大規模言語モデルによる日本語診療テキストからの人名抽出

岩瀬裕哉 [1] 柴田大作 [1] 大西颯真 [1] 辻川剛範 [1] 渡辺純子 [1]
石井亮 [2] 中川敦寛 [2] 香取幸夫 [2] 久保雅洋 [1]
[1] 日本電気株式会社 [2] 東北大学病院
{yuya-iwase,daisaku-shibata,soma-onishi,tujikawa,junta_n,masahirokubo}@nec.com

概要

診療テキストの二次利用において，テキスト中の個人識別情報を処理し，個人を識別不能にすることは重要である．本稿では識別情報の中でも特に人名を対象とし，エンコーダ型事前学習済言語モデル（BERT）を微調整したモデルと，zero-shot 設定のデコーダー型大規模言語モデル（LLM）の間で，疑似人名を埋め込んだ日本語診療テキストからの人名抽出を比較した．実験の結果，BERT は評価データ全体で F1 が 0.971 と高い性能を示した一方，BERT 微調整時の学習データに含まれない人名に対して LLM が BERT を上回ることが明らかとなった（0.899 vs 0.876）．この実験結果に分析と考察を加え，人名抽出モデルの実応用に向けた知見を示す．

1 はじめに

病院には経過記録などの診療テキストが大量に蓄積されており，これらの診療テキストを二次利用することで，臨床研究などの促進を図り，医療の質の向上に寄与することが期待されている [1, 2]．しかし，診療テキストには，患者や医師の氏名に代表される，個人を識別できる情報が多く含まれている．したがって，個人情報保護の観点から，診療テキスト中の個人識別情報に対して，仮名化または匿名化を施し，個人を特定不可能にする処理（非識別化）を施すことが二次利用の前に求められる場合がある．一方で，人手による非識別化は時間的・金銭的なコストが大きく [3]，診療テキストの二次利用を推進する上での障壁となっている．

そのため，自然言語処理技術により診療テキストの非識別化を試みる研究がいくつか報告されている．これまでには，例えば Bidirectional Encoder Representations from Transformers（BERT）[4] をはじめとする事前学習済みエンコーダモデルを，識別情報に関する情報がアノテーションされたデータにより微調整を行った非識別化モデルが多数提案されている [5, 6, 7]．加えて，非識別化モデルの実応用を推進するためには，モデルが入力データの性質によらず頑健に動作する必要があるが，診療テキストは施設や文書種別によって記載様式や語彙が異なるため，あるデータセットで微調整したモデルを別施設・別データセットへそのまま適用すると性能が低下することが報告されている [8, 9, 10]．

さらに近年，大規模言語モデル（Large Language Model: LLM）を用いた非識別化手法も提案されている [11, 12, 13]．これらの手法は，追加学習を行わない zero-shot あるいは few-shot 設定であっても，高い性能で非識別化を行えると報告している [11, 14]．しかし，LLM による非識別化は，二次利用に有益な臨床情報まで過剰に抽出してしまうことや，再現性・実用性についての課題も指摘されている [15]．

このように診療テキストの非識別化に関する研究はいくつか報告されているが，その多くは英語の公開データセットで実施されたものであり，日本語の診療テキストを対象とした検討は限られている [16]．加えて，日本語診療テキストに対する非識別化タスクにおいて，エンコーダモデルと LLM の長所・短所の比較と分析を行った研究は，筆者らの調査の限りでは見つけられなかった．

そこで本研究では，東北大学病院で作成された日本語の経過記録を対象に，個人識別情報の中でも診療テキストに頻出する人名に注目して，BERT と LLM を用いた人名抽出モデルをそれぞれ構築し，その性能を比較する．また，入力データの性質による実応用上の課題を捉えるため，BERT の微調整に用いた学習データに含まれる人名の集合に基づいて評価データを分割し，未学習人名を含む入力に対する各モデルの頑健性に焦点を当てた分析を行う．[1)2)]

1） 本稿で説明する内容は東北大学病院医学系研究科倫理委員会の承認 (承認番号: 2022-1-851) を得て実施された．
2） 本研究で使用した診療テキストは事前に東北大学病院内で仮名加工した上で，NEC 内部へ転送し，実験を行った．

