固有表現抽出におけるタスク特化型 BERT と大規模言語モデルの性能比較と実用性評価

黒澤 研二 [1]　市川 聖 [2]　原口 昌也 [2]　狭間 美祐希 [2]
桜井 駿 [3]
[1] 株式会社リクルート
[2] 株式会社キュリオスビークル
[3] ワークスアイディ株式会社
kenji_kurosawa@r.recruit.co.jp
{ichikawa,haraguchi,hazama}@curicle.jp
shu_sakurai@worksid.co.jp

概要

本研究では，固有表現抽出におけるタスク特化型 BERT モデルと大規模言語モデル（Large Language Models，LLM）の一つである Gemini との性能と実用性を比較評価した．独自に構築したデータセットを用いて，few-shot 学習と fine-tuning の両方の学習方法での性能を分析し，タスク適合性やコストなどを考慮に入れて実用面での評価も行った．結果として，タスク特化型 BERT の fine-tuning において最も高い性能を示し，Gemini は few-shot 学習での柔軟性と，fine-tuning による性能向上の可能性を示した．本研究の知見は，実用的な固有表現抽出システムの設計と実装に貢献すると期待される．

1　はじめに

固有表現抽出（Named Entity Recognition, NER）は，自然言語処理（NLP）の中核的なタスクの一つであり，人名・地名・組織名・日付・金額などの特定の情報をテキストから抽出することを目的としている．この技術は，情報検索・文書分類・質問応答など多岐にわたる分野で重要な役割を果たしており，多くのアプリケーションで実適用されている．

近年，BERT（Bidirectional Encoder Representations from Transformers）[1] や LLM など Transformer[2] ベースのモデルが自然言語処理に革命をもたらした．BERT は双方向の文脈理解能力を備え，多くの NLP タスクで従来技術を凌駕する性能を示した．また，タスク特化型 BERT が開発され，各専門分野における固有表現抽出精度が向上した．

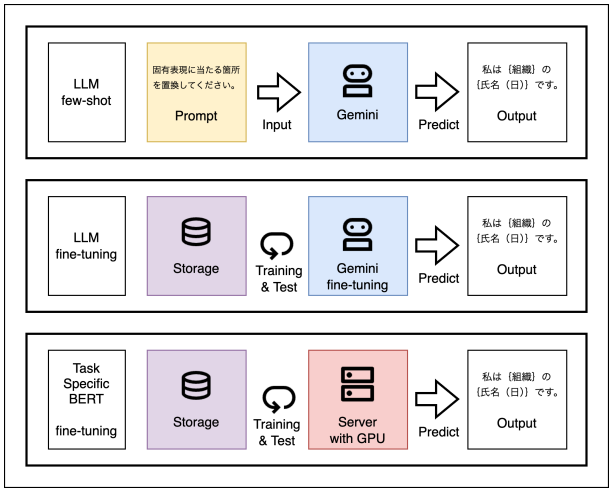

図 1　比較パターン

一方で，LLM は膨大なデータセットで事前学習されており，高い汎用性と柔軟性を持つことから，多様なタスクへの迅速な適用が可能である．しかし，LLM は自己回帰により順次トークンを生成する Transformer の Decoder[2] がベースとなっているため，余分な表現や形式不整合といった課題が存在する．

本研究では，図 1 に示す様に，Gemini[3]（Gemini-1.5-Pro-002）を few-shot で学習する，Gemini を fine-tuning する，BERT を fine-tuning するの 3 通りで固有表現抽出における Gemini とタスク特化型 BERT の性能および実用性を比較・評価する．具体的には，独自にニュースやブログから抽出した文章に対し氏名や住所などの固有表現に当たる箇所をアノテーションしたデータセットを用いて，モデル作成及び推論を行う．テストデータの F 値を算出することで

モデル性能を比較する．また，タスク適合性やアノテーションコスト等を算出することでモデルの実用性も比較する．

2 関連研究

近年は，大規模データの活用と最新の言語モデルの応用に焦点を当てている研究が増えている．

福島ら [4] は，50 億文規模の巨大なウェブコーパスから「カテゴリ名-固有名」というペアの集合である固有名リストを抽出し，これを固有表現抽出タスクに利用する手法を提案した．この研究は，大規模データを活用して固有表現抽出の性能を向上させる可能性を示した点で画期的であった．

一方，楢木ら [5] は，LLM を用いた自動アノテーション手法を提案し，人手によるアノテーションと統合することで，NER タスクの性能向上とコスト効率化を同時に達成した．この研究は，LLM の活用がアノテーションプロセスを大幅に改善できることを示唆している．

