathbf{e}}_{style} = \mathrm{MLP}_{\mathrm{style}}(\mathbf{e}) \tag{2}$$

これらの MLP を，DREAM の定義に従い，式 (3) に示す損失を用いて訓練する．本研究では，3.1 節で構築したパラレルコーパスを用いて，これらの損失を計算する．以降では，元の文埋め込みを s とし，折り返し翻訳後の文埋め込みを t とする．

$$L = L_{\mathrm{re}} + L^{\mathrm{sim}}_{\mathrm{mean}} + L^{\mathrm{dis}}_{\mathrm{mean}} + L^{\mathrm{sim}}_{\mathrm{style}} + L^{\mathrm{id}}_{\mathrm{style}} \tag{3}$$

再構成損失　再構成損失では，抽出したスタイル表現と意味表現の和が元の表現に近づくように、式 (4) で定義する．

$$L_{\mathrm{re}} = \frac{1}{d}||\mathbf{e} - (\hat{\mathbf{e}}_{\mathrm{style}} + \hat{\mathbf{e}}_{\mathrm{mean}})||^2_2 \tag{4}$$

ここで，d は埋め込みの次元数である．

意味損失　元の文埋め込みから得た意味表現と，折り返し翻訳後の文埋め込みから得た意味表現は等しくあるべきである．したがって，意味損失を式 (5) のように定義する．

$$L^{\mathrm{sim}}_{\mathrm{mean}} = 1 - \cos(\hat{\mathbf{s}}_{\mathrm{mean}}, \hat{\mathbf{t}}_{\mathrm{mean}}) \tag{5}$$

表 1　ご当地キャラクターの発話データ

	発話例	訓練	検証	評価
キャベッツさん	こんにちにゃべっ	2,000	234	233
ちぃたん☆	大漁ですっ☆ちぃたん☆ですっ☆	2,000	138	138
レルヒさん	オハ 雨ノ月曜日 ゆーうつヤー	60,000	163	163

表 2　一般話者の発話データ

	訓練	検証	評価
既知の話者	262,984	32,864	32,978
未知の話者	42,634	5,328	5,347

表 3　発話文のベクトル類似度に基づく Precision@k

	キャベッツさん			ちぃたん☆			レルヒさん			既知の話者 (200 人)			未知の話者 (33 人)		
	k=10	k=100	k=1000	k=10	k=100	k=1000	k=10	k=100	k=1000	k=10	k=20	k=40	k=10	k=20	k=40
BERT-base	0.982	0.929	0.761	0.986	0.964	0.894	0.918	0.863	0.612	0.016	0.014	0.013	0.080	0.075	0.068
意味	0.897	0.791	0.572	0.907	0.774	0.552	0.810	0.661	0.415	0.013	0.011	0.010	0.060	0.054	0.050
スタイル	**0.994**	**0.989**	**0.981**	**0.999**	**0.998**	**0.995**	**0.990**	**0.986**	**0.973**	**0.025**	**0.022**	**0.019**	**0.110**	**0.100**	**0.089**
BERT-large	0.978	0.946	0.786	**0.997**	0.976	0.832	0.955	0.899	0.614	0.019	0.016	0.014	0.087	0.079	0.072
意味	0.969	0.932	0.803	0.952	0.845	0.553	0.879	0.749	0.451	0.013	0.012	0.010	0.063	0.057	0.052
スタイル	**0.994**	**0.991**	**0.972**	0.996	**0.995**	**0.993**	**0.990**	**0.991**	**0.986**	**0.028**	**0.024**	**0.021**	**0.114**	**0.103**	**0.093**
ModernBERT 130m	0.954	0.880	0.619	0.977	0.961	0.864	0.953	0.906	0.771	0.014	0.012	0.011	0.069	0.062	0.055
意味	0.928	0.849	0.666	0.970	0.959	0.911	0.794	0.662	0.461	0.015	0.013	0.012	0.070	0.061	0.055
スタイル	**0.978**	**0.931**	**0.683**	**0.982**	**0.976**	**0.943**	**0.982**	**0.949**	**0.853**	**0.018**	**0.016**	**0.014**	**0.086**	**0.077**	**0.069**
ModernBERT 310m	0.892	0.749	0.488	0.973	0.938	0.648	0.910	0.852	0.662	0.013	0.012	0.010	0.060	0.055	0.051
意味	0.899	0.793	0.551	0.962	0.896	0.677	0.699	0.550	0.394	0.014	0.013	0.011	0.064	0.058	0.053
スタイル	**0.991**	**0.985**	**0.915**	**0.997**	**0.997**	**0.944**	**0.995**	**0.990**	**0.968**	**0.023**	**0.020**	**0.017**	**0.090**	**0.080**	**0.072**