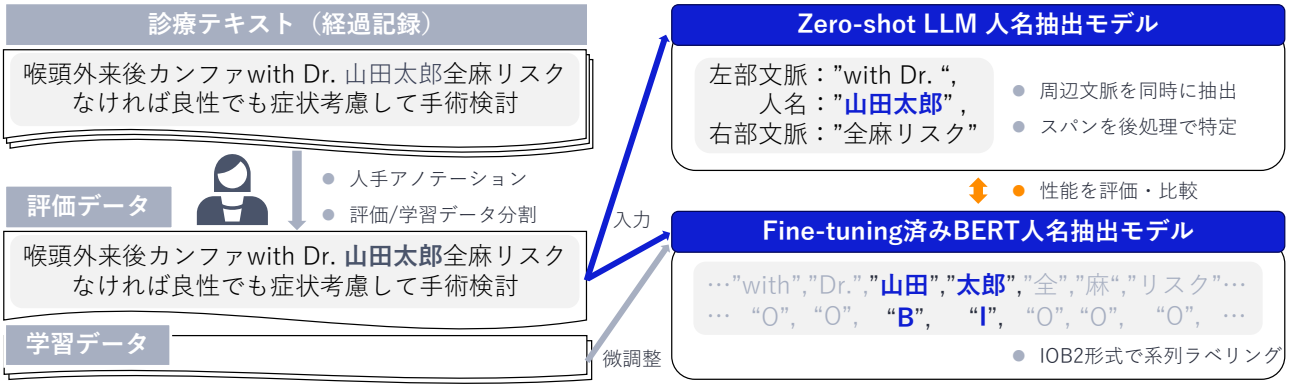

図 1　本研究で扱う人名抽出タスクの概要

2　実験データ

2.1　データセットの概要

東北大学病院で作成された 38 診療科の，既に個人識別情報が削除された経過記録に対し，4 名の作業者が個人識別情報の種類を細分化するアノテーション作業を行った．その後，個人識別情報が削除されていた部分に，疑似的な個人識別情報をランダムに埋め込み，テキスト中の人名部分を自動的に特定できるよう，文字区切りのインデックスで正解のスパン情報を付与した．学習データとして 18,755 件，評価データとして 6,715 件を使用する．評価データに関してはさらに 1 名の作業者がすべてのアノテーション内容を確認し，修正を行った．

2.2　評価データセットの構築

個人識別情報のうち本研究で扱う人名に注目すると，日本人の氏名のうち，苗字の 8 割は地名に由来する [17]．したがって，ある病院に通院する人々は近隣の住人が多くなると予想されるため，病院の所在によって診療テキストに現れる人名の分布が異なり，非識別化モデルの性能低下につながる可能性がある．そこで，BERT における未学習語彙への頑健性を分析するため，学習データ全レコードから記号除去等の正規化を施した人名集合を作成し，評価データ中の 1 レコードから同様に抽出した人名集合と照合し，評価データのサブセットを作成した．すなわち各レコードは，自身に含まれる人名集合に基づき，次の 3 つの区分に分類される．

- **既知**: すべて学習データに含まれる
- **未知**: すべて学習データに含まれない
- **混在**: 上記 2 種の人名が混在

表 1　評価データ統計情報（() 内はユニーク数）

区分	-	レコード数	人名数
既知	-	5,593	11,095（893）
未知	-	390	409（359）
混在	既知	732	3,075（633）
	未知		894（626）
全体	-	6,715	15,473（1,802）

なお，人名を含まないレコードは今回の評価データから除外されている．区分ごとのレコード件数，人名の出現数およびユニーク人名数は表 1に示す．

3　実験

実験の概要を図 1に示す．本節では，BERT・LLM のそれぞれを用いて人名抽出モデルを構築し，性能を評価・比較する．BERT の微調整には，前節で述べた 18,755 件の学習データを使用し，性能評価には BERT・LLM ともに 6,715 件の評価データを用いた．LLM については，公開されている学習済みモデルに対して追加の学習を行わず，zero-shot 設定で推論のみを実施した．

3.1　人名抽出モデル構築

BERT 人名抽出モデル　長文入力への対応と高速な推論を特徴とする事前学習済みモデルである，ModernBERT-Japanese 310M[18] を微調整することで人名抽出モデルを構築する．ここでは，人名抽出を固有表現認識（Named Entity Recognition; NER）とみなし，IOB2（Inside–outside–beginning）形式ラベルを付与する系列ラベリングとして定式化する．入力テキストはまず ModernBERT 付属の SentencePiece トークナイザによりトークン列へ変換される．次に ModernBERT の最終層から各トークンの埋め込み表現を得て，線形層により各ラベル（B・I・O）に対

