固有表現候補の用語情報を取得する LLM を用いた固有表現抽出

与那覇竜馬　三輪誠
豊田工業大学
{sd21101,makoto-miwa}@toyota-ti.ac.jp

概要

生物医学分野における大規模言語モデルを用いた固有表現抽出を対象に，抽出モデルに合わせた用語情報の取得を行うために，一度，抽出した候補について用語情報を取得し，その用語情報を利用して再度抽出を行う 2 段階の抽出手法を提案する．また，取得した用語情報の種類及びその組み合わせが抽出性能に与える影響を調査する．統合医学用語システム UMLS を用語辞典として，BC5CDR データセットを対象に評価を行った結果，用語情報を利用することで，提案手法による抽出性能の向上を確認できた．特に用語の定義とカテゴリを，用語の定義がない場合に親概念の定義を利用しながら，組み合わせる方法が有効であることがわかった．

1　はじめに

大規模言語モデル (Large Language Models; LLM) は様々な自然言語処理タスクで高い性能を誇っている．しかし，生物医学分野の文書から事前に定義された用語を抽出する固有表現抽出 (Named Entity Recognition; NER) においては，ファインチューニングを行う手法においても，未だ従来手法の性能に追いついていない [1].

生物医学分野における LLM を用いた固有表現抽出の性能向上のために，統合医学用語システム (Unified Medical Language System; UMLS) [2] から用語の定義やカテゴリといった用語情報を取得し，LLM の入力に付与する研究が注目を集めている [1, 3, 4]. しかし，これらの既存手法では，抽出モデルの判断に必要な情報か否かを情報取得時に考慮できていない．さらに，用語の定義やカテゴリの情報を個別に利用しているため，有効な情報の種類やその組み合わせは明らかになっていない．

これらの課題を解決するために，本研究では，抽出モデルの予測に合わせた用語情報の利用の有効性の検証と有効な用語情報の組み合わせの発見を目的とする．抽出モデルの判断に必要な情報を取得するために，1 段階目で LLM を用いた固有表現抽出で予測した固有表現候補について用語情報の取得を行い，取得した用語情報を用いて，再度，固有表現抽出を行う 2 段階の抽出手法（図 1）を提案する．また，用語辞典から取得する情報の種類及びその組み合わせが抽出性能に与える影響を調査する．

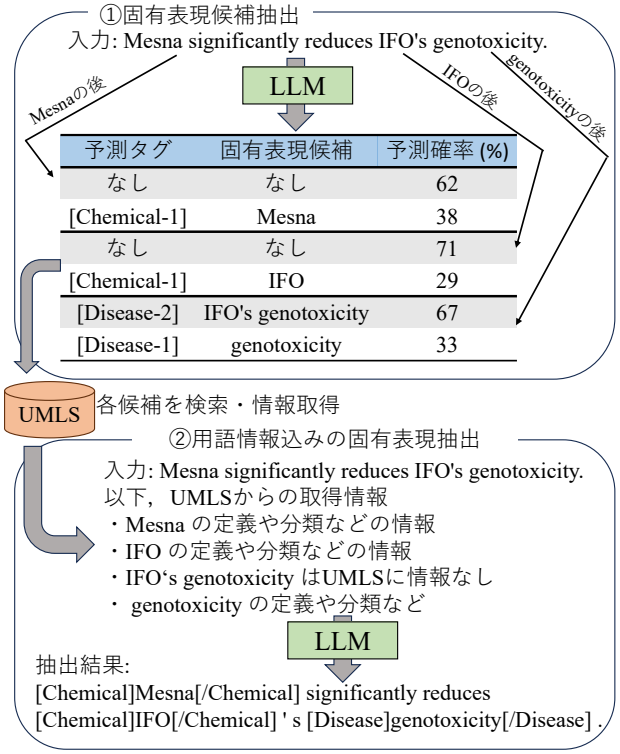

図 1　提案手法の概要図

本研究の貢献は以下のとおりである．

- 固有表現抽出で抽出した抽出候補について取得した用語情報を用いて，再度，抽出を行う 2 段階の LLM を用いた固有表現抽出手法を提案.
- LLM を用いた固有表現抽出において固有表現

Input: Recurrent reversible acute renal failure
from amphotericin.
Output: Recurrent reversible acute renal failure[Disease-3]
from amphotericin[Chemical-1] .

図 2　訓練データの例

候補を抽出するための新たな出力形式を提案.

- LLM を用いた固有表現抽出における用語情報として，用語の定義とカテゴリを，用語の定義がない場合に親概念の定義を利用しながら組み合わせる方法が特に有効であることを示した.

