固有表現抽出タスクの形式の学習と様々なドメインへの適用

石井奏人 [1]　新妻巧朗 [1]　田森秀明 [1]

[1] 株式会社朝日新聞社メディア研究開発センター
{ishii-k10,niitsuma-t,tamori-h}@asahi.com

概要

新聞社では，分析の目的や対象文書のドメインに応じた固有表現抽出（NER）が求められる．従来どおり，特定のドメインごとに NER モデルをファインチューニングするには，教師データの作成などに高いコストがかかる．一方，大規模言語モデル（LLM）を用いた Few-shot 推論なども検討できるが，精度は十分とはいえない．そこで本研究では，LLM に対する NER プロンプトのフォーマットと，NER タスクそのものの形式を学習させるファインチューニング手法を提案する．このモデルで Few-shot 推論を行い，少量のデータのみで特定のドメイン・ラベルへ適用させた NER を行う．

1　はじめに

MUC-6 [1] で定義されるような，人名や地名など一般的なカテゴリを対象とした従来の NER タスクに対しては，すでに高い精度を達成するモデルが存在する．しかし，一般のドメインだけでなく，特定のドメインやラベルを対象とする NER のニーズも多い [2]．例えば新聞社では，「首相動静 [1)]」のような特定のコンテンツにおいて，政治家名のみを抽出するといったニーズがある．こうした要求を満たすためには，従来のように大量の教師データを用いた学習が必要となるが，データの作成コストが大きいという課題がある．

近年，GPT-4 [3] をはじめとする LLM の有用性が自然言語処理のさまざまなタスクで示されているが，データが豊富にある領域の NER においては従来モデルに精度で劣る場合がある [4]．大規模な訓練データを用意できない環境で特定のドメインやラベル設定が必要な状況では，LLM による Few-shot 推論は有用となり得るが，Few-shot 推論のみでは十分な精度が得られないケースも存在する．

1) https://www.asahi.com/rensai/list.html?id=256

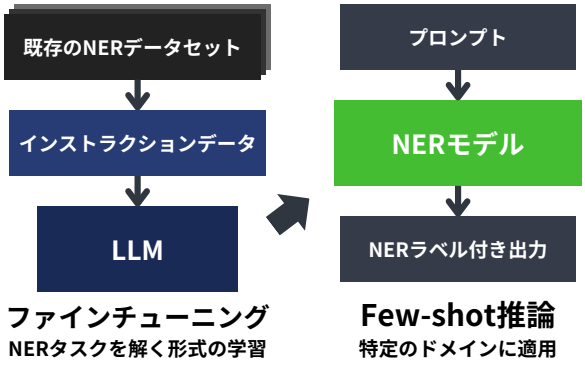

図 1　提案法の概要．既存の NER データセットをもとにインストラクションデータを作成して LLM をファインチューニングし，タスク形式を学習させる．その後，少量のデータを Few-shot で与えることで，特定のドメイン・ラベルへの適用を図る．

そこで本研究では，特定のドメインやラベルそのものを事前に学習するのではなく，LLM に「NER タスクの形式」を学習させることを目的とする．具体的には，既存の複数の NER データセットを組み合わせてインストラクションデータを作成し，それを用いて LLM をファインチューニングする．このようにして NER タスク全般を解く知識・形式を事前に学習したモデルを用いると，新たなドメインやラベルに対し Few-shot のみで推論ができる．図 1 に提案法の概要を示す．

実験では，既存の複数の NER データセットを組み合わせてインストラクションデータを作成し，LLM をファインチューニングした．その後，インストラクションデータには含まれていない 2 種類のデータセットに対して Few-shot 推論を行い，NER タスクの形式の学習による効果を検証した．

2　関連研究

近年，GPT-4 [3] をはじめとする LLM が様々なタスクで高い性能を示しており，特に Few-shot による推論など，大規模な教師データを利用できない領域

