GPT を用いた退院サマリの自動生成に関する性能評価について

中谷亮太[1] 廣淵亮太[1] 佐藤貴俊[1] 篠原恵美子[2] 岡本康宏[1]
有田悠人[1] 有田隼也[1] 佐藤敏紀[1] 河添悦昌[2] 大江和彦[2]
[1]ソフトバンク株式会社 [2]東京大学大学院 医学系研究科
[1]{ryota.nakatani01, ryota.hirobuchi,takatoshi.sato01, yasuhiro.okamoto01, haruto.arita,
shunya.arita, toshinori.sato}@g.softbank.co.jp
[2]{emikoy, kawazoe}@m.u-tokyo.ac.jp [2]kohe@hcc.h.u-tokyo.ac.jp

概要

退院サマリは，退院後の患者のケアを引き継ぐ医療関係者や医療機関に対して，診断結果や入院中の治療に関して正確な情報を共有するのに重要な役割を果たす．その一方で，退院サマリの作成には大量のデータから必要な情報を取捨選択する必要から多くの時間と労力がかかり，多忙な医師にとっては削減したい作業の一つである．本研究では，大規模言語モデルの一つである GPT を用いて，患者情報及び看護記録等のデータから退院サマリを自動生成することを目指す．具体的には，患者基本情報及び看護記録等から引用元を明らかにしつつ，退院サマリを自動生成し，医師が作成した退院サマリと比較することで，生成された退院サマリの評価を行う．検証の結果，GPT が生成する退院サマリは比較的簡素に生成され，データ全体から満遍なく情報を抽出するセクションに関しては精度が低いことが確認された．

1 はじめに

少子高齢化により医師の人材不足と患者数の増加が懸念されており，医師の業務負担が増加することが問題視されている．また，2024 年 4 月より施行されている医師の働き方改革により，時間外労働の上限規制が設けられたことから，医療現場における業務の効率化が一層求められている．

医療現場の過重労働の要因の一つとして，診療には直接関係しない間接的業務の多さが挙げられる．医療現場では診察記録や経過記録，紹介状などの多岐にわたる医療文書の作成に日々多くの時間が費やされており，瀬戸らの報告によると，医師の間接的業務が 1 日当たり 3 時間を占め，その約 4 分の 1 が電子カルテの操作に費やされている[1]．また，全国医師ユニオンが実施した勤務医労働実態調査によると，勤務医の診療環境改善に必要な国の政策として，医師の事務的作業を代行する医療クラークなどの医療補助職の増員が 1 位と重要視されており，間接的業務の軽減が強く望まれていると言える[2].

中でも，退院サマリの作成は医療現場において重要なものの，その作業には多くの時間と労力がかかる．本研究では，大規模言語モデルの一つである GPT を用いて，患者情報及び看護記録などのデータから退院サマリを自動生成し，医師が作成した退院サマリと比較することで，生成された退院サマリの評価を行う．患者情報などの医療データには多くの個人情報が含まれているため，GPT をはじめとしたクラウドサービスの利用には慎重な検討が必要である．本研究では，個人が特定できないようにマスキングされた医療データを用いることで，検証を可能としている．

2 関連研究

近年，自然言語処理の技術を利用した医療分野への応用は盛んに行われている．河添らは，1 億 2000 万件の日本語で記載された診療記録をもとに BERT の事前学習を行なった UTH-BERT を公開しており，医療ドメインに特化したモデルの精度を報告している[3]．宇野らは，大規模言語モデルを利用して診療録から治療経過サマリの作成支援アプリの試作を行っており，石川らはその評価を行なっている[4, 5]．また，Emre らは，入力及び音声の文字起こしにより得られたテキスト情報から，NER 技術を用いて薬剤や症状の情報を抽出した結果を報告している[6]．このように，自然言語処理の技術を利用した医療分野への応用が盛んに行われている．

また，退院サマリに応用した研究も行われている．Arne らは，電子健康記録から ChatGPT-4 を用いて精神科における退院サマリの生成を行い，研修医によって作成された退院サマリとの比較を行った．その結果，研修医が作成した退院サマリの方が高品質で

