様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２５ 日現在

機関番号：２２７０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１９６７

研究課題名（和文）音声に内在する個人性の言語的側面に関する研究

研究課題名（英文）Improving Speaker Recognition by Using Linguistic Information Inherent in Speech

研究代表者
越仲　孝文（Koshinaka, Takafumi）

横浜市立大学・データサイエンス学部・教授

研究者番号：６０８９５９２８

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：顔認証や静脈認証などと比べて実用化が進んでいない音声認識(話者認識)の技術水準向上を目指し，音声に含まれる言語情報の活用方法について調べ，近年の深層学習モデルの有効性を示した．関連して，近年目覚ましく発展する大規模言語モデルを含む生成AIと人間の識別可能性を検証した．さらに生成AIの能力や振舞いを明らかにするため，大規模言語モデルが出力するテキストデータの品質を比較し，GPT-4のような最高水準のモデルが人間と同等以上のテキスト生成能力を持つことを示した．画像キャプショニングモデルについても検証を行い，最新のモデルが画像分類タスクでその能力を定量的に測定した．

研究成果の学術的意義や社会的意義
ディジタル社会の進展に伴い，ユーザの本人確認を安全かつ簡便に行う技術が求められている中で，顔認証や静脈認証などと並んで普及が期待される音声認証の精度を改善し，より安全で便利な社会の実現に貢献する．加えて，近年急速に発展して社会的な注目度も高い，大規模言語モデルなどのいわゆる生成AIの性質を明らかにすることにより，AIの社会への普及を促進し，AI技術の健全な発展に貢献する．

研究成果の概要（英文）：Aiming to improve speaker recognition technology, which has not yet been put into practical use compared to facial recognition or vein recognition, we investigated methods for utilizing the linguistic information contained in speech and demonstrated the effectiveness of recent deep learning models. We also examined the discriminability of generative AI, including large-scale language models, which have made remarkable progress in recent years, and humans. Furthermore, to clarify the capabilities and behavior of generative AI, we compared the quality of text data produced by large-scale language models and demonstrated that state-of-the-art models, such as GPT-4, have text generation capabilities equal to or greater than humans. We also examined image captioning models, quantitatively measuring the capabilities of the latest models in image classification tasks.

研究分野：知能情報学，知覚情報処理

キーワード：筆者認識　話者認識　生体認証　深層学習　大規模言語モデル　生成AI　ディープフェイク検出

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

ディジタル社会の進展に伴い，企業や官公庁が提供する各種サービスにおいてユーザの本人確認を確実に行うことがますます重要になっている．本人と他人を正しく区別する安全性もさることながら，ユーザにとっての簡便性もこの種のサービスでは軽視できないため，顔認証や静脈認証に代表される生体認証技術が注目され，実用化も進んでいる．音声の時間-周波数パターン(いわゆる声紋)によって個人を識別する音声認証(声紋認証)もその一つであり，声を発するだけで認証されるという高い簡便性から今後の普及が期待される．しかしながら現状では，顔認証や静脈認証などと比較して音声認証の精度は低く，雑音がない良好な条件であっても 99～99.9%程度である．4 桁の暗証番号(原理上の精度は 99.99%)に劣るようでは，認証方式としての実用化の見込みはないと言っても過言ではない．

他方では，深層ニューラルネットワーク(DNN)の Transformer アーキテクチャが 2017 年に登場して以来，大規模なニューラル言語モデルの研究開発が急速に進展，これらのモデルが人間と区別がつかないような受け答えを見せるようになってきた．本研究課題の申請(2020 年 11 月)は，OpenAI の GPT-3 発表から半年ほどが経過した時期で，米国の人気掲示板 Reddit で GPT-3 ボットが人間に紛れて，1 分おきというハイペースで大量の投稿を 1 週間続けたという衝撃的な事件[1]が報じられた頃である．それから現在に至るまで，生成 AI あるいは大規模言語モデル(LLM)と総称される DNN モデルが続々と公開され，悪用の危険性が世界各国で共有されている現状は周知の通り．現在の，特に ChatGPT 以降の状況が当時正確に想像できていたとは言い難いが，生成 AI によって作られる本物のような偽物，いわゆるディープフェイク(deepfake)の悪用を防ぐ方策が必要となることは，研究開始当初から予見されていた．