において注目されている [5]．一方，NER においては，BERT [6] などを用いたエンコーダーモデルによる推論の優位性は依然として高い一方で，特定のドメインやラベル向けのモデル学習には大量のデータ作成にかかるコストが課題となる．少量のデータのみで特定のドメイン・ラベルに対応するため，LLM による NER の研究が行われており，GPT-NER [4] や，PromptNER [7] などが提案されている．一方で，これらは，系列ラベリングをベースにした NER とは異なり，エンティティの抽出のみを行うため，スパンを取得できない．スパンを取得してテキスト内の複数の同名のエンティティを識別できれば，正確な関係抽出，エンティティリンキングのために有用となる．

LLM を用いた教師データの生成 [8] が検討されている．アノテーション対象のラベルを指定せずオープンに生成させることで，多様なエンティティを持つデータを作成する手法 [9, 10] も提案されている．ただ，ドメイン知識を必要とする場合など，LLM のみで十分な精度の教師データを作成することは難しい [11]．そこで LLM を用いた事前アノテーションによるコスト削減 [12] も検討されている．

NER タスクに関連するデータセットとしては，CoNLL-2003 [13]，拡張固有表現 [14, 15] に基づいた『現代日本語書き言葉均衡コーパス』（BCCWJ [16]）に対する拡張固有表現タグ付きコーパス [17] など，多くのデータセットが存在する．これらのデータを NER のためのインストラクションデータ [18] に利用することで，LLM に NER の知識や形式を与えたモデルを学習できる．本研究では，ファインチューニングしたモデルと Few-shot 推論を組み合わせることで，データが少ない条件下で特定のドメイン・ラベルに対して適用させた NER を行う．

3 提案法

LLM を用いた NER のためのプロンプトフォーマットと，インストラクションデータを用いた NER タスクの形式の学習，Few-shot 推論による特定のドメイン・ラベルへの適用を提案する．

3.1 NER のためのプロンプトフォーマット

プロンプトは，学習時，推論時に同様のフォーマットを用いる．ただし，Few-shot 事例は推論時のプロンプトのみに含め，インストラクションデータのプロンプトには含めない．以下に，プロンプトの構成を示す．

instruction で入力にエンティティのタグをつけるよう指示をする．また，ここで推論の対象となるエンティティのラベルを定義する．出力形式には，エンティティの対象文字列とラベルだけでなく，スパンを正しく抽出できるように，入力のテキストに XML 形式のタグを用いてエンティティの情報を与える方法を採用する．出力を output のタグで囲むこと，入力テキストに対する書き換え，削除，挿入などは行わず，入力に対するタグの挿入のみを行うことなど，出力形式に関する注意もここに記載する．

define the following tags に対象となるエンティティの定義を記載する．

example（推論時のみ使用）に，入力文と，出力の例として入力文にエンティティの予測結果をタグで示したものをそれぞれ<input></input>, <output></output>で囲み，複数個示す．

result には入力文を<input></input>のタグで囲んで示し，モデルに<output></output>で囲んだタグ付きのテキストを出力させる．

以下に実際のプロンプトフォーマットを示す．

プロンプトフォーマット

```
# instruction
Tag the entities contained in the input. Entities are [エンティティ名], ···. The output is enclosed in an output tag. No rewriting, deletion, or insertion is performed, only the insertion of tags for the input. If the input does not contain any entities, output as is.
# define the following tags
[エンティティ名]: [エンティティの定義]
# example
<input>[入力例]</input>
<output>[出力例]</output>
# result
<input>[入力テキスト]</input>
```

3.2 インストラクションデータを用いた NER タスクの形式の学習

3.1 のプロンプトフォーマットを既存の複数の NER データセットに適用し，LLM に対してファインチューニングをすることで，NER タスクの知識や形式を学習させる．既存の NER データセットは，BIO タグを用いた系列ラベリングを前提としているため，エンティティを XML 形式のタグで囲うように整形する．NER タスク全般の知識や形式の学習

