大規模言語モデルを用いた病名予測の検討

宇都宮 和希 [1]　坂野 遼平 [1]
[1] 工学院大学
em23007@ns.kogakuin.ac.jp banno@cc.kogakuin.ac.jp

概要

近年，自然言語処理技術を活用した医療支援や大規模言語モデルを用いた研究が盛んに行われており，少子高齢化による医者や病院の不足などから AI の遠隔診療への応用も挙げられている．診療の重要なプロセスとして，診断，すなわち患者の診察により病気の種類や名称を判断する行為がある．患者の曖昧な症状表現から病名を診断することは現状では医師が担っており，AI の活用，特に大規模言語モデルによる病名診断の可否や精度は明らかではない．

本研究では，大規模言語モデルによる医療支援の一環として，GPT を用いた患者表現からの病名予測の精度を検証する．また，GPT のバージョンおよびファインチューニングの有無による影響を分析する．

患者表現辞書を用いた実験の結果，ファインチューニングを行うことで予測精度が向上すること等を確認できた．

1　研究背景

近年，自然言語処理技術を活用した医療支援が始まりつつある．例えば，新型コロナウイルス感染症の動向を把握する発生届を自然言語処理を用いて自動化検討を行う取り組み [1] や，脳波を AI で解析し，認知症の診断や重症度の評価を行う実証研究 [2] などがある．少子高齢化による医者や病院の不足や AI を活用する未来として AI の遠隔診療への応用なども挙げられており [3]，医療分野における AI の需要が上昇している．

AI の医療応用に関する既存研究として，分類モデルによる疼痛表現の抽出 [4] や専門医試験における AI のパフォーマンスを評価する手法 [5] などが行われている．一方で，近年急速な発達を遂げている大規模言語モデルによる病名診断の可否や精度については明らかではない．

本研究では，患者の曖昧な症状表現から病名を予測するタスクを対象とし，データセットとして患者表現辞書 [6] を用いて，GPT による予測の精度を検証する．また，GPT のバージョンおよびファインチューニングの有無による影響を分析する．これにより，大規模言語モデルの医療応用の可能性を探る．

2　関連研究

柴田ら [4] は診療記録より事前学習した BERT と fast-TEXT，日本語 Wikipedia により事前学習を行った BERT を用いて疼痛表現の抽出を行った．結果として，診療記録で事前学習を行った BERT の F 値が最も高く，対象タスクと同じドメインのテキストで事前学習を行うことの重要性が示唆された．

野田ら [5] は日本語での自然言語処理技術と医療分野における有効性についての報告は少ないという観点から，耳鼻咽喉科専門医試験の選択肢問題に関して日本語のプロンプトと英語のプロンプト，GPT-3.5，GPT-4 などを組み合わせて，多角的に検証，評価を行い，日本語の耳鼻咽喉科領域においての有効性と AI 活用の課題について検討を行った．結果として英語プロンプトの GPT-4 が最も精度が良かった．また，日本語でも耳鼻咽喉科領域において一定の水準を達成できることが確認された．

以上のように AI の医療応用に関して研究が為されている一方，日本語を用いた大規模言語モデルによる病名の予測に関する研究は少なく，その精度や予測傾向は明らかではない．本研究では，大規模言語モデル GPT を用いて，患者表現からの病名予測に関する精度の分析を行う．

3　分析方法

3.1　大規模言語モデルによる病名予測

初期的な調査として，特定の病名と患者表現の対応関係データを学習データとテストデータに分け，大規模言語モデルによる病名予測の評価を行

う．データセットとして患者表現辞書 [6] を用いる．データの例を表 1 に示す．TYPO_出現形とは打ち間違えや言い間違えの正式な表現，ICD10 コードとは標準病名に対応する ICD10 対応標準病名マスターに記載されている ICD10 コード，Web 頻度とは出現形の Yahoo!検索における検索結果件数を表している．学習データは図 1 のように大規模言語モデルがファインチューニングを行えるフォーマットにする．作成した学習データフォーマットを用いて大規模言語モデルのファインチューニングを行い，ファインチューニングを行ったモデルにテストデータの患者表現を入力し，標準病名を出力させ，正解病名とのコサイン類似度を算出する．本研究では標準病名のみを出力させるために，学習データフォーマットに「Please answer the name of the disease you entered.」という初期設定を加える．