また，異なる意味を持つ文から分離された意味表現同士は異なる表現を持つべきであるため，式 (5) に加えて式 (6) も考慮する．

$$L^{\text{dis}}_{\text{mean}} = \max(0,\ \cos(\hat{\mathbf{s}}_{\text{mean}}, \hat{\mathbf{s}}'_{\text{mean}})) + \max(0, \cos(\hat{\mathbf{t}}_{\text{mean}},\ \hat{\mathbf{t}}'_{\text{mean}})) \tag{6}$$

ここで，$\hat{s}'_{\text{mean}}$ は折り返し翻訳前の文 s のうち，t と対訳関係にないランダムな文の意味表現であり，$\hat{t}'_{\text{mean}}$ は折り返し翻訳後の文 t のうち，s と対訳関係にないランダムな文の意味表現である．

スタイル損失　同一スタイルを持つ文埋め込みから得たスタイル表現は等しくあるべきである．したがって，式 (7) によって，同一スタイルのスタイル表現を近づけるように訓練する．

$$L^{\text{sim}}_{\text{style}} = 2 - \cos(\hat{\mathbf{s}}_{\text{style}}, \hat{\mathbf{s}}'_{\text{style}}) - \cos(\hat{\mathbf{t}}_{\text{style}}, \hat{\mathbf{t}}'_{\text{style}}) \tag{7}$$

さらに，異なるスタイルは異なるスタイル表現を持つべきである．よって，式 (8) および式 (9) に示すように，スタイル識別用の MLP_{id} も同時に訓練することで，スタイルを区別する能力を獲得する．

$$\hat{\mathbf{y}} = \text{softmax}(\text{MLP}_{\text{id}}(\hat{\mathbf{e}}_{\text{style}})) \tag{8}$$

$$L^{\text{id}}_{\text{style}} = -\sum_{j}^{J} \mathbf{y}_j \log \hat{\mathbf{y}}_j \tag{9}$$

ここで，$\hat{\mathbf{y}}$ は各スタイルの予測確率であり，$\mathbf{y}$ は正解のスタイル，J はスタイルの数である．

4　内的評価

4.1　実験設定

本研究では，スタイル表現の品質を確認するために，内的評価を実施した．網羅的な評価のために，ご当地キャラクターと一般話者の発話文を用いた．

データセット　ご当地キャラクターには表 1 に示す 3 名の発話 [15] を用い，一般話者には 233 人の対話から構築された RealPersonaChat[1] [16] を用いた．後者では，既知および未知のスタイルに対する性能をそれぞれ評価するために 233 人を 200 人と 33 人に分割し，200 人を訓練に用いた．対話データから各話者の発話を取得し，重複する発話を削除した後，表 2 に示すように各話者の発話数が訓練用・検証用・評価用で 8:1:1 となるように分割した．

スタイル抽出器の訓練　折り返し翻訳の中間言語には英語を使用し，翻訳器にはオープンソースの PLaMo 翻訳モデル[2]を用いた．埋め込みを取得するモデルには，日本語で事前訓練された BERT [17] の base[3] および large[4]，日本語と英語で事前訓練さ

1) https://github.com/nu-dialogue/real-persona-chat
2) https://huggingface.co/pfnet/plamo-2-translate
3) https://huggingface.co/tohoku-nlp/bert-base-japanese-v3
4) https://huggingface.co/tohoku-nlp/bert-large-japanese-v2

れた ModernBERT [18] の 130m[5]，310m[6] を用いた．意味抽出器およびスタイル抽出器，言語識別器には 1 層の MLP を用いた．最適化に Adam [19] を使用し，バッチサイズは 512，学習率は 1e-5 とし，検証用データに対する損失が 3 エポック改善されなければ訓練を終了した．

評価方法　スタイル表現の品質は，類似度の高い top-k 件の訓練データのラベルから精度を算出する Precision@k で評価した．なお，ご当地キャラクターは 1 名あたりの最小の発話数が 2,000 件であるのに対し，既知の話者では 46 件，未知の話者では 60 件であったため，類似度の計算に用いる訓練データの各話者の埋め込みはそれぞれこの発話数に統一した．それに伴い，ご当地キャラクターでは k を $k \in \{10, 100, 1000\}$，一般話者では $k \in \{10, 20, 40\}$ と変動させたときの精度を報告する．また，比較手法として元の埋め込みおよび意味表現を用いて算出した Precision@k も報告する．

4.2　実験結果

表 3 にご当地キャラクターおよび一般話者の各種埋め込みを Precision@k によって評価した結果を示す．BERT-large におけるちぃたん☆の k=10 を除き，すべての設定においてスタイル表現を用いた場合が最高性能を達成した．これらの結果から，未知の話者にも有効で高品質なスタイル表現を獲得できたことを確認した．また，評価結果は全体的にスタイル表現 > 元の埋め込み > 意味表現となる傾向にあり，意図した分離が実現できたことを確認した．さらに，元の埋め込みや意味表現では，k が大きくなるにつれて精度が大幅に低下する傾向にあった一方で，特にご当地キャラクターのスタイル表現は，ほとんどの場合において 0.9 付近の精度を維持できた．