表 1 インストラクションデータに用いるデータセット

	データ数	ドメイン	ラベル数
BCCWJ	1585	一般	23
AnEM	1819	解剖学	11
SEC-filings	659	金融	4
re3d	508	防衛	10

表 2 評価に用いるデータセット

	データ数	ドメイン	ラベル数
Stockmark	535	一般	8
MedTxt-CR	140	医療	13

と，未知のドメイン・ラベルへの適用が目的であるため，推論時に使用するデータセットは学習にはあえて含めない．

3.3 Few-shot 推論による特定のドメイン・ラベルへの適用

ファインチューニングによって得られた NER の形式を学習したモデルを用いて，ターゲットのドメインやラベルに適用するための Few-shot 推論を行う．あらためて推論時のプロンプトに，対象となるエンティティを定義とともに記載することで，インストラクションデータに含まれないエンティティに対する NER ができる．

タグ付きの出力から XML をパースし，定義したラベルと一致しないタグを削除するなどの処理をして対象のラベルとスパンを抽出する．元のテキストの書き換えが起こった場合は，入力文と，タグを除いた出力文の編集距離を計算し，入力文と同じになるよう出力文を修正する．

4 実験

NER の形式を学習したモデルの性能を測定するため，LLM のファインチューニング前後における NER の精度を比較する．

4.1 ファインチューニング

学習に用いるモデルは，OpenAI[2] の GPT-4o mini (`gpt-4o-mini-2024-07-18`)，Llama 3.1 Swallow 8B [19, 20] とする．

インストラクションデータには，国立国語研究所『現代日本語書き言葉均衡コーパス』(BCCWJ) に対する拡張固有表現タグ付きコーパス [16, 17], Relationship and Entity Extraction Evaluation Dataset (re3d) [21]，Anatomical Entity Mention (AnEM) [22], Security and Exchange Commission (SEC) filings [23] の 4 種類のデータセットを用いる．インストラクションデータに含まれるデータ数とドメイン・ラベル数を表 1 に示す．

2) https://platform.openai.com/

表 3 GPT-4o mini のファインチューニングにおけるパラメータと学習ステップ数

Epochs	Batch size	LR multiplier	Step
1	4	0.1	1143

表 4 Swallow 8B のファインチューニングにおけるパラメータと学習ステップ数

Epochs	Batch size	Learning rate	Step
3	1	1e-6	1716

4.2 Few-shot 推論の評価結果

評価用データとして，Stockmark 社の Wikipedia を用いた日本語の固有表現抽出データセット [24]，MedTxt-CR: 症例報告 (Case Reports) コーパス [25] の 2 種類を用いる．評価用データセットのデータ数とドメイン・ラベル数を表 2 に示す．本実験の設定では，NER の形式の学習時には，後に適用させるドメインとラベルが決まっていないことを想定するため，評価に用いるデータセットはインストラクションデータには含まれないことに注意する．

表 3，表 4 に，ファインチューニング時のパラメータと，学習ステップ数を示す．

表 5 に，Stockmark データに対して 8-shot の条件で推論した結果を示す．GPT-4o mini においては，ファインチューニングにより 8 ポイントの精度向上が見られた．また，ファインチューニング前のスコアが低かった Swallow においては，24 ポイントの精度向上が見られた．

表 6 に，MedTxt-CR に対して 2-shot の条件で推論した結果を示す．MedTxt-CR においては，GPT-4o mini で 5 ポイント，Swallow で 2 ポイントの精度向上が見られた．

以上の結果から，ファインチューニングで NER タスクの形式を学習したことによる精度の向上が確認できた．

5 議論

Few-shot の事例の数による性能の変化と，推論時のデータセットのドメイン・ラベルによる精度の違いに関して議論する．

表 5 Stockmark データセットにおける 8-shot の性能比較（FT はファインチューニングしたモデルを表す）.