２．研究の目的

前述の背景を踏まえ，第一に音声認証の実用化推進に資する研究を行う．具体的には，従来の話者認識(speaker recognition; 音声認証に対応する学術分野の名称)が話者の発声器官の形状に個人差があることを前提として，音声の音響的な特徴を手掛かりとして推論を行うアプローチ[2]を基本とするのに対して，本研究は音声の言語的な特徴を手掛かりとする．人にはそれぞれよく使う語彙や言い回し，口癖のようなものがあり，それらを個人差としてとらえることができれば，発声器官の個人差と同様に，話者を識別する際の手がかりになるはずである．テキストデータから筆者の特徴をとらえて誰が書いたかを推論する技術は，自然言語処理分野で筆者認識(authorship recognition)と呼ばれ，それほど多くはないがいくつかの研究が報告されている．ただ近年発展著しい深層学習の手法が活用されているとは言い難い[3]．本研究は，筆者認識を話者認識の精度向上の一手段と位置付けて，最新の深層学習手法を用いてアップデートすることを目指す．

第二に，筆者認識の手法を人間の識別ではなく人間と機械の識別にも適用し，有効性を検証する．いわゆる生成 AI によって生成されたテキストがどのような点で人間が書いたテキストと異なるのか(あるいは見分けがつかないのか)を明らかにする．さらに視点を広げて，GPT シリーズのような純粋な言語モデルのみならず，画像キャプショニング(画像説明文生成)モデルのように画像からテキストを生成するようなモデルも含めて，生成 AI の一般的な性質を明らかにする．本研究を通して，生成 AI の実社会での活用指針を提供したい．

３．研究の方法

まず，テキストと筆者の関係について主に以下の 2 点から考察する．

(1) テキストを手掛かりとした筆者認識モデルの構築と精度評価
テキストを入力として筆者を推論する深層ニューラルネットワーク(DNN)モデルを構築する． 精度評価においては単語の重み付き出現頻度(TF-IDF)に基づく非 DNN なモデルと比較し，DNN モデルの有効性を検証する．

(2) 生成 AI が出力したテキストと人間が書いたテキストの識別実験
同様の DNN モデルを用いて， 与えられたテキストの筆者が人間か否かを推論する DNN モデルを構築し，人間と機械(ニューラル言語モデル)の際について考察する．

また，生成 AI の性質について以下の 2 点から考察する．

(3) 生成 AI が出力したテキストと人間が書いたテキストの品質比較
生成 AI として大規模言語モデル(LLM)を活用し，生成 AI が作成したテキストと人間が作成したテキストの品質を測るモデルを作成し，最新の生成 AI の能力を評価する．

(4) 生成 AI が出力するテキストが有する情報の定性/定量的評価
ここでは画像キャプショニング(画像説明文生成)モデルに着目し， 説明文を画像分類タスクに用いることで，説明文が持つ情報の定量的測定を試みる．

当初は(1)に主眼を置いていたが， GPT-3 やそれに続く大規模モデルの隆盛を鑑みて(2)に， さらに(3)～(4)に注力する方向にシフトした．

４．研究成果

前節に示した(1)～(4)のそれぞれについて以下に研究成果を示す．

(1) テキストを手掛かりとした筆者認識モデルの構築と精度評価

筆者認識をテキストからその筆者を予測する文書分類問題とみなし，いくつかの分類モデルを構築した:

- TFIDF+RL: テキスト中の内容語(名詞，動詞など)の重み付き出現頻度(TF-IDF)で表現し他クラスロジスティック回帰により分類するモデル
- TFIDF+MLP: ①でロジスティック回帰を多層パーセプトロン(MLP)に置き換えたモデル
- bLSTM: テキスト中の全単語を双方向 LSTM に入力して分類するモデル
- Transformer: ③の双方向 LSTM を Transformer に置き換えたモデル