{“role”: “system”, “content”: “Please answer the name of the disease you entered.”}, {“role”: “user”, “content”: “できものができている”}, {“role”: “assistant”, “content”: “腫瘤"}

{“role”: “system”, “content”: “Please answer the name of the disease you entered.”}, {“role”: “user”, “content”: “できものがある”}, {“role”: “assistant”, “content”: “腫瘤"}

{"role": "system", "content": "Please answer the name of the disease you entered."}, {"role": "user", "content": "放心"}, {"role": "assistant", "content": "急性一過性精神病性障害"}

図 1　学習用フォーマット

3.2　バージョンによる差異の検証

3.1 と同じように患者表現辞書から患者表現と標準病名を学習データとテストデータの 2 つに分けて，図 1 のような学習データフォーマットを作成した後に，図 2 のように異なるバージョンの GPT モデルでファインチューニングを行う．その後，テストデータに含まれている患者表現をそれぞれの学習済みモデルに入力し，患者表現辞書の標準病名とモデルの出力病名のコサイン類似度から評価を行い，類似度やプロンプト変化によるモデルごとの差異調査を行う．

4　評価

実験環境として Google Colaboratory[7] を用いた．患者表現辞書から標準病名が「幻覚」，「疼痛」，「発熱」であるもの計 7 件をテストデータとして取り除いた．テストデータ以外の 6387 件を学習データとして GPT-3.5[8] でファインチューニングを行い，ファインチューニングを行ったモデルと行っていないモデルに，テストデータの患者表現を入力し，病名を出力させた．出力病名とコサイン類似度の結果を表 2 に示す．なお表 2 ではファインチューニング無しを N、ファインチューニング有りを F として「GPT-3.5(N)」のように表記している．

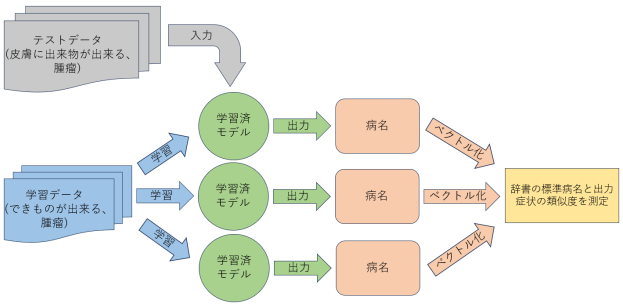

図 2　調査手法

また，バージョンによる差異の検証として患者表現辞書を GPT-3，GPT-3.5 にファインチューニングし，辞書にある標準病名「運動性失語症」，「体感異常」，「疼痛」に対応した患者表現を各バージョンのモデルに入力し，病名を出力した．入力した患者表現と出力結果，コサイン類似度を表 3，表 4 に示す．なお，ファインチューニングを行わず，初期設定に「You are a doctor.」という性格を付与したモデルを GPT-3.5(C) と記載している．

4.1　評価結果

実験において明確な病名が出力されなかったケースがあり，表中では「N/A」と記載している．例えば「あなたが言及している病名はありません」などがある．表 2 の結果から，ファインチューニングを行うことでより正確に病名を予測できていることがわかる

表 3, 表 4 の結果から，GPT3.5(N)，GPT-3.5(C) では「ジクジク」という感覚的な患者表現に対して病名を出力できていない一方，ファインチューニングを行った GPT-3.5(F) では正確に出力できており，他の病名についても比較的高いコサイン類似度が得られている．GPT-3 と GPT-3.5 のいずれにおいても，ファインチューニングを行うことでより正解病名に近い出力ができる傾向を確認することができた．

また，表 4 からファインチューニングを行った GPT-3 のコサイン類似度があまり向上しなかった理由として，データセットの品質が関係していると考えられる．患者表現辞書には患者表現と対応した標準病名が同じ語句になっている重複プロンプトがい

表 1　患者表現辞書 (一部抜粋)

出現形（患者表現）	TYPO 出現形	ICD-10	標準病名	Web 頻度
できものができている		R229	腫瘤	3930000000
できものがある		R229	腫瘤	3520000000
放心		F239	急性一過性精神病性	3420000000
できもん	できもの	R229	腫瘤	3140000000
できものができる		R229	腫瘤	2730000000
おでき		L029	せつ	2720000000