Model	Precision	Recall	F1-Score
GPT-4o mini	0.66	0.63	0.63
GPT-4o mini FT	0.72	0.72	**0.71**
Swallow	0.46	0.50	0.45
Swallow FT	0.73	0.71	0.69

表 6 MedTxt-CR データセットにおける 2-shot の性能比較.

Model	Precision	Recall	F1-Score
GPT-4o mini	0.51	0.37	0.42
GPT-4o mini FT	0.57	0.40	**0.47**
Swallow	0.34	0.20	0.22
Swallow FT	0.52	0.17	0.24

5.1 Stockmark データに対する結果

図 2 に，Stockmark データに対して Few-shot の事例数を変えながら各モデルで推論をした結果を示す．全てのモデルにおいて，Few-shot の事例数に応じた精度の向上が見られた．0-shot における精度の向上は，ファインチューニングによって，エンティティにあたる単語の特徴や，出力の形式を正しく学習できていることなどが要因として考えられる．

5.2 MedTxt-CR データに対する結果

図 3 に，MedTxt-CR データに対する結果を示す．MedTxt-CR データは，入力文字数が多く，事例を増やした際に正しく出力ができない例があったため，2-shot までの条件で実験した．GPT-4o mini では，ファインチューニングの有無に関わらず，事例数を増やすことで精度が向上する傾向が見られた．一方，Swallow においては，1-shot に比べて 2-shot の条件で精度が落ちる結果となった．

5.3 データセットによる性能向上の差

Stockmark データは対象のドメインが近く，一部のエンティティがインストラクションデータに含まれるものと重なるため，ファインチューニングによる精度向上が比較的大きかった．一方で，MedTxt-CR データはインストラクションデータには含まれない医療ドメインかつ，時間表現を除いたラベルが新規であるため，ファインチューニングによる精度向上の幅が比較的小さい．しかし，0-shot では 7 ポイント以上の上昇が見られ，提案法の効果が確認できた．

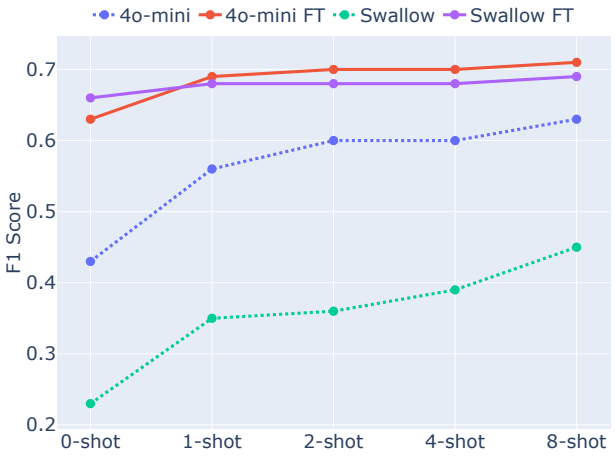

図 2 Stockmark データに対する推論の f1-score.

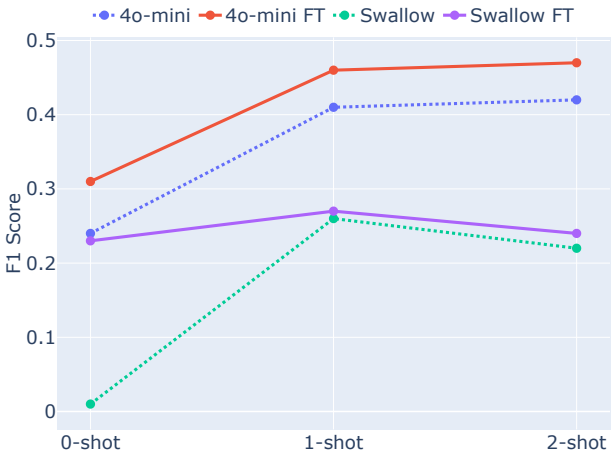

図 3 MedTxt-CR データに対する推論の f1-score.