2　関連研究

言語モデルを用いた固有表現抽出は，言語モデルが学習時に取り込んだ静的な知識に依存しているため，生物医学分野など専門分野の専門用語を適切に抽出できない課題がある．この課題に対して，外部知識源から用語の情報を取得し，言語モデルの知識を補強する手法が提案されている [1, 3, 4, 5]

Kim ら [1] は，従来の言語モデルである ConNER [6] を用いた固有表現抽出結果について用語の定義を取得し，用語情報と ChatGPT を用いて結果を修正することで NER 性能を向上させるフレームワークを提案している．Biana ら [3] は入力文の用語を AutoPhrase [7] で取得し，用語のカテゴリと同義語を訓練データに加えて LLM を学習する手法を提案している．Munnangi ら [4] は入力文を ScispaCy [8] のエンティティリンキングモデルを使用して，文中の固有表現を抽出し，取得した固有表現について定義を取得して元に用語の定義を取得し，LLM の入力に付与する NER 手法を提案している．Amalvy ら [5] は，入力文の元となった論文の文脈情報を扱うために，元論文から入力文と文脈の近い文を取得し，LLM を用いた固有表現抽出で利用する手法を提案している．これらの研究は，抽出モデルの判断への情報の必要性を考慮した情報取得ができておらず，また，用語情報の組み合わせも考慮できていない.

3　提案手法

本研究では，LLM を用いた固有表現抽出の予測確率を用いて選んだ固有表現候補について用語情報を取得し，取得した用語情報を入力に付与して最終的な固有表現を抽出する 2 段階の固有表現抽出手法を提案する．ここでは 2 段階抽出モデルの予測，学習，また，用語情報の比較について説明する.

表 1　図 2 の訓練データに対する候補抽出結果の例

failure 後の予測結果	固有表現候補	予測確率 (%)
[Disease-2]	renal failure	53
[Disease-3]	acute renal failure	26
[Disease-1]	failure	21
Amphotericin 後の予測結果	固有表現候補	予測確率 (%)
タグなし	なし	57
[Chemical-1]	amphotericin	43

3.1　予測

予測においては，1 段階目の固有表現抽出では，用語情報を用いない抽出モデルを利用して，入力文中の固有表現候補とその予測確率を抽出する．LLM を用いた固有表現抽出では，入力文中の固有表現を “[type]”，“[/type]” のような固有表現の種類を表すタグで囲む出力形式を用いた抽出が高い性能を示すことが報告されている [9] が，このような抽出では固有表現全体の予測確率を計算できない．この問題に対処するため，図 2 に示すような固有表現の後に固有表現のタイプと範囲を表すタグを生成する新たな固有表現抽出の出力形式を提案する．このタグの予測確率を用いることで，表 1 のように，範囲のオーバーラップや抽出できていない偽陰性の例も含めて，固有表現候補を複数抽出し，情報取得を行う.

2 段階目の固有表現抽出では，1 段階目で抽出された固有表現候補について統合医学用語システムから用語情報を取得し，取得した用語情報を入力文に付与して，入力文中の固有表現を “[type]”，“[/type]” で囲む出力形式で最終的な固有表現を抽出する.

3.2　学習

学習においては，それぞれの抽出モデルを LoRA (Low-Rank Adaptation) [10] を用いて，学習データ上で個別にファインチューニングする．この際，2 段階目の用語情報を用いた固有表現抽出の学習のために，1 段階目の固有表現候補の抽出を学習データ上で行う必要がある．しかし，学習データ全体で学習したモデルを用いると，学習データと候補を抽出する対象のデータが同じになり，学習データ上での固有表現候補の確率分布が，予測データ上での確率分布と異なったものになる．この問題を緩和するために，学習データ上で 10 分割交差検証を行い，交差

検証の予測におけるタグの予測確率を固有表現候補の抽出に利用する．

3.3 利用する用語情報

統合医学用語システムに含まれる用語情報には，定義，カテゴリ，同義語・類義語，さらに用語間の階層関係が含まれている．本研究では，定義，カテゴリ，同義語・類義語を利用し，さらに，用語の定義がない場合に親概念の定義を利用する場合についても考慮しながら，用語情報の種類及び組み合わせについて，抽出性能を比較する．

4 実験

4.1 実験設定

LLM としては現在の最新モデルの一つである Llama-3.2-1B-Instruct [11] を使用した．また，実験及び評価では生物医学分野のタイトルと要旨に疾患と化学物質の固有表現がラベル付けされた BC5CDR [12] データセットを使用した．データセットの統計を付録 A の表 7 と 8 に示す．

