1V-05

SNS のテキストデータを用いた BERT による投稿者の属性推定

堂前拓生† 上田芳弘† 坂本一磨† 池田理玖‡
公立小松大学生産システム科学部生産システム科学部
公立小松大学大学院 サステイナブルシステム科学研究科 生産システム科学専攻‡

1. はじめに

近年，多くの人が SNS(Social Networking Service)の投稿を通じて，多種多様な情報が発信されるようになった. SNS を活用し職業や年代など属性を推定することができれば，マーケティングに役立てることが可能である．本稿では, SNS のテキストデータから投稿者の属性の一つである職業の推定に着目する． 本稿においては，既存研究[1]と同様に文章分類において優れたモデルである BERT （Bidirectional Encoder Representations from Transformers）[2]を用いて，属性推定を行う． BERT の特徴としては，文章の分類問題，穴埋め問題等の複数タスクへの対応が，少ないデータによるファインチューニングで可能なことや，文章を双方向の Transformer によって学習することにより，文脈理解に優れた処理が可能であることが挙げられる．本稿では BERT を用いて，マイクロブログの投稿から投稿者の職業を学習・推定し、良好な精度で推定可能かを検証する．なお, 東北大学の研究室が公開している Wikipedia で事前学習された日本語 BERT モデル[3]を用いた.

2. 研究の概要

本稿でのシステムの概要を図 1 に示す．本システムでは，ファインチューニング無しの場合は，Wikipedia によって事前学習済み BERT モデルを用いて，評価用テキストデータによって評価し，正解ラベルと一致するかどうかの属性推定行う．ファインチューニング有りの場合のシステムの流れを図 1 に示す．図 1 に示すとおり，Wikipedia によって事前学習された BERT のモデルに訓練用の SNS テキストデータとテキストラベルを学習させ，訓練されたモデルに評価用の SNS テキストデータによって評価し，正解ラベルと一致するかどうかの属性推定を行うことで分類精度を比較するというシステムである．文章関係学習機能では，Web 上の文章群として，日本語の Wikipedia を使用し，BERT に事前学習をさせることで， BERT 事前学習モデルを構築する．文章特徴学習機能では，BERT 事前学習モデルを参照し，整形された文章を入力として，BERT でファインチューニングをし，BERT 訓練済みモデルを構築する．文章分類機能では，評価用投稿を入力として，属性推定をする.

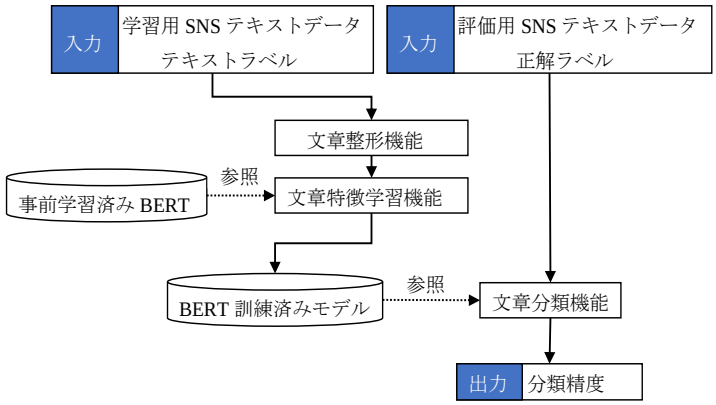

図 1 研究の流れ

表 1 ツイートデータ

		カテゴリ	
		投稿者数(人)	投稿数（ツイート）
職業	美容師	37	101,330
	カメラマン	40	195,230
	エンジニア	20	48,011
	学生	20	47,247

3. 検証実験

Wikipedia で事前学習された日本語 BERT モデルを用いて美容師，カメラマン，エンジニア，学生の 4 つの職業の Twitter の投稿について分析する．投稿を学習データ，検証データ，テストデータに分割（6:2:2）して，図 1 の流れに沿って全体の分類精度とそれぞれの職業についての分類精度を比較し分析した．ファインチューニング無しの場合は，学習データ，検証データ，を使用せず，テストデータを用いて，Wikipedia で事前学習された日本語 BERT モデルで評価す

Attribute Estimation of Contributors by BERT Using Microblogs
†Takumi Doumae, Yoshihiro Ueda , Kazuma Sakamoto and Riku Ikeda
Faculty of Production Systems Engineering and Sciences. Komatsu University