するスコアを出力する．そして，このスコア列の上に条件付き確率場 [19] を重ねることで，ラベル間遷移を一括して最適化し，IOB2 形式における構造的一貫性を保ちながら系列全体のラベル列を推定する．予測されたラベル列から連続スパンを抽出し，各スパンを構成する最初のトークンの開始位置と最後のトークンの終了位置をトークナイザが返す offset_mapping から取得することで，トークン単位の範囲を文字インデックスの範囲へ変換した．

微調整時の学習パラメータは，最大シーケンス長を 1,024，バッチサイズを 16，学習率を 1e-5, エポック数を 30 に設定し，他の値はデフォルトを使用した．

LLM 人名抽出モデル　BERT 人名抽出モデルとの比較と，人手アノテーション済みの学習データを得られない状況における実応用を想定して，LLM を用いた zero-shot 人名抽出モデルを構築する．ここでは人名抽出タスクを，入力テキストに含まれる人名エンティティ候補と周辺文脈を同時に構造化して生成するタスクとして定式化する．実験では，個人情報保護の観点から，ローカル環境で動作可能なモデルである Qwen3-32B[20] を選択し，推論時の temperature は 0.6，top-k は 20，top-p は 0.95 に設定した．また，Singh ら [12] により提案された手法に着想を得て，人名と左右文脈を構造化して出力させる指示とともに，人間のアノテーション作業時に使用した基準に相当する指示を含むプロンプトを作成した．実験では常にこのプロンプトを用い，処理対象のテキストをプロンプトに埋め込み，LLM に入力することで出力を得た．その後，出力された人名は入力文へ照合し，文字スパンを復元した．原文中に同一表記が複数回出現する場合，左右文脈の文字列マッチングに基づき尤もらしいスパンを推定した．

3.2　評価指標

本実験では，抽出されたエンティティの開始位置と終了位置が，正解アノテーションと完全に一致した場合のみを正解とみなし，Precision・Recall・F1-score（F1）の三評価指標を用いて評価する．ここで TP（True Positive）は正しく抽出されたエンティティ数，FP（False Positive）は不要なエンティティを誤って抽出した数，FN（False Negative）は本来抽出すべきエンティティを見逃した数を表す．Precision は TP/(TP + FP)，Recall は TP/(TP + FN) と定義する．F1 は Precision と Recall の調和平均と定義する．

表 2　評価結果（評価データ全体および各分割）

Model	区分	Precision	Recall	F1
BERT	overall	**0.973**	**0.970**	**0.971**
	既知	**0.974**	**0.973**	**0.974**
	未知	**0.875**	0.878	0.876
	混在	**0.980**	**0.970**	**0.975**
LLM	overall	0.914	0.949	0.932
	既知	0.912	0.949	0.930
	未知	0.873	**0.927**	**0.899**
	混在	0.947	0.948	0.948

3.3　実験結果

表 2に BERT 人名抽出モデルおよび LLM 人名抽出モデルの評価結果を示す．評価データ全体（overall）では，BERT の Precision が 0.973，Recall が 0.970，F1 が 0.971 とすべての指標において高い性能であることが確認された．一方，LLM は，Precision が 0.914，Recall が 0.949，F1 が 0.932 となり，適合率に比べ再現率は高い傾向が見られるが，BERT と比較して全指標で性能が低下する結果となった．

区分別に見ると，学習データに含まれる人名のみからなる既知データでは，BERT の F1 が 0.974，LLM の F1 が 0.930 と，BERT が一貫して高性能であった．これに対し，学習データに含まれない人名のみからなる未知データでは，BERT の F1 が 0.876 であるのに対し，LLM は F1 が 0.899 であり，zero-shot LLM の方が未学習語彙への汎化性能に優れる結果となった．混在データにおいては，両モデルとも高い性能を示したものの，BERT の F1 が 0.975，LLM の F1 が 0.948 と，既知データを多く含む設定では引き続き BERT が優位であった．

3.4　分析・考察

LLM 人名抽出モデルの誤り分析　表3に，評価データ全体に対する LLM 人名抽出モデルの誤りを類型化し，著者が目視で確認した各類型の代表例を示す．ここでは，LLM の誤りを，正解アノテーション（正解）と予測スパン（予測）の対応関係に基づいて三類型に整理した．具体的には，正解と予測について，いずれの予測とも対応が成立しない正解を**見落とし**，いずれの正解とも対応が成立しない予測を**誤抽出**と定義した．また，両者が部分的に重なるものの完全一致しない場合は，抽出境界の過不