実験では，候補抽出手法としてタグの種類と文脈情報として文外の元の文章の有無による比較，予測確率の閾値設定，用語情報の種類及び組み合わせについて段階的に開発データでの評価を行い，最良の条件でテストデータを用いて，最先端のモデルと比較した．1 段階目の候補抽出では固有表現の範囲について，2 段階目の固有表現抽出では固有表現の範囲とラベルタイプについて，適合率，再現率，F 値を指標として，評価した．まず，候補抽出手法を決定するために，2 種類のタグと文脈情報の有無における，学習データでの 10 分割交差検証における候補抽出性能を比較した (4.2 節)．次に，予測確率の閾値を決定するために，閾値を変えながら候補抽出モデルと 2 段階目の固有表現抽出モデルにおける抽出性能を比較した (4.3 節)．さらに，3.3 節で述べた用語情報について，その利用と組み合わせについて評価をした (4.4 節)．最後に，最良の条件を用いて SOTA モデルである VerifiNER [1] との性能比較を行った (4.5 節).

4.2 タグ形式の違いと文脈情報の有無における候補抽出性能の比較

候補抽出手法を決定するために，タグ形式の違い及び文脈情報の有無における候補抽出性能を比較した．ここで，予測確率 0.01 以上のものを候補として，10 分割交差検証で学習データから抽出した候補の範囲について評価を行った．結果として表 2 に示すように “[type-n]” 形式で，データに文脈情報を加える手法が候補抽出において優れていることがわかった．このため，このタグ形式を用いた候補について用語情報の取得を行うこととした．

表 2 タグ形式の違いと文脈情報の有無における学習データでの候補抽出の評価 (%)

条件	適合率	再現率	F 値
[n] (文脈情報なし)	63.78	84.54	72.68
[type-n] (文脈情報なし)	68.75	89.41	77.73
[type-n] (文脈情報あり)	76.78	91.75	83.60

表 3 閾値の違いによる開発データでの候補抽出の評価 (%)

閾値	適合率	再現率	F 値
0.1 以上	79.82	84.81	82.24
0.01 以上	68.60	85.02	75.93
0.001 以上	51.54	85.50	64.31
0.0001 以上	33.26	86.20	48.00
0.00001 以上	5.00	91.42	9.48

固有表現候補の抽出において，トークン数のみを表すタグ “[n]” より，ラベルタイプとトークン数を表すタグ “[type-n]” の方が優れていた理由として，タグに用語の種類の情報が含まれているためだと考えられる．さらに，文脈情報をデータに加えることで，LLM の文脈理解を助け，固有表現候補の抽出性能に寄与した考えられる．

4.3 候補抽出時の閾値による候補抽出性能及び固有表現抽出への影響

提案手法における予測確率の適切な閾値を決定するために，開発データを用いて，各閾値における候補抽出性能及び固有表現抽出性能を比較した．結果として，表 3 に示すように，閾値を 0.00001 としても抽出できない用語が約 9%存在した．

閾値を小さくすることで，固有表現の網羅性は向上するが，過剰な用語情報がノイズ・計算時間の両面において，2 段階目の抽出における妨げとなる懸念があるため，各閾値により抽出した候補を用いて 2 段階目の抽出性能を評価し，適切な閾値を決定した．2 段階目の抽出性能評価では，1 段階目の抽出候補について，用語情報として統合医学用語システムから用語の定義とカテゴリを，用語の定義がない場合に親概念の定義を取得し，利用した．表 4 より，予測確率の閾値を 0.001 以上としたとき，抽出

表 4　閾値の違いによる開発データでの固有表現抽出性能評価 (%)

閾値	適合率	再現率	F 値
0.1 以上	89.43	86.94	88.16
0.01 以上	89.62	86.83	88.20
0.001 以上	90.42	87.20	88.78
0.0001 以上	90.50	87.41	88.93
0.00001 以上	90.57	87.22	88.86

プロンプト

```
Your task is to identify diseases and chemicals in the input.
Please refer to the provided knowledge.
Knowledge: {information_from_UMLS}
```

入力文

```
 Input:  {input_sentence}
Answer: {model_output}
```

図 3　使用したプロンプト

性能が最も高く，過剰な用語情報を含まずに正しい固有表現を候補として抽出できる設定として適当であることを確認した．

4.4　用語情報追加による抽出性能評価

提案手法における各用語情報と組み合わせの評価のために，用語情報の種類及び組み合わせを変更し，開発データにおける抽出性能を比較した．計算時間の削減のため，文脈情報は付与せずに比較を行った．表 6 の通り，個別の情報としてはカテゴリが有効であり，組み合わせでは用語の定義とカテゴリ，用語の定義がない場合に親概念の定義を合わせて使用する方法が有効であった．結果として，一部の組み合わせを除き，抽出性能の向上が確認でき，提案手法における用語情報取得と利用の有効性を示した．