る．なお，BERT の処理には不適合な URL や記号等の情報は，予め投稿から削除して使用した．使用したデータの詳細は表 1 のようになっている．BERT にて実験を行った結果は，表 2 に示したようにファインチューニングを行わなかった場合は全体の平均精度は 28.0%であり，最も高精度なカテゴリはエンジニアで 37. 0%，逆に最も低精度なカテゴリは学生で 5. 0%であった． 学習回数 100 回でファインチューニングを行った場合は全体の平均精度は 61.4%であり，最も高精度なカテゴリはカメラマンで 97.5%，逆に最も低精度なカテゴリは，学生で 17.5%となった．

4. 考察

図 2 に表 2 の結果を図に示した．表 2，図 2 に示すように全てのカテゴリにおいて投稿によるファインチューニングによって分類精度は向上することがわかった．これは，Wikipedia で事前学習された BERT モデルをファインチューニングによって SNS に対応できる可能性を示唆しているといえる．一方，学生の分類精度はファインチューニング前後で極端に低いものであった．これは，例えばカメラマンやエンジニアでは，表 3 の正解ラベル B，C のように特徴的な単語あるいは専門的な単語を含んだ投稿が多かったため高精度な結果得られたが，美容師や学生では，表 3 の正解ラベル A，D のように特徴的な単語を含む投稿が少数であり，趣味などの投稿が多かったため，低精度な結果になったと考えられる．

5. 終わりに

本稿では BERT を用いて SNS の投稿を 4 種類の職業カテゴリに分類してその精度を分析した．その結果，Wikipedia で事前学習した BERT モデルを，SNS の投稿でファインチューニングすることより，分類精度が向上することがわかった．また，カテゴリによっては，十分な分類精度を得られるが，逆に極端に低い精度のカテゴリも存在する．今後は，この原因について更に分析する予定であり，SNS の投稿によって投稿者の職業を推定できるカテゴリと，推定が困難なカテゴリを明らかにすることを目指す．

表 2 実験結果

		条件	
		ファインチューニング無し (%)	学習回数 100 回 (%)
カテゴリ	全体	28.0	61.4
	美容師	35.5	44.0
	カメラマン	23.3	97.5
	エンジニア	37.0	74.9
	学生	5.0	17.5

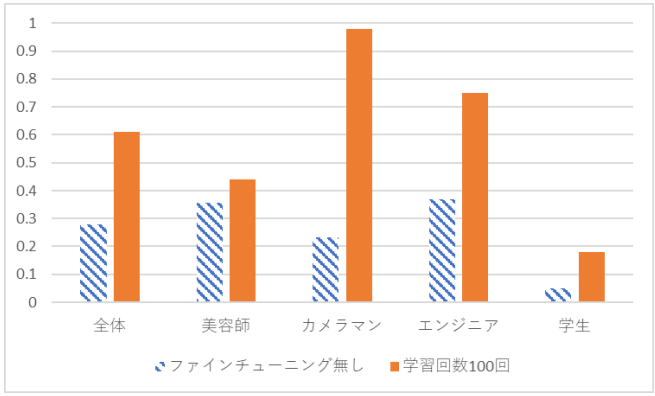

図 2 実験結果比較

表 3 エンジニアのツイート例

		カテゴリ	
		ツイート例	推定結果
正解ラベル	A	寒い よ 満喫 に 行き たい	B
	B	発売 の レンズ を お 借り しまし た	B
	C	Y mobile（ワイ モバイル で iPhone 7 を 分割 で 契約 する 方法 と は Apple ローン を 利用 すれ ば 金利 0 円 の 料金 も お 得	C
	D	これ まで 茨城 は 魅力 度 ランキング ワースト でし た が ついに 脱 最下位 これから も 推し ます 茨城 …	A

A:美容師 B:カメラマン C:エンジニア D:学生

参考文献

［1］ 中川嵩将,上田芳弘,坂本一磨:BERT を用いたマイクロブログユーザの興味推定に関する研究, 情報処理学会第 84 回全国大会講演論文集,Vol.2022,No.12022 ,pp729-730(2022)

［2］ Devlin, J. Chang, M, W. Lee, K. Toutanova, K. : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (2018).

［3］ 鈴木正敏:Pretrained Japanese BERT models released /日本語 BERT モデル公開,東北大学乾研究室知能情報科学講座自然言語処理学分野, https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/