日本語の小説のコレクションとして知られる公開データセット「青空文庫」から，作品数の多い著名筆者 10 人を選び，段落単位での分類実験を実施した．段落総数は約 32,946 で 80%を訓練に，残りの 20%を評価にそれぞれ用いた．分類精度は下表 1 の通り．

表 1: 各モデルの筆者認識精度

	層数	パラメタ数	分類精度(%)
TFIDF+LR	-	6.8K	52.1
TFIDF+MLP	-	679K	50.8
bLSTM	2	11M	**64.8**
	4	24M	59.0
	6	36M	50.0
Transformer	2	13M	63.7
	4	19M	**64.7**
	6	25M	64.1

深層ニューラルネットワークに基づく end-to-end 分類器である bLSTM や Transformer の精度が最大で約 65%ともっとも高く，TF-IDF に基づく旧来型モデルの 52%を大きく上回った．ただし，モデルの規模と分類精度が必ずしも比例していないことから，訓練データ量が不十分である可能性があり，また学習率やエポック数などのハイパーパラメタは調整の余地があると思われる．分類誤りの傾向を見たところでは，特定の品詞，例えばフィラーの多寡が筆者の特徴として重視されていると見られ，ある種のフィラーを多用する筆者の間で混同が起こりやすいことなどが明らかになった．

(2) 生成 AI が出力したテキストと人間が書いたテキストの識別実験[4]

日本語 GPT-2 をベースとして，ドメインを限定した小規模な生成 AI を作成した．具体的には，楽天トラベルの宿泊施設レビューと施設側の応答を訓練データとして，レビューを与えると施設側の応答を出力するモデルを構築した．このモデルが生成する応答文は一見自然で，宿泊施設の担当者が書いたものとして読んでも違和感のないものが多い．

モデルが生成したフェイク応答文と，宿泊施設が書いた本物の応答文を各 3000 件用意し，分類実験を行った．分類モデルには BERT を用いたところ，98%の分類精度を確認した(表 2)．つまり本物とフェイクをほぼ完ぺきに見分けることができた．初期の ChatGPT に対して類似の実験を行った報告[5]でも同様の結果が示されている．しかし，条件を少し変えると違った結果が得られることも判明した．すなわち，文生成の条件をビーム探索から Top-K 探索に変更すると，分類精度は 70%にまで低下した．つまり，文全体としての自然性や一貫性を少し緩めることで，生成 AI と人間の区別がつきにくくなるということである．具体的な現象としては，ビーム探索時には比較的よく出現していた「この度は当ホテルをご利用いただきありがとうございます」のような定型的な言い回しが，Top-K 探索では使われなくなり，より多様な応答文が生成される傾向となった．

表 2: 生成 AI と人間の識別精度

	モデル	ビーム探索	Top-K 探索
本研究	日本語 GPT-2	0.98	0.70
Mitrović **et al.[5]**	ChatGPT	0.98	-

生成された文の言い換え(rephrasing)で生成 AI と人間の区別が困難になることはよく知られているが，文生成時の経路探索アルゴリズムのような些少な設定も生成 AI の特徴に影響を与えていることがこの実験を通して明らかになった．

(3) 生成 AI が出力したテキストと人間が書いたテキストの品質比較[6]

最新の大規模言語モデル(LLM)が生成するテキストの性質については，ChatGPT 登場以来，様々な評価報告があるが，本研究では検索エンジン最適化(SEO)という応用場面に着目して評価を行った．グーグル検索に代表される検索エンジンで上位にランクされる Web ページにはいくつかの良い特徴があり，それを人間の被験者が評価して 10 段階等で評価するユーザ評価と呼ばれる取組みが SEO 業界では一般的に行われている．本研究では，①ユーザ評価によって 10 段階のスコアを付与された Web ページのテキストデータを用いて，テキストからユーザ評価に準ずるスコアを予測するモデルを作成し，②そのモデルを用いて LLM が生成するテキストデータの品質を定量的に評価した．LLM としては OpenAI の GPT-4，GPT-3.5 (いずれもパラメタ数非公開)，およびサイバーエージェント社が公開する CALM2 (70 億パラメタ)を用いた．