さらに，西山ら [6] は医療テキストの固有表現抽出において，生成モデルと既存の分類モデルを比較した．生成モデルが少数のデータでも高い性能を示し，特に精度の高い生成モデルがアノテーション支援に有用であることが明らかになった．

これらの先行研究は，大規模データや LLM の活用が固有表現抽出タスクの性能向上に寄与することを示している．しかし，タスク特化型 BERT と最新 LLM を直接比較し，その実用性を評価した研究はまだ少ない．そこで本研究では，これらのモデルの性能比較と実用性評価を行うことで，固有表現抽出タスクにおける最適なモデル選択の指針を提供することを目指す．

3 手法

3.1 モデル構成

本実験では，スクラッチでモデルを作成したタスク特化型 BERT と Gemini の 2 種類のモデルを使用した．

BERT は Transformer の Encoder[1] 部分を使用しており，Self-Attention[2] により長距離の依存関係を捉えることができる．これは双方向型で，LLM で採用されている自己回帰型の GPT モデル [7] と比べて固有表現の文脈依存関係を捉えることができる．そのため，タスク特化型モデルとして BERT を採用し，特にその中でも高精度な東北大学が公開している日本語学習済み BERT モデル [8] を使用することとした．

さらに，ニュース記事やブログなどの文章データを収集し人手によりアノテーションを行い，独自に日本語の固有表現抽出データセットを作成した．本データセットを用いて BERT モデルの fine-tuning を行った．

一方，Gemini については，Gemini-1.5-Pro-002[3] を選定した．本モデルは 2024 年 12 月時点で最新の LLM で，大規模かつ多様なデータセットで学習されていると推察されており，幅広い知識と柔軟な言語理解能力を持つ．また，クラウド環境上で few-shot や fine-tuning を容易に実施可能である．few-shot 学習による迅速なタスク適応が可能であり，プロンプトエンジニアリングを通じて固有表現抽出タスクに対応させることができる．

表 1 データセットの内訳

データの種類	サンプル数
教師データ	10,067
検証データ	1,000
テストデータ	320

3.2 実験設計

本研究では，Gemini を few-shot で学習する，Gemini を fine-tuning する，BERT を fine-tuning するの 3 通りの学習方法でモデルの性能を評価した．データセットは，ニュース記事，ブログ記事などから収集し，

- 定型：電話番号，郵便番号，口座番号，メールアドレス，URL
- 非定型：氏名（日），氏名（英），住所，日付，時間，ID，パスワード，組織
- 専門：法律，医療用語，緯度経度

の合計 16 種類の固有表現にあたる語句に対し人手でアノテーションを行った．few-shot で学習させるデータ件数は 64 件である．fine-tuning に使用した各データセット件数は，表 1 に示す．

また，評価指標として，Precision，Recall，F1-score を採用し，各固有表現タイプによりモデル間の特性の違いを検討した．

3.3 実装詳細

タスク特化型 BERT および Gemini の fine-tuning では，事前学習済みモデルに対して独自に作成し

表 2　モデル別の固有表現抽出結果

固有表現	教師データ内の固有表現数	Gemini-1.5-Pro-002 (zero-shot)			Gemini-1.5-Pro-002 (few-shot)			Gemini-1.5-Pro-002 (fine-tuning)			タスク特化型 BERT		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
電話番号	417	0.950	0.950	0.950	0.950	0.950	0.950	0.900	0.900	0.900	0.889	0.800	0.842
郵便番号	390	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.952	0.952	0.952
口座番号	445	0.722	0.520	0.605	0.762	0.640	0.696	0.750	0.783	0.766	0.818	0.720	0.766
メールアドレス	422	1.000	1.000	1.000	0.950	0.950	0.950	1.000	1.000	1.000	1.000	1.000	1.000
URL	445	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
氏名（日）	3997	0.588	0.769	0.667	0.846	0.846	0.846	0.667	0.769	0.714	0.923	0.923	0.923
氏名（英）	377	0.950	0.950	0.950	0.950	0.950	0.950	0.813	0.650	0.722	0.842	0.800	0.821
住所	2738	1.000	0.708	0.829	1.000	0.833	0.909	0.714	0.833	0.769	1.000	0.833	0.909
日付	2034	1.000	0.607	0.756	0.950	0.679	0.792	0.958	0.821	0.885	1.000	0.786	0.880
時間	655	0.900	0.643	0.750	0.870	0.714	0.784	0.920	0.821	0.868	0.926	0.893	0.909
ID	631	0.538	0.350	0.424	0.688	0.550	0.611	0.889	0.800	0.842	0.833	0.750	0.789
パスワード	384	0.909	1.000	0.952	0.909	1.000	0.952	0.870	1.000	0.930	1.000	1.000	1.000
組織	2289	0.500	0.296	0.372	0.640	0.593	0.615	0.905	0.704	0.792	0.955	0.778	0.857
法律	513	0.667	0.261	0.375	0.808	0.913	0.857	0.833	0.870	0.851	0.913	1.000	0.955
医療用語	742	0.000	0.000	0.000	0.500	0.133	0.211	0.778	0.700	0.737	1.000	0.633	0.776
緯度経度	1092	0.556	0.345	0.426	0.696	0.552	0.615	1.000	0.966	0.982	0.741	0.690	0.714
平均		0.768	0.650	0.691	0.845	0.769	0.796	0.875	0.851	0.860	0.925	0.847	0.880