くつかあり，重複プロンプトの削除などデータセットの質を向上させることで，ファインチューニングを行った GPT-3 のパフォーマンスがより向上する可能性がある．

5　おわりに

本研究では，曖昧な患者表現から病名を予測するタスクに着目し，GPT を用いて予測精度の検証を行った．GPT のバージョンやファインチューニングの有無による影響を分析し，ファインチューニングを行った GPT-3.5 が他のパターンと比べより正確に病名を出力できることを確認した．

今後の課題としては以下の点が挙げられる．

- GPT 以外の生成モデルやプロンプトの変化による差異の調査
- 病名予測結果に対する人手による評価
- アンサンブルモデルの適応検討

参考文献

[1] 福本拓也, 坂根亜美, 村松俊平, 五十嵐正尚, 狩野芳伸, 荒牧英治, 堀口裕正, 奥村貴史. 新型コロナ感染症発生届の分析-記載における非効率と自然言語処理による解決への課題と展望-. 人工知能学会全国大会 (第 37 回), 2023.

[2] 琢史, 貴島晴彦, 原田達也, 数井裕光, 吉山顕次, 吉村匡史, 西田圭一郎, 畑真弘. 脳波・脳磁図を用いた ai 解析による認知症の診断・重症度評価に関する実証研究. 科学技術情報発信・流通総合システム, 2020.

[3] 中尾睦宏. Ai，ict，vr を活用する未来に向けて. バイオフィードバック研究, Vol. 48, No. 1, pp. 12–15, 2021.

[4] 柴田大作, 河添悦昌, 嶋本公徳, 篠原恵美子, 荒牧英治. 診療記録で事前学習した bert による疼痛表現の抽出. 第 24 回日本医療情報学会春季学術大会 シンポジウム 2020 web, 2020.

[5] 野田昌生, 上野貴雄, 甲州亮太, 島田 Dias 茉莉, 伊藤真人, 矢本成恒, 吉崎智一, 野村章洋. 耳鼻咽喉科専門医試験における generative pretrained transformer の有効性に関する検討. 科学技術情報発信・流通総合システム, 2023.

[6] 西谷実紘, 矢田竣太郎, 若宮翔子, 荒牧英治. 生成アプローチによる患者表現の標準化. 科学技術情報発信・流通総合システム, 2021.

[7] Colaboratory へようこそ, (2023 年 1 月 7 日閲覧). https://colab.research.google.com/?hl=ja.

[8] OPEN AI documentation Models, (2024 年 1 月 7 日閲覧). https://platform.openai.com/docs/models.

表 2　GPT-3.5(N)(F) による出力病名とコサイン類似度

入力テキスト	正解の標準病名	GPT-3.5(N) 出力病名	GPT-3.5(F) 出力病名	GPT-3.5(N) 類似度	GPT-3.5(F) 類似度
幻	幻覚	N/A	知覚障害	0	0.775
あるはずの無いものが感じられる	幻覚	N/A	幻覚	0	1.0
ねつ	発熱	発熱	発熱	1.0	1.0
熱がある	発熱	発熱	発熱	1.0	1.0
グリグリ	疼痛	N/A	疼痛	0	1.0
ズキンズキン	疼痛	N/A	疼痛	0	1.0
痛みがある	疼痛	N/A	疼痛	0	1.0
平均コサイン類似度				0.286	0.968

表 3　GPT のバージョン等による差異

入力テキスト	正解標準病名	GPT-3(N)	GPT-3(F)	GPT-3.5(N)	GPT-3.5(C)	GPT-3.5(F)
運動後に言葉が出ない	運動性失語症	運動後無言症	マイクロアビリティ障害	運動性失語症	運動後発作	運動性呼吸困難
感覚がわからない	体感異常	感覚不全症	振戦	感覚障害	感覚消失	感覚障害
ジクジク	疼痛	ジストニア	腹部痛	N/A	N/A	疼痛

表 4　GPT のバージョン等による差異（コサイン類似度）

入力テキスト	正解の標準病名	GPT-3(N)	GPT-3(F)	GPT-3.5(N)	GPT-3.5(C)	GPT-3.5(F)
運動後に言葉が出ない	運動性失語症	0.908	0.776	1.0	0.820	0.874
感覚がわからない	体感異常	0.852	0.712	0.855	0.803	0.855
ジクジク	疼痛	0.408	0.722	0	0	1.0
平均コサイン類似度		0.723	0.737	0.618	0.541	0.910