表 3　LLM 誤り類型と代表例（下線：正解スパン，【】：モデル予測スパン）．

誤り類型（割合）	代表例	説明
見落とし（29.3%）	参加者 Dr【鈴木】、Ns【山田】、佐藤PT、【吉田】	人名が連続する箇所で見落としが起こる．
誤抽出（55.6%）	【鈴木】病院，【山田】神経内科，【佐藤】クリニック	病院など施設名およびそれに準ずる名称の一部を人名として誤って抽出する．
スパン誤り（15.1%）	【耳鼻科鈴木】，【山田初診】，胸部 CT【後佐藤】Dr	漢字が連続する表現において，人名スパン境界の誤りが生じる．

足や分割の違いに起因する**スパン誤り**として分類した．分類の結果，誤り総数に占める内訳は，見落とし 29.3%，誤抽出 55.6%，スパン誤り 15.1%であった．見落としは，人名が連続して列挙される箇所で発生しやすく，非識別化において致命的な誤りとなり得るため，より見落としの少ないモデルを構築する必要がある．誤抽出については，病院名のような施設名によく含まれる，人名に見られる文字列を，人名として過剰に抽出するケースが多かった．実験において，BERT 人名抽出モデルに比べ，LLM の Precision が大きく低下していた主な要因は，このような誤抽出であった．そして，スパン誤りは，複数の漢字が連続する表現において，境界が前後にずれることで生じやすかった．

未知データにおける誤りの傾向　未知データに対しては，BERT・LLM ともにすべての指標で性能が低下している．BERT については，既知データと比較して未知データにおいて，フルネームの人名を途中で区切ってしまうスパン誤りと，同じ文字列で一般名詞の意味を持つ苗字の見落としが増加していた．また，LLM については，山田（太）のような人名の略記に対する見落としの増加等が原因で Recall が低下し，IC（Informed Consent）といった医療現場で用いられる用語・略語の誤抽出の増加等が原因で Precision が低下したが，前節で類型化した代表的な誤りの傾向と一致していた．

非識別化モデルの実応用に向けた考察　本研究で行った評価データの既知・未知・混在への分割と評価は，ある施設で作成したアノテーション済み学習データで微調整したモデルを，別施設や別期間の診療テキストへ適用する状況を想定し，人名表記のカバレッジの違いを擬似的に再現するものである．表2に示す通り，既知・混在データでは BERT が高精度であり，処理対象となるデータセットに対して十分な学習データを準備し，人名表記のカバレッジを高められる場合には，従来の微調整に基づく手法が依然として有力であることが確認された．一方で，未知データでは LLM の性能が相対的に良好であり，学習データで人名表記を十分に網羅できない状況では，LLM の使用が有効となり得る．ただし，表3より LLM は誤抽出が主要な誤りであり，施設名など人名に類似した文字列を過剰抽出しやすい傾向が観察されている．また，個人情報保護の観点にのみ着目すれば，非識別化では見落としが最も重大なリスクとなる一方，誤抽出は相対的に影響が小さい．したがって，例えば LLM を単独で用いるのではなく，LLM の予測を人名等の候補生成として用い，人間による確認で誤抽出を抑制しつつ，漏れのないアノテーション作業の支援に活用できる可能性がある．

4　おわりに

本稿では，診療テキストにおける人名抽出を対象として，微調整した BERT と，zero-shot LLM 人名抽出モデルの性能を比較した．東北大学病院 38 診療科で作成された経過記録から構築した人名アノテーション付きデータを用いた評価の結果，評価データ全体では BERT が最も高い F1 を達成した一方で，学習データに含まれない人名のみからなる入力に対しては LLM が BERT を上回る性能を示した．この結果から，非識別化モデルの実応用において，処理対象の診療テキストに出現する人名を網羅できるだけの学習データが準備できる場合は BERT の方が優位である一方で，十分に学習データが準備できず入力されるデータに含まれる人名をカバーできない場合，LLM が有効に働く可能性が示唆された．

今後の課題として，本稿では人名のみを抽出対象としたが，年齢・住所・施設名など他の識別情報についても同様に調査を行う必要がある．今後，複数病院のデータを使用した評価や，初診時記録など他形式のテキストについても評価を行う予定である．