あるものの，一部のケースに関してはChatGPT-4が生成した退院サマリの品質も適切であると報告している[7]. Arneらは，退院サマリの中から入院中の経過についてまとめた入院経過のセクションに焦点を当て，様々なモデルによる精度検証の結果を報告している[8]．このように，大規模言語モデルを利用した退院サマリの自動生成に関する研究は存在するものの，退院サマリの一部のセクションに限定されることが多く，医師が作成した精度には及ばない結果となっている．

3 方法

以下に述べる実験は，東京大学大学院医学系研究科の倫理委員会の承認を得て，関連する倫理指針及び規則に従って実施した (審査番号:2023340NI).

3.1 データセット

東京大学医学部附属病院の電子カルテから 10 症例の患者基本情報と入院時記録，退院サマリ (以下，医師サマリ) を抽出した．これらデータを匿名加工医療情報作成事業者が匿名加工することで，氏名や組織名，日付，電話番号といった個人を特定しうる情報をマスキングされたものを研究利用した．表 1 に，患者基本情報と入院時記録の詳細を示す．

3.2 実験手法

本研究では，大規模言語モデルの一つであるGPTを用いて，患者基本情報及び入院時記録 (以下，入力データ) から退院サマリ (以下，GPT サマリ) の自動生成を行う．GPTのモデルとしては，GPT-4oを利用する．図1は検証の流れを示している．

3.2.1 セクション別の精度検証

退院サマリには退院時診断や薬歴，現病歴といった複数のセクションが含まれる．本研究では，日本医療情報学会の「退院サマリー作成に関係するガイダンス」によって提案されるセクション毎に精度検証を行う．実サマリにおけるセクションは表記揺れが生じているため，対応付けによってセクション名を正規化した (表 2).

3.2.2 出典元の提示

GPT が生成する GPT サマリはハルシネーションを引き起こす可能性があるが，ハルシネーションの有無を確かめるには，大量の入力データの中から

表 1　患者基本情報と入院時記録の詳細

データの種類	定義
患者基本情報	身長・体重，喫煙歴，飲酒歴，アレルギー歴
処方オーダ	処方された薬剤情報
注射オーダ	処方された注射情報
検査記録	実施した診断結果の記録
看護記録	実施したケアや観察結果，患者の状態，看護計画や評価の記録
経過記録	患者の診療経過や治療の進捗状況の記録

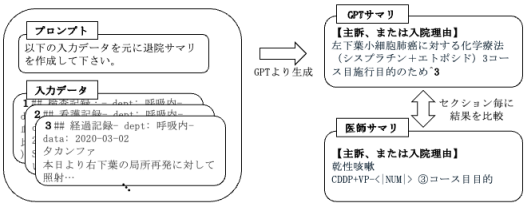

図 2　検証の流れ

GPT サマリの作成に利用されたデータを探しだす必要があるため，検証には時間を要する．そこで，本研究では入力データの書類毎にタグ付けを行い，GPT サマリを生成する際に利用したデータのタグを同時に出力することで，出典元を提示する．これにより，GPTサマリにおけるハルシネーションの確認が容易になることが期待される．

3.3 評価指標（定量評価）

本研究には，評価指標として BLEU[9]，ROUGE[10], BERTScore[11]の 3 つの指標と，医師サマリとGPTサマリの出典元の精度評価としてRecallを用いる．

3.3.1 BLEU

BLEUスコアはPrecisionベースの評価指標であり，0~1 の範囲の実数で表され，類似度が高いほど値が高くなる．Tokenizerとしてはjanomeを利用し，1~4-gram の平均を求めた．

3.3.2 ROUGE

ROUGEスコアはRecallベースの評価指標であり，0~1 の範囲の実数で表され，類似度が高いほど値が高くなる． Tokenizerとしては Mecab を利用し，医師サマリと GPT サマリ間の最長共通部分列を評価する，ROUGE-L を利用した．