4.5　提案手法における抽出性能評価

4.4 節の結果より，定義とカテゴリ，定義がない場合に親概念の定義を組み合わせてたデータを用いてファインチューニングしたモデルで，テストデータでの抽出性能を評価した．

テストデータにおいても，用語情報を用いることでの性能向上が確認できた．一方で，最先端の BERT ベースのモデル VerifiNER [1] には性能では及ばなかった．このため，今後は候補の抽出方法やより良い用語情報の活用など，さらなる手法の改善が必要だと考える．

表 5　用語情報を用いた開発データでの固有表現抽出性能評価 (%)

情報の種類	適合率	再現率	F 値
なし	87.95	86.92	87.44
定義	88.77	87.12	87.93
カテゴリ	90.02	87.12	88.54
同義語	89.79	86.73	88.23
定義，カテゴリ	88.22	86.24	87.22
定義，同義語	89.03	85.56	87.26
定義，親概念	89.28	86.92	88.08
カテゴリ，同義語	88.91	85.70	87.28
定義，カテゴリ，同義語	91.39	83.90	87.49
定義，カテゴリ，親概念	90.50	87.41	88.93
定義，同義語，親概念	90.81	85.56	88.11
定義，カテゴリ，同義語，親概念	91.39	84.90	88.02

表 6　テストデータでの固有表現抽出性能評価 (%)

手法	適合率	再現率	F 値
用語情報なし	84.08	85.71	84.88
用語情報あり	86.10	85.88	85.99
VerifiNER [1]	94.77	91.61	93.16

5　おわりに

本研究では，LLM を用いた固有表現抽出において，予測に合わせた情報取得の有効性の検証，及び，有効な情報の種類やその組み合わせの発見を目的に，LLM の予測確率を用いて選んだ固有表現候補について用語情報を取得し，取得した用語情報を入力に付与して最終的な固有表現を抽出する 2 段階の抽出手法を提案した．また，各用語情報とその組み合わせについて固有表現抽出における有効性を評価した．実験では，提案した 2 段階の抽出手法により最終的な抽出性能が向上することを示し，特に，用語の定義とカテゴリを，用語の定義が見つからない場合に親概念の定義を利用しながら，組み合わせて使用したとき抽出性能に寄与することを示した．

今後の課題として，1 段階目の候補抽出において約 13%の固有表現が抽出できていないため，候補抽出における網羅性を高める工夫や，過剰な候補用語とその情報が含まれる懸念を解消するための選別の工夫，さらなる用語情報の活用が挙げられる．

参考文献

[1] Seoyeon Kim, Kwangwook Seo, Hyungjoo Chae, Jinyoung Yeo, and Dongha Lee. VerifiNER: Verification-augmented NER via knowledge-grounded reasoning with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2441–2461, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic Acids Research**, 2004.

[3] Junyi Biana, Weiqi Zhai, Xiaodi Huang, Jiaxuan Zheng, and Shanfeng Zhu. Vaner: Leveraging large language model for versatile and adaptive biomedical named entity recognition. In **27TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE**, 2024.

[4] Monica Munnangi, Sergey Feldman, Byron Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. On-the-fly definition augmentation of LLMs for biomedical NER. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 3833–3854, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[5] Arthur Amalvy, Vincent Labatut, and Richard Dufour. Learning to rank context for named entity recognition using a synthetic dataset. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10372–10382, Singapore, December 2023. Association for Computational Linguistics.

[6] Minbyul Jeong and Jaewoo Kang. Consistency enhancement of model prediction on document-level named entity recognition. **Bioinformatics**, Vol. 39, No. 6, p. btad361, 06 2023.

[7] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. Automated phrase mining from massive text corpora. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 30, No. 10, pp. 1825–1837, 2018.

[8] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, **Proceedings of the 18th BioNLP Workshop and Shared Task**, pp. 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[9] 鬼頭泰清, 牧野晃平, 三輪誠, 佐々木裕. 固有表現抽出における大規模言語モデルの lora ファインチューニングの学習設定の調査. 第 30 回言語処理学会年次大会, 2024.

[10] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, and et al. Angela Fan. The llama 3 herd of models, 2024.

[12] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. **Database**, Vol. 2016, p. baw068, 05 2016.

A　データセットの統計

本研究で使用したデータセット BC5CDR における学習，開発，テストのそれぞれのデータ数を表 7 に示す．また，固有表現ラベルごとの事例数を表 8 に示す．

表 7　BC5CDR のデータ数の統計

訓練	開発	テスト
5330	5330	5870

表 8　BC5CDR の固有表現ラベルごとの事例数

ラベル	訓練	開発	テスト
Disease	4,182	4,244	4,424
Chemical	5,203	5,347	5,385