E-044

大規模言語モデルを用いたテキスト入力に向けた個人メール文体への適応
Fine-tuning of Large Language Model to Individual E-mail Style toward Text Entry

猪飼 宗樹† 加藤 恒夫† 田村 晃裕†
Soki Ikai Tsuneo Kato Akihiro Tamura

1. はじめに

スマートフォンなどに搭載されているテキスト入力システムはフリック入力や予測変換によって使いやすいものであるが，基本的にユーザが全て入力する．これに対して，テキスト入力支援にコンテキストを考慮して流暢なテキストを生成できる大規模言語モデル(Large Language Model, LLM)を用いれば，半自動でメール文の入力などを行えるようになることが期待できる．しかし，LLM が生成する文は通常説明的なフォーマルな文章であり，個人の個性が表れたカジュアルな表現ではない．

LLM をファインチューニング(FT)することで特定のタスクやドメインに特化する取り組みは多数行われているが，テキスト入力支援に応用する試みは少ない．タブレットにおけるテキスト入力に，入力文を構成する各単語の頭文字を手書きで入力すると完全な文の候補を提示するインタフェース[1]が提案されている．

本研究では個人用の LLM をメール入力支援に用いることを想定し，特定の個人による数千のメール文を用いて LLM を個人のメール文体に適応し，その次単語予測精度を評価する．LLM の適応手法としては，パラメータ効率の良い FT 手法である Low-Rank Adaptation (LoRA)[2]を用いる．次単語予測の精度は，単語正解率，PPL，MRR の 3 種類の評価指標で測る．また，使用するメール文コーパスは主に送信メールを収録し，受信メールをほとんど含まないため，gpt-4o を用いて送信メールに挟まれた受信メールを推定して訓練，評価用のデータセットを作成する．

2. LoRA を用いた大規模言語モデルの個人メール文体への適応と評価

LLM は Transformer デコーダで構成され，自己回帰的な単語出力の過程で，与えられた文脈や過去の出力単語との関連性を自己注意機構により捉える． ここでは，メール文の入力を想定し，一つ前の送信メールもしくは直前の受信メールをコンテキストとして，続く送信メール文を予測する．

LoRA は，学習可能パラメータの変化分を少数のパラメータで表現する効率的な適応手法である．今回は Transformer の自己注意機構にあるクエリ，キー，バリューとマルチヘッド統合後の線形変換行列に LoRA を導入する．また，使用するメール文コーパスは主に送信メールで構成され，受信メールをほとんど含まないため，そのままでは直前の受信メールをコンテキストとする送信メール文の予測を評価できない．そこで，gpt-4o を用いて受信メールを推定して学習，評価を行う．

†同志社大学大学院 理工学研究科
Graduate School of science and engineering,
Doshisha University

表 1 個人メール文体学習用データセットの例
（同一人物による連続送信メール対）

例 1	一つ前送信メール	こっちは雨降り始めたよ～
	送信メール	なんだか降ったり止んだりしてる
例 2	一つ前送信メール	確かに
	送信メール	上手い人の後に歌いとうない
例 3	一つ前送信メール	わー！
	送信メール	まだ取ってなかったか

表 2 実験に使用したデータセットの分量

Sender (Context)	Train	Test
女性 A（一つ前の送信メール）	3623	402
女性 A（gpt-4o の推定受信メール）	3623	402
女性 B（一つ前の送信メール）	2871	317
女性 B（gpt-4o の推定受信メール）	2871	317

3. 実験

3.1 データセット

個人のカジュアルな文体のメール文を収録したコーパスとして「加藤安彦ケータイメイルコーパス」[3]を用いる．同コーパスには，大学生が公開を許諾したメールが匿名化された宛先とともに時系列で収録されている．そこで，LLM の訓練・評価用データセットとして，特定の女性 2 名による，前後で宛先が揃った送信メールを対にして抽出する．表 1 に前後の送信メール対の例を示す．友人とのメール交換の様子が掴めるが，受信メールが欠けているため正確な話題の展開は掴めない．収録された送信メールの数が多かった 2 名の女性 A と B についてデータセットを作成した．表 2 に各データセットの構成を示す．訓練セットと評価セットがおおよそ 9:1 となるようにコーパスから抽出した対をランダムに分けた．なお，コーパス中の「＜人物＞」のように匿名化された固有名詞は括弧タグをそのまま使用し，絵文字や記号は削除して用いた．

評価セットの基本的な統計量を測った．送信メール文の平均文字数が A: 10.9，B: 13.5，平均形態素数(MeCab)が A: 6.68，B: 8.05，distinct-1(Mecab)が A: 0.31，B: 0.30，distinct-2(Mecab)が A: 0.76，B: 0.75 であった．

3.2 gpt-4o による受信メールの推定

メール文生成のコンテキストとして一つ前の送信メールを用いるよりも直前の受信メールを用いる方が，予測精度

が高くなる可能性が高いと考え，コーパスには収録されていない直前の受信メールを，gpt-4o に前後の送信メールを与えて推定する．gpt-4o に対して，システムメッセージにはタスクと出力形式，入出力例を指定し，続いて受信メールを挟む前後の送信メールを指定する．