表 2　セクション一覧と表記揺れの例及び各セクションの定義

セクション名	表記揺れの例	定義
退院時診断	診断名，problem list	主たる病名と関連する合併症などを記載
アレルギ・不適応反応	アレルギー，Allergy	アレルギーの有無を記載
主訴、または入院理由	主訴，入院目的	患者が発症した主発症，または入院した主な理由を記載
既往歴	既往歴，併存症	これまでに罹った病気について記載
家族歴	家族歴，家族	親族や同居者の既往歴について記載
社会生活歴	生活歴，職業歴	飲酒歴や喫煙歴，生活習慣，仕事などについて記載
薬歴	内服歴，入院時使用薬	入院時に利用している内服薬や常備薬について記載
現病歴	現病歴，病歴	入院に至るまでの経緯や入院適応の判断内容について記載
入院時身体初見	入院時現象，Vital	入院時の身長・体重やバイタルサイン，全身の状態について記載
入院時検査初見	検査初見，血液検査	血液検査や尿検査といった検体検査の結果について記載
入院経過	入院後経過，経過	入院中の経過についてまとめたものを記載
退院時使用薬剤情報	退院処方，退院後処方	退院時に処方された薬剤情報について記載
退院時方針	今後の方針，今後予定	療養生活上の注意点や今後の受診計画などについて記載

3.3.3 BERTScore

BLEU 及び ROUGE は n-gram に基づく評価指標であるため，文章の意味を適切に評価するには限度がある．そのため，文章の意味を考慮した評価指標である BERTScore でも評価を行う．BERTScore は 0~1 の範囲の実数で表され，類似度が高いほど値が高くなる. BERTScore を求める数式は以下の通りである．

$$R_{BERT} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{x}}_j$$

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{x}}_j$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

3.3.4 Recall

退院サマリの各セクションには記載すべき内容が定められており，情報の抜けがないことが求められる．したがって，医師サマリに記載されている情報が GPT サマリでどれだけ網羅されているかを示す Recall が，評価の重要な指標となる．本研究では，GPT が出力した出典元と医師サマリの出典元から Recall を算出することで，GPT サマリ内に記載すべき情報がどの程度含まれているかを評価する．医師サマリの出典元のアノテーションは 2 症例に対して 2 名体制で実施した．

4 結果・考察

4.1 結果

実験の結果を表 3 に示す．表 3 はセクション毎の精度結果をまとめており，BLEU，ROUGE，BERTScore は 10 症例の平均の結果を表している．また，Recall は 2 症例の平均の結果を表している．表の結果から，BLEU，ROUGE，BERTScore に関しては，「主訴、または入院理由」や「社会生活歴」などで精度が高く，「家族歴」や「退院時方針」などで精度が低いことがわかる．また，Recall に関しては「アレルギ・不適応反応」や「薬歴」で精度が高く，「現病歴」や「入院経過」などで精度が低いことがわかる．

4.2 考察

表 3 の結果から，BLEU，ROUGE，BERTScore に関しては，「主訴、または入院理由」や「社会生活歴」などで精度が高く，「家族歴」，「退院時方針」などで精度が低いことが明らかとなった．「主訴、または入院理由」や「社会生活歴」で精度が高い理由の一つとして，正解データである医師サマリが比

表 3　患者基本情報と入院時記録の詳細

セクション	BLEU (n=10)	ROUGE-L (n=10)	BERTScore (n=10)	Recall (n=2)
退院時診断	0.067	0.274	0.722	–
アレルギ・不適応反応	0.287	0.316	0.715	1
主訴、または入院理由	0.171	0.365	0.776	0.6
既往歴	0.124	0.206	0.721	0.5
家族歴	0.06	0.168	0.638	–
社会生活歴	0.157	0.296	0.737	0.5
薬歴	0.049	0.211	0.704	1
現病歴	0.016	0.122	0.66	0.393
入院時身体所見	0.012	0.164	0.656	0.75
入院時検査所見	0.017	0.141	0.676	0.643
入院経過	0.014	0.134	0.669	0.072
退院時使用薬剤情報	0.121	0.226	0.708	0.5
退院時方針	0.017	0.119	0.633	0.75

