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

鬼頭 泰清　牧野 晃平　三輪 誠　佐々木 裕
豊田工業大学
{sd20037,sd21505,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

大規模言語モデルを低コストでファインチューニングする LoRA と呼ばれるファインチューニング手法が注目を集めている．一方で，LLM の固有表現抽出 (NER) に対する性能は低く，未だ BERT を用いた最先端モデルの性能に追いついていない．本研究では，生命医学分野の NER に有効な LLM の LoRA を用いたファインチューニングの学習設定を調査した．結果として，最先端モデルの性能を上回る結果は得られなかったが，(1) プロンプトが不要であること，(2) タグ付けによる NER が高性能であること，(3) 学習可能な全ての層に LoRA を適用することで少ない学習パラメタ数でも Full Fine-tuning と同等の性能を達成することを明らかにした．

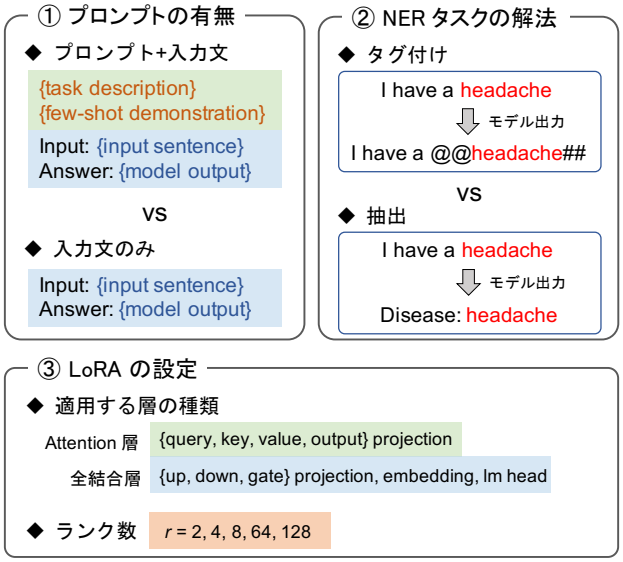

図 1　研究の調査内容

1　はじめに

固有表現抽出 (Named Entity Recognition; NER) は，文章中から人名や組織名などの特定の固有表現を抽出する基礎的な自然言語処理タスクで，文章の内容を捉える上で重要である．NER は古くから取り組まれ，その適用先は新聞記事などの一般的な文書のみならず，生命医学や化学などの専門分野の文書にも広がっている [1]．近年では，ニューラルネットワークを用いた手法が主流で，事前学習済み言語モデルの一つである BERT (Bidirectional Encoder Representations from Transformers) [2] を基盤とした教師あり学習による手法が高い性能を示している．

最近では，BERT よりも大規模なパラメタを持つモデルを大規模に事前学習した大規模言語モデル (Large Language Model; LLM) が自然言語処理分野で強力な存在感を見せており，質問応答や常識推論などのタスクで最高性能を発揮している [3]．NER においても LLM を用いた手法が研究されているが，現状では NER における LLM の性能は低いという問題があり [3, 4]，特に生命医学分野においては一般分野と比較して大幅に低い性能が報告されている [5]．性能を向上させるために LLM をファインチューニングする研究もされているが，未だ BERT を用いた最先端モデルの性能を下回っており [6, 7]，高い性能を発揮するための学習設定は明らかでない．

モデルの全パラメタを更新する通常のファインチューニング (Full Fine-tuning) を大規模なパラメタを持つ LLM に対して行うのは膨大なコストがかかるため，最近では少数のパラメタのみを更新してコストを低減する LoRA (Low-Rank Adaptation) [8] と呼ばれるファインチューニング手法が注目