その結果，表 1 の 3 例については，それぞれ「天気が変わりやすいね」（例 1），「でも、上手い人の後ってプレッシャーあるよね？」（例 2），「見落としてたところあった？」（例 3）という受信メールが推定された．

品質を確認するため，女性 A の評価セット 402 文について gpt-4o が推定した受信メールを評価者 3 名で評価した．前後の送信メールに挟まれた受信メールの自然さを「1:自然である，2:許容範囲，3:許容できない」の 3 段階で評価した．3 名による評点の内訳は，1 が 31.3%，2 が 40.7%，3 が 28.0%であった．品質に問題のある受信メールも少なからず含まれたが，gpt-4o が推定した直前の受信メールと直後の送信メールを対にしたデータセットを作成し，LLM の FT と評価に用いる．

3.3 実験条件

LLM として Llama-3-ELYZA-JP-8B[4]を用いる．同モデルの Transformer 全 32 ブロックの自己注意に含まれるクエリ，キー，バリューとマルチヘッド統合後の線形変換行列に対して LoRA を適用する．LoRA の低ランク次元数 r は 32，スケーリング係数は 64 とした．ドロップアウト率を 0.1，バイアスを none とした．バッチサイズは 4，最適化アルゴリズムは AdamW，初期学習率を 2e-5 とし，ウォームアップ付きのコサインスケジュールを採用した．エポック数を 10 とした．

3.4 評価指標

評価は，Teacher Forcing を用いて常に直前までの正解を与え，次のトークンの予測に限定する．単語正解率(WA)，テストセットパープレキシティ(PPL)，平均逆数ランク(MRR)の 3 種類の指標を測る．

単語正解率WAは正しく予測したトークンの割合である．

$$WA = \frac{N_H}{N_H + N_W} \tag{1}$$

ここで，N_Hは正しく予測されたトークンの数，N_Wは誤って予測されたトークンの数である．

テストセットパープレキシティ$\boldsymbol{PPL}$は，言語モデルの次単語予測精度を測る指標である．

$$PPL = e^{\bar{L}} \tag{2}$$

$$\bar{L} = \frac{\sum_{i=1}^{N} L_i}{N} \tag{3}$$

ここで，$\boldsymbol{N}$はトークン総数，$\boldsymbol{L_i}$ は各トークンのクロスエントロピ損失である．

$\boldsymbol{MRR}$は正解トークンが出現した順位を測る指標である．

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R_i} \tag{4}$$

表 3　女性 2 名の評価セットに対する
ファインチューニング前後の次単語予測精度

モデル（文生成コンテキスト）	WA	PPL	MRR
女性 A			
FT 前（一つ前の送信メール）	0.323	18.5	0.430
FT 後（一つ前の送信メール）	**0.456**	10.8	**0.568**
FT 前（gpt-4o 推定受信メール）	0.378	14.6	0.481
FT 後（gpt-4o 推定受信メール）	0.429	**10.0**	0.538
女性 B			
FT 前（一つ前の送信メール）	0.337	20.0	0.442
FT 後（一つ前の送信メール）	**0.473**	11.8	**0.577**
FT 前（gpt-4o 推定受信メール）	0.388	16.5	0.490
FT 後（gpt-4o 推定受信メール）	0.451	**10.5**	0.553

ここで，$\boldsymbol{N}$はトークン総数，R_iは i 番目のトークンにおける正解の順位である．

3.5 実験結果

表 3 に，2 名の女性の評価セットに対する FT 前後の単語正解率，PPL，MRR を示す．メール文生成のコンテキストとして，一つ前の送信メールと gpt-4o が推定した直前の受信メールの 2 種類を評価している．FT により全条件で単語正解率，PPL，MRR の全ての評価指標が改善され，単語正解率は 0.5 弱，PPL は約 10，MRR は 0.5 を超えた．

次に，gpt-4o が推定した直前の受信メールをコンテキストにすることで，FT 前でも WA の上昇，PPL の低下，MRR の上昇が確認され，次単語予測タスクが易しくなった．しかしながら，gpt-4o が推定した直前の受信メールをコンテキストとする FT による次単語予測精度の改善は限定的で，PPL が最小となった以外は単語正解率も MRR も最高には至らなかった．

4. おわりに

LLM をテキスト入力支援に用いることを想定し，特定の個人による数千通のメール文により LLM を個人のメール文体に適応し，次単語予測精度を評価した．単語正解率が 0.5 弱，テストセットパープレキシティが約 10，MRR が 0.5 を超え，有望な精度を確認した．gpt-4o を用いた受信メールの推定は次単語予測タスクを容易にする一方，推定した受信メールをコンテキストとする LLM の適応はそれほど効果がなかった．

参考文献

[1] Z. Xu et al., “SkipWriter: LLM-Powered Abbreviated Writing on Tablets”, UIST '24, 2024.

[2] E.J.Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”, ICLR 2022.

[3] 田中他, “加藤安彦ケータイメイルコーパス”, 計量国語学, 34(2), pp. 111-121, 2023.

[4] https://note.com/elyza/n/n360b6084fdbd ,2024 年 6 月 26 日