謝辞

本研究の実施にあたり，ご協力いただいた東北大学病院の先生方，オープン・ベッドラボのスタッフの皆様，臨床研究推進センターバイオデザイン部門のスタッフの皆様に厚く御礼申し上げます．

参考文献

[1] Anatol-Fiete Näher, Carina N Vorisek, Sophie AI Klopfenstein, Moritz Lehne, Sylvia Thun, Shada Alsalamah, Sameer Pujari, Dominik Heider, Wolfgang Ahrens, Iris Pigeot, et al. Secondary data for global health digitalisation. **The Lancet Digital Health**, Vol. 5, No. 2, pp. e93–e101, 2023.

[2] 武田理宏, 真鍋史朗, 松村泰志. 電子カルテデータ二次利用の現状と課題. 生体医工学, Vol. 55, No. 4, pp. 151–158, 2017.

[3] Ishna Neamatullah, Margaret M. Douglass, Li-wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. Automated de-identification of free-text medical records. **BMC Medical Informatics and Decision Making**, Vol. 8, p. 32, 2008.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **NAACL-HLT**, pp. 4171–4186, 2019.

[5] Christopher Meaney, Wali Hakimpour, Sumeet Kalia, and Rahim Moineddin. A comparative evaluation of transformer models for de-identification of clinical text data. **arXiv preprint arXiv:2204.07056**, 2022.

[6] Callandra Moore, Lucas Bulgarelli, Tom Pollard, and Alistair Johnson. Transformer-deid: Deidentification of free-text clinical notes with transformers. PhysioNet, 2023.

[7] Jiyong An, Jiyun Kim, Leonard Sunwoo, Hyunyoung Baek, Sooyoung Yoo, and Seunggeun Lee. De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained bert. **BMC Medical Informatics and Decision Making**, Vol. 25, No. 1, p. 82, 2025.

[8] Xi Yang, Tianchen Lyu, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. A study of deep learning methods for de-identification of clinical notes at cross institute settings. In **2019 IEEE International Conference on Healthcare Informatics (ICHI)**, pp. 1–3. IEEE, 2019.

[9] Xiang Yue and Shuang Zhou. PHICON: Improving generalization of clinical text de-identification models via data augmentation. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, **Proceedings of the 3rd Clinical Natural Language Processing Workshop**, pp. 209–214, Online, November 2020. Association for Computational Linguistics.

[10] Woojin Kim, Sungeun Hahm, and Jaejin Lee. Generalizing clinical de-identification models by privacy-safe data augmentation using gpt-4. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 21204–21218, 2024.

[11] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. **arXiv preprint arXiv:2303.11032**, 2023.

[12] Praphul Singh, Charlotte Dzialo, Jangwon Kim, Sumana Srivatsa, Irfan Bulu, Sri Gadde, and Krishnaram Kenthapadi. RedactOR: An LLM-powered framework for automatic clinical data de-identification. In Georg Rehm and Yunyao Li, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)**, pp. 510–530, Vienna, Austria, July 2025. Association for Computational Linguistics.

[13] Samuel Sousa, Michael Jantscher, Mark Kröll, and Roman Kern. Large language models for electronic health record de-identification in english and german. **Information**, Vol. 16, No. 2, p. 112, 2025.

[14] Bayan Altalla, Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. Evaluating gpt models for clinical note de-identification. **Scientific Reports**, Vol. 15, No. 1, p. 3852, 2025.

[15] Kiana Aghakasiri, Noopur Zambare, JoAnn Thai, Carrie Ye, Mayur Mehta, J Ross Mitchell, and Mohamed Abdalla. Not what the doctor ordered: Surveying LLM-based de-identification and quantifying clinical information loss. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 32187–32203, Suzhou, China, November 2025. Association for Computational Linguistics.

[16] Kohei Kajiyama, Hiromasa Horiguchi, Takashi Okumura, Mizuki Morita, and Yoshinobu Kano. De-identifying free text of japanese electronic health records. **Journal of Biomedical Semantics**, Vol. 11, No. 1, p. 11, 2020.

[17] 丹羽基二. 日本人の苗字 三〇万姓の調査から見えたこと. 光文社新書, No. 054. 光文社, 2002.

[18] Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. ModernBERT-Ja, 2025.

[19] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In **Proceedings of the 18th International Conference on Machine Learning**, pp. 282–289, 2001.

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.