表 3　モデル別の推論速度結果

	Min(s)	25%tile(s)	Mean(s)	75%tile(s)	Max(s)	Var	SD
Gemini-1.5-Pro-002 (zero-shot)	5.785	6.237	7.418	7.710	34.671	7.492	2.737
Gemini-1.5-Pro-002 (few-shot)	6.101	6.772	6.880	6.966	8.656	0.036	0.191
Gemini-1.5-Pro-002 (fine-tuning)	6.908	7.882	8.694	9.125	16.609	1.616	1.271
タスク特化型 BERT (CPU)	0.220	0.264	0.299	0.310	0.699	0.004	0.062

た固有表現抽出タスク用のデータセットを用いて fine-tuning を行った．

Gemini の few-shot では，タスク説明と数例の入出力サンプルを含むプロンプトテンプレートを作成した．プロンプトには，文章から先述の固有表現を抽出して指定文字列に置換させるタスク指示と，正しく置換された例を含めた．

4　実験結果

4.1　固有表現抽出の精度比較

表 2 に，Gemini の few-shot 学習および fine-tuning，タスク特化型 BERT の性能比較を示す．参考までに，Gemini の zero-shot の結果も併記した．Gemini は few-shot 学習でも一定の精度を示し，fine-tuning によってさらに精度が向上している．また，タスク特化型 BERT が最も高い精度となっている．

4.2　固有表現タイプ別の性能分析

各モデルの固有表現タイプ別の F1-score を分析したところ，以下のような特徴が観察された．

- zero-shot の推論結果を考慮すると，「郵便番号」，「メールアドレス」，「URL」などの定型的な固有表現タイプの識別に関して，Gemini は元々高い性能を持っている．
- Gemini は，few-shot 学習で「郵便番号」，「URL」，「組織」，「法律」，「医療用語」，「緯度経度」などの定型的または専門的な固有表現タイプで高い性能を示した．
- Gemini は，fine-tuning を行うことで多くの固有表現タイプで性能が向上し，特に「緯度経度」や「医療用語」などの専門的な固有表現タイプでの改善が顕著であった．
- タスク特化型 BERT は，特に「氏名（日）」，「住所」，「組織」，「法律」，「医療用語」などの非定型・専門的な固有表現タイプで高い性能を示した．

4.3　推論速度比較

表 3 に，Gemini の few-shot 学習および fine-tuning，タスク特化型 BERT の平均推論速度の比較を示す．これは，320 回推論を行い，各回毎に結果が出るまでにかかった時間の平均である．なお，推論時の平均入力トークン数は，Gemini の zero-shot が約 366 トークン，few-shot 学習が約 2789 トークン，その他は共通で約 18 トークンである．

タスク特化型 BERT は平均 0.299 秒で推論が完了し，Gemini に比べて高速であった．標準偏差も小さく安定した応答速度を示している．

5 考察

5.1 各モデルの実用性評価

本研究の結果から，Gemini とタスク特化型 BERT の実用性を以下の観点から評価した．

- 精度：

  Gemini は，few-shot 学習でも一定の性能を示し，特に定型的または専門的な固有表現で高い精度を達成した．fine-tuning を行うことで多くの固有表現タイプで性能が向上し，タスク特化型 BERT に近い精度となる．

  タスク特化型 BERT は，高い精度と多様な固有表現タイプに対応可能であることを示した．特に，非定型的な固有表現タイプや日本語特有の表現において優れた性能を発揮した．