①では 100 クエリについて各々グーグル検索で上位 50 件の Web ページを収集し，計 5000 件の Web ページについて 10 段階のユーザ評価を実施，これを用いてユーザ評価予測モデルを作成した．②では同じクエリを LLM に提示して Web ページに相当するテキストを出力させ，それらをユーザ評価予測モデルに与えてスコアを算出した．結果は図 1 のボックスチャートのようになった．左の 6 列(評価 4～評価 9)は人間が書いたテキストで，各々がユーザ評価 4～9 に相当する．右の 3 列が LLM によって生成されたテキストである．世間で言われているように，GPT-4 の能力の高さが示される結果となった．GPT-4 が生成するテキストの品質は，ユーザ評価で 7～8 に相当し，平均的な人間が各テキストと同等かそれ以上の水準にある．

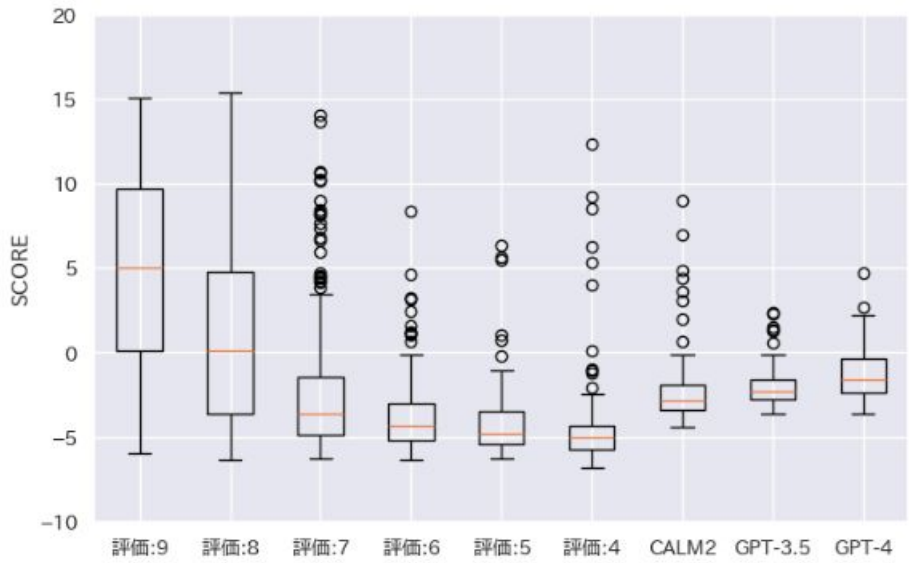

図 1: LLM が生成したテキストと人間が書いたテキストの品質比較

今回はユーザ評価モデルを品質評価の手段として用いたのみであるが，これを報酬モデルとみなして LLM のアラインメント(RLHF など)を行えば，SEO 観点で最適な LLM，つまり検索エンジンで上位にランクされるコンテンツを生成する能力を備えた LLM を作ることもできるだろう．そのような観点で現在研究を進めている．

(4) 生成 AI が出力するテキストが有する情報の定性/定量的評価[7][8]

生成 AI の能力を測るという視点でもう一つ研究を行った．Vision and Language の大規模モデルを用いて画像の説明文(キャプション)を生成させたとき，モデルは画像中のどんな情報をとらえてキャプションに反映させるのだろうか? このようなリサーチクエスチョンを設定し，その第一歩として図 2 のような画像分類システムを試験的に構築した．画像分類では昨今，大量の画像データで事前学習された大規模モデル(ResNet，ViT など)を用いれば高い分類精度が得られることが知られている．図の左半分のフローがそれである．しかし，画像キャプショニングを用いて図の右半分のような構成で画像分類を行うこともできる．このような構成は概念ボトルネックモデル，より正確には言語ボトルネックモデルとも呼ばれる．近年発表されている BLIP や CLIP Interrogator などの画像キャプショニングモデルは非常に饒舌で，与えられた画像を事細かに説明してくれる．画像から生成されたキャプションを BERT などのテキスト分類器で分類した場合，ResNet や ViT のような強力な画像分類器にどこまで迫れるだろうか?