較的簡素に記載されていることが挙げられる．GPT サマリは比較的簡素に表現する傾向があるため,「現病歴」や「入院経過」といった長文で記載されるセクションに比べて精度が高くなったと考えられる．

一方で，「家族歴」の精度が低い理由としては，そもそも家族歴に該当する情報が 10 症例中 2 症例しか入力データに含まれておらず，該当するデータが入力データに含まれていない割合が多いことが挙げられる．大規模言語モデルは入力データに依存した生成を行うため，入力データに該当する情報が含まれていない割合が高い「家族歴」に関しては，生成が困難であったと考える．

必要な情報を抽出できているかの指標である Recall に関しては「アレルギ・不適応反応」や「薬歴」で精度が高く，「現病歴」や「入院経過」などで精度が低いことがわかった．「アレルギ・不適応反応」や「薬歴」で精度が高い理由としては，それぞれ入力データの中で「アレルギー歴」，「処方オーダ」が該当する出典元であり，出典元が明確であったことが要因であると考える．一方で,「現病歴」と「入院経過」の精度が低い理由として，これらのセクションは入力データ全体から満遍なく情報を抽出して記載する必要がある点が挙げられる．それら以外のセクションに関しては，入力データのごく一部のみで十分な情報が得られるものが多い．しかし，「現病歴」と「入院経過」に関しては，入力データ全体から参照する必要があり，そのまとめ方も医師個人に大きく影響される．そのため，必要な情報を効果的に抽出できているかを示す指標である Recall の結果が悪かったと考えられる．

5 おわりに

本研究では，大規模言語モデルの一つである GPT を用いて，患者情報及び看護記録等のデータから退院サマリを生成し，医師が作成した退院サマリと比較することで，生成された退院サマリの評価を行った．その結果，GPT が生成する退院サマリは比較的簡素に生成され，データ全体から満遍なく情報を抽出するセクションに関しては精度が低いことが確認された．今後の展望として，医師サマリの出典元をアノテーションした症例が現時点では 2 症例に限られているため，更なる症例数の増加を図った上で Recall を求めることが挙げられる．また, zero-shot プロンプティングから few-shot プロンプティングへの移行による精度検証も今後の展望として挙げられる．

参考文献

1. 瀬戸僚馬, 津村宏. 医師が電子カルテ操作に費やす業務時間に関する調査. 医療情報学, Vol. 32, No. 2, pp. 59-63, 2012.
2. 植山直人, 勤務医労働実態調査 2022 概要, 2024 年 12 月 25 日閲覧, http://union.or.jp/wordpress/wp/wp-content/uploads/2022/10/%E5%8B%A4%E5%8B%99%E5%8C%BB%E5%8A%B4%E5%83%8D%E5%AE%9F%E6%85%8B%E8%AA%BF%E6%9F%BB2022%E6%A6%82%E8%A6%81-%E3%80%80%E6%9C%80%E7%B5%82%E7%89%88%E3%80%802022.10.21.pdf.
3. Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific BERT developed using a huge Japanese clinical text corpus. Plos one, Vol. 16, No. 11, 2021.
4. 宇野裕, 石井亮, 柴田大作, 石川開, 定政邦彦, 渋谷恵, 辻川剛範, 中川敦寛, 小山田昌史, 久保雅洋, 香取幸夫. 大規模言語モデルを利用した治療経過サマリの作成支援アプリの試作. 人工知能学会全国大会論文集, 2024.
5. 石川開, 宇野裕, 石井亮, 定政邦彦, 柴田大作, 辻川剛範, 中川敦寛, 小山田昌史, 久保雅洋, 香取幸夫. 大規模言語モデルを用いて診療録から生成した治療経過サマリの評価. 人工知能学会全国大会論文集, 2024.
6. Emre Sezgin, Syed-Amad Hussain, Steve Rust, and Yungui Huang. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. JMIR Formative Research, Vol. 7, 2023.
7. Arne Schwieger, Katrin Angst, Mateo de Bardeci, Achim Burrer, Flurin Cathomas, Stefano Ferrea, Franziska Grätz, Marius Knorr, Golo Kronenberg, Tobias Spiller, David Troi, Erich Seifritz, Samantha Weber, and Sebastian Olbrich. Large language models can support generation of standardized discharge summaries–A retrospective study utilizing ChatGPT-4 and electronic health records. International Journal of Medical Informatics, Vol. 192, 2024.
8. Thomas Searle, Zina Ibrahim, James Teo, and Richard J.B. Dobson. Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. Journal of Biomedical Informatics, Vol. 141, 2023.
9. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Association for Computational Linguistics, pp. 311-318, 2002.
10. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. Text summarization branches out, 2004.
11. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. International Conference on Learning Representation, 2020.