  精度面では，両者ともに高い精度を達成できるが、データに定型的固有表現が多い場合は Gemini を，非定型的固有表現が多い場合はタスク特化型 BERT を選択することが適している．
- 速度：

  推論速度の観点では，タスク特化型 BERT は Gemini よりも高速であり，一定の応答速度を保つ安定性を示した．タスク特化型 BERT の方が，大規模なデータ処理やリアルタイム処理が求められるアプリケーション，リソース制約のある環境において有利である．

5.2 モデル選択の指針

本研究の結果から，以下のようなモデル選択指針が提案できる．

- データ量および精度要求：

  少ないデータしかない環境下では，Gemini の few-shot 学習が有効である．タスク特化型 BERT は多様な固有表現タイプに対して平均的に高い精度を示していることから，大規模なアノテーション済みデータセットが利用可能である場合は，タスク特化型 BERT が最適である．

  また、精度の観点からみると，Gemini も fine-tuning によってタスク特化型 BERT に比肩する高い精度を達成できるため，非定型的固有表現タイプが少なければ Gemini も選択可能である．しかし，高い精度と多様な固有表現タイプへの対応が求められるアプリケーションでは，タスク特化型 BERT がより適している．
- リアルタイム性：

  タスク特化型 BERT と比べて Gemini は推論速度が遅く，応答速度にも若干ばらつきがある．これは，NW 速度やクラウド上のリソース利用状況などの影響によると考えられる．そのため，Gemini は，バッチ処理などリアルタイム性を重視しないシナリオや遅延が許されるシステムでの利用が望ましい．

  リアルタイム処理や低レイテンシ要件がある場合は，高速で安定な推論性能を持つタスク特化型 BERT が適している．
- 開発コスト：

  Gemini は，few-shot 学習による迅速なシステム構築が可能である．大量のアノテーションデータを必要とせずに一定の性能を発揮できる点で初期開発コストが低い．開発コストを安価に抑えたい場合には Gemini が適している．

  一方，タスク特化型 BERT は，fine-tuning 時に高品質なアノテーションデータを大量に必要とするため，データ作成などの初期開発コストが高くなる懸念がある．
- 運用環境：

  タスク特化型 BERT は，オンプレミス環境での運用が可能であり，セキュリティやデータプライバシー保護の観点から有利である．

  一方，Gemini も堅牢なセキュリティ機能を備えているため，クラウド環境で運用可能であれば Gemini も検討できる. ただし，API 利用料やネットワーク遅延など運用コストへの配慮が必要である．

6 結論

本研究では，タスク特化型 BERT と LLM（Gemini）の性能と実用性を比較評価した．タスク特化型 BERT は高精度かつ高速な推論性能を示し，一貫した結果を提供するため，本番環境向けアプリケーションに適している．

一方，Gemini は少量データ環境下で迅速に適応できる．また，自動アノテーション機能による初期開発コスト削減にも寄与する．

これらの知見は，固有表現抽出タスクにおけるモデル選択や設計戦略に関する重要な指針となる．

謝辞

本論文を執筆するに当たり，研究計画や原稿内容に関する推敲やアドバイスを頂いた株式会社リクルートの高橋 諒氏，中田 百科氏に感謝致します．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, pp. 5998–6008, 2017.

[3] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens. **arXiv preprint arXiv:2403.05530**, 2024.

[4] 福島健一, 鍛治伸裕, 喜連川優. 日本語固有表現抽出における超大規模ウェブテキストの利用. 電子情報通信学会第 19 回データ工学ワークショップ（DEWS2008）, 2008.

[5] 楢木悠士, 山木良輔, 池田愛和, 堀江孝文, 長沼大樹. 固有表現抽出における大規模言語モデルを用いた自動アノテーション. 言語処理学会第 30 回年次大会発表論文集, 2024.

[6] 西山智弘, 柴田大作, 宇野裕, 辻川剛範, 北出祐, 久保雅洋, 矢田竣太郎, 若宮翔子, 荒牧英治. 生成モデルは医療テキストの固有表現抽出に使えるか？ 言語処理学会第 30 回年次大会発表論文集, 2024.

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. **OpenAI**, 2018.

[8] Tohoku NLP Group. Bert base japanese (unidic-lite with whole word masking, cc-100 and jawiki-20230102). https://huggingface.co/tohoku-nlp/bert-base-japanese-v3, 2023. Accessed: [2024-12-02].