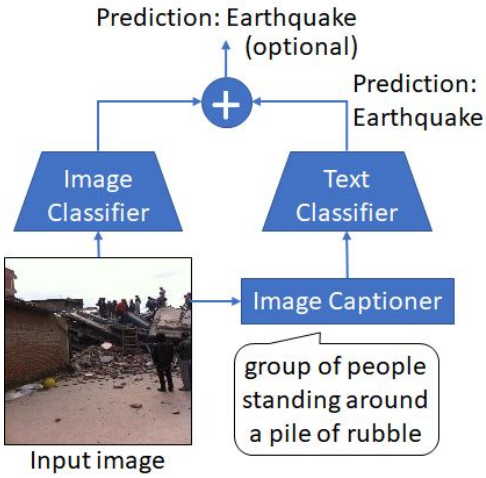

図 2: 画像分類タスクを通して画像キャプショニングモデルの振舞いを見る

災害画像データセット CrisisNLP を用いて行った実験において，言語モデルネックモデルに基づく画像分類(図 2 の右半分)は平均的な画像分類モデル(MobileNetV2，EfficientNet)を楽々と上回る分類精度を達成し，画像キャプショニングモデルによっては最高水準の画像分類モデル(ResNet-50，ViT)を凌ぐ精度を示した．

この結果は査読付き国際会議(IEEE ICIP2024)で採択され，10 月に発表予定なので詳細はそちらを参照されたい．いずれにせよこの研究はまだ初期段階であり，調べるべきことがたくさんある(災害画像以外ではどうなのか，等々)．引き続き検討を進める予定である．

参考文献

[1] W. D. Heaven, “A GPT-3 bot posted comments on Reddit for a week and no one noticed,” MIT Technology Review, Oct 2020. https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/

[2] B. Desplanques, J. Thienpondt, K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” Proc. of INTERSPEECH, Oct 2020.

[3] G. Doddington, “Speaker Recognition based on Idiolectal Differences between Speakers,” Proc. of EUROSPEECH, Sep 2001.

[4] 益子，越仲，“感情付与を用いた低評価レビューに対する応答生成，” 第 37 回人工知能学会全国大会(JSAI2023)，2023 年 6 月．

[5] S. Mitrović, D. Andreoletti, O. Ayoub, “ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text,” arXiv:2301.13852, Jan 2023.

[6] 益子，木村，越仲，“LLM 生成コンテンツの SEO 観点での品質評価，” 第 30 回言語処理学会年次大会(NLP2024)，2024 年 3 月．

[7] 有働，越仲，“画像キャプショニングは画像そのものよりも多くを語る，” 第 37 回人工知能学会全国大会(JSAI2023)，2023 年 6 月．

[8] H. Udo, T. Koshinaka, “Image Captioners Sometimes Tell More Than Images They See,” arXiv:2305.02932, May 2023.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 益子怜，越仲孝文
２．発表標題 感情付与を用いた低評価レビューに対する応答生成
３．学会等名 2023年度人工知能学会全国大会(JSAI2023)
４．発表年 2023年

１．発表者名 有働帆乃璃，越仲孝文
２．発表標題 画像キャプショニングは画像そのものよりも多くを語る
３．学会等名 2023年度人工知能学会全国大会(JSAI2023)
４．発表年 2023年

１．発表者名 小林，越仲
２．発表標題 EC サイトのレビューテキストからのレーティング予測と購買者評価の分析
３．学会等名 2022年度人工知能学会全国大会(JSAI2022)
４．発表年 2022年

１．発表者名 益子，木村，越仲
２．発表標題 LLM生成コンテンツのSEO観点での品質評価
３．学会等名 言語処理学会第30回年次大会(NLP2024)
４．発表年 2024年

１．発表者名 H. Udo, T. Koshinaka
２．発表標題 Reading Is Believing: Revisiting Language Bottleneck Models for Image Classification
３．学会等名 2024 IEEE International Conference on Image Processing (ICIP2024)（国際学会）
４．発表年 2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関