A 付録

表 4　医師サマリと GPT サマリの比較

セクション名	医師サマリ	GPT サマリ
アレルギ・不適応反応	food (-)、drug (-)、asthma (-)、contrast medium (-)	アレルギー歴：記載なし^4
主訴、または入院理由	乾性咳嗽 CDDP+VP-16 ③コース目目的	左下葉小細胞肺癌に対する化学療法 3 コース目施行のため入院^10
既往歴	高血圧(<|NUM|>より薬物治療開始)、前立腺癌(<|NUM|>歳 全摘出)	肝機能障害^11, 21 前立腺癌^11, 21 高血圧^11, 21
家族歴	父：心筋梗塞、母：胃潰瘍、脳溢血。 兄：脳溢血。	なし
社会生活歴	喫煙：30 本/日×50 年 飲酒：7 回/週×50 年（日本酒・焼酎 100ml）	喫煙歴：30 本/日 × 50 年（<|NUM|>歳まで）^2 飲酒歴：日本酒 100ml^3
薬歴	コデインリン酸塩酸 1％ 40mg2x, 酸化マグネシウム錠 250mg 4T2x, アジルサルタン(20mg) 1T1x	コデインリン酸塩散 1%^5, 27
退院時方針	<|NUM|> 呼吸器内科 入院予定 CDDP＋ETP④ 施行予定	次回の化学療法（4 コース目）を予定し、経過観察を継続する。食事や活動の制限はなし。化学療法による有害事象が出現した場合には早めに受診を^34, 37.

表 5　出典元となった入力データの例

出典元 ID	内容
2	喫煙歴:30 本 ×50 年(<|NUM|>歳まで)
3	飲酒歴:日本酒 100ml
4	アレルギ歴:[薬剤]:記載なし, [食物]:記載なし, [造影剤]:記載なし, [その他]:記載なし
5	コデインリン酸塩散１％ 40.0 ｍｇ 1 日 2 回 朝夕食後 24 日分
10	S:前回は 5 日くらいで退院できたんだよ。今回も早く退院できるといいね。今からもうマスクしてるよ。 O:左下葉小細胞肺癌に対し化学療法（シスプラチン＋エトポシド）3 コース目施行目的で入院された。化学療法は<|DATE|>より開始となる予定。時折乾性咳嗽見られている。 (以下省略)
11	＃左肺下葉原発肺門部小細胞癌　cT4N3M1a→IVA 期、進展型（左肺上下葉と右肺下葉に癌性リンパ管症）<|NUM|>- CDDP+ETP①　<|NUM|>- CDDP+ETP②　<|NUM|>- CDDP+ETP③(予定)　＃肝機能障害　＃HBcAb(+)(HBV-DNA 検出せず(<|NUM|>))　＃高血圧　＃前立腺癌よろしくお願いします。BT: 35.8℃, HR: 95 , BP: 132/76 mmHg, RR: 18/min, SpO2: 95%(r/a)　[ECG] HR 96bpm, NSR, ST-T change(-)　[CXR] CTR 47%, CP angle sharp　<|NUM|>- CDDP+ETP③(予定)
34	＃左下葉原発小細胞癌 cT4N3M1aIVA 期 全然変わりないですよ。いつもそうです。今回も入院したら次の入院の日すぐ決まってね。 vital 著変なし ＃左下葉原発小細胞癌 cT4N3M1aIVA 期　<|DATE|>- CDDP+VP-16③,day3.　本日も問題なく投与終了。予定通り明日退院。 次回<|DATE|>（4 週回し）入院予約済