6 おわりに

様々なドメインの NER を解くために，LLM に NER タスクの形式を学習させる手法を提案した．既存の複数の NER データセットを用いて作成したインストラクションデータで LLM のファインチューニングを行い，NER タスクを解くための知識と形式を学習させることで，少量のデータを用いた Few-shot 推論により特定のドメインに対する NER を行うことができる．実験では，インストラクションデータに含まれない特定のドメイン・ラベルに対する NER を検証し，NER のタスク形式を学習させたことによる精度の向上を確認した．

今後の展望として，NER のコンテキストを活用してマルチステップに対話をすることで，単一のモデルで行う属性抽出やエンティティリンキングの実現を検討している．

謝辞

研究アドバイザーとしてご助言をいただいたMBZUAI・東北大学の乾健太郎教授および東京科学大学の岡崎直観教授に感謝いたします.

参考文献

[1] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6: A brief history. In **COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics**, 1996.

[2] 福田美穂, 関根聡. 国内におけるドメイン依存の固有表現抽出の応用技術の調査. 自然言語処理, Vol. 30, No. 2, pp. 800–815, 2023.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[4] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models, 2023.

[5] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In **Proceedings of the 36th International Conference on Neural Information Processing Systems**, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, Vol. abs/1810.04805, , 2018.

[7] Dhananjay Ashok and Zachary C. Lipton. Promptner: Prompting for named entity recognition, 2023.

[8] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a good data annotator? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11173–11195, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. Universalner: Targeted distillation from large language models for open named entity recognition, 2024.

[10] Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne P Bernard. NuNER: Entity recognition encoder pre-training via LLM-annotated data. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 11829–11841, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] Yuxuan Lu, Bingsheng Yao, Shao Zhang, Yun Wang, Peng Zhang, Tun Lu, Toby Jia-Jun Li, and Dakuo Wang. Human still wins over llm: An empirical study of active learning on domain-specific annotation tasks, 2023.

[12] 石井奏人, 新妻巧朗, 田口雄哉, 山野陽祐, 杉野かおり, 田森秀明. 大規模言語モデルによる事前ラベリングを活用した系列ラベリングのアノテーション. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 4Xin2114–4Xin2114, 2024.

[13] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In **Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003**, pp. 142–147, 2003.

[14] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In Manuel González Rodríguez and Carmen Paz Suarez Araujo, editors, **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

[15] Satoshi Sekine. Extended named entity ontology with attribute information. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[16] 前川喜久雄（監修）. 書き言葉コーパス—設計と構築—. 講座日本語コーパス 2. 朝倉書店, 2014.

[17] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会研究報告 = IPSJ SIG technical reports, Vol. 2008, No. 113, pp. 113–120, 11 2008.

[18] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **International Conference on Learning Representations**, 2022.

[19] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[20] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[21] Defence Science and Technology Laboratory. Relationship and entity extraction evaluation dataset (documents), 2024. Accessed: 2024-12-17.

[22] Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. Open-domain anatomical entity mention detection. In Antal Van Den Bosch and Hagit Shatkay, editors, **Proceedings of the Workshop on Detecting Structure in Scholarly Discourse**, pp. 27–36, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[23] Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. Domain adaption of named entity recognition to support credit risk assessment. In Ben Hachey and Kellie Webster, editors, **Proceedings of the Australasian Language Technology Association Workshop 2015**, pp. 84–90, Parramatta, Australia, December 2015.

[24] 近江崇宏. Wikipedia を用いた日本語の固有表現抽出のデータセットの構築. 言語処理学会 第 27 回年次大会, 2021. PDF.

[25] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-mednlp: Overview of real document-based medical natural language processing task. In **Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)**, pp. 285–296, 2022.