5　外的評価

5.1　実験設定

本研究で得たスタイル表現の有用性を評価するために，スタイル表現から話者を推定する話者識別器を構築し，分類性能を評価した．本実験でも，ご当地キャラクターおよび一般話者の発話を用いた．スタイル抽出器には 4 章で構築したものを使用した．

5）https://huggingface.co/sbintuitions/modernbert-ja-130m
6）https://huggingface.co/sbintuitions/modernbert-ja-310m

表 4　各種埋め込みで学習した分類器の正解率

	ご当地キャラクター	一般話者	
	キャベッツさん	既知	未知
BERT-base	0.953	0.126	0.274
意味	0.914	0.099	0.215
スタイル	**0.970**	**0.131**	**0.287**
BERT-large	**0.983**	0.141	0.303
意味	0.948	0.113	0.239
スタイル	0.979	**0.152**	**0.317**
ModernBERT 130m	0.966	0.126	0.256
意味	0.944	0.127	**0.301**
スタイル	**0.974**	**0.131**	0.298
ModernBERT 310m	0.974	0.135	0.287
意味	0.936	0.123	0.262
スタイル	**0.987**	**0.145**	**0.311**

話者識別器には 2 層の MLP を用い，中間層の活性化関数として ReLU，中間層の次元数は各モデルの半分の次元数とした．スタイル抽出器と同様の実験設定で分類器を訓練し，検証用データに対する損失が 3 エポック改善されなければ訓練を終了した．話者識別器の性能は，推定した話者ラベルの正解率で評価した．比較手法として，元の埋め込みおよび意味表現で訓練した分類器を用いた．

5.2　実験結果

ご当地キャラクターのうち，ちぃたん☆およびレルヒさんはどのモデルでもほぼ正解できたため，本稿ではキャベッツさんと一般話者の評価結果のみを報告する．ご当地キャラクターおよび一般話者の各種埋め込みから構築した話者識別器の評価結果を表 4 に示す．多くの設定において，スタイル表現で分類器を訓練した場合に最高性能を達成し，本研究で得たスタイル表現が話者識別に有用であることを確認できた．

6　おわりに

本研究では，高品質なスタイル表現獲得のために，発話文の埋め込みからの意味表現とスタイル表現の分離に取り組んだ．ご当地キャラクターと一般話者の発話を用いた評価実験の結果，高品質なスタイル表現を獲得できたことを確認した．また，本研究で得たスタイル表現が話者識別に有用であることを明らかにした．

謝辞

データを提供いただいた，西東京商工会青年部 広報・親善大使のキャベッツさん様，ちぃたん☆様，新潟県エヌキャラネットのレルヒさん様に，深く感謝いたします．本研究は，JST BOOST（課題番号：JPMJBY24036821）の支援を受けたものです．

参考文献

[1] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A Trainable Agent for Role-Playing. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13153–13187, 2023.

[2] Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. Characteristic AI Agents via Large Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, pp. 3016–3027, 2024.

[3] Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, pp. 11836–11850, 2024.

[4] Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, pp. 7828–7840, 2024.

[5] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14743–14777, 2024.

[6] 稲葉通将. 対話システムのための対照学習によるキャラクタ性の評価. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 1I5OS31b01–1I5OS31b01, 2024.

[7] 銭本友樹, 古俣槙山, 長谷川遼, 宇津呂武仁. 日本語の口調類似度評価データセットの作成および口調埋め込みモデルの構築・評価. 人工知能学会論文誌, Vol. 40, No. 5, pp. MO25–D_1–9, 2025.

[8] Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7764–7774, 2021.

[9] Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. Unsupervised Learning of Style-sensitive Word Vectors. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, pp. 572–578, 2018.

[10] Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. Fundamental Exploration of Evaluation Metrics for Persona Characteristics of Text Utterances. In **Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 178–189, 2021.

[11] 石川和樹, 小川浩平, 佐藤理史. 口調エンコーダを用いた小説発話の話者推定. 自然言語処理, Vol. 31, No. 3, pp. 894–934, 2024.

[12] Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 5240–5245, 2022.

[13] Keita Fukushima, Tomoyuki Kajiwara, and Takashi Ninomiya. Reversible Disentanglement of Meaning and Language Representations from Multilingual Sentence Encoders. In **Proceedings of the 5th Workshop on Multilingual Representation Learning**, pp. 265–270, 2025.

[14] Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 8042–8049, 2020.

[15] 近藤里咲, 梶川怜恩, 梶原智之, 二宮崇. スタイル変換による雑談対話へのキャラクタ性の付与. 情報処理学会論文誌, Vol. 65, No. 3, pp. 657–666, 2024.

[16] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. RealPersonaChat: A Realistic Persona Chat Corpus with Interlocutors' Own Personalities. In **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 852–861, 2023.

[17] 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**, pp. 4171–4186, 2019.

[18] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. **arXiv:2412.13663**, 2024.

[19] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In **Proceedings of the 3rd International Conference for Learning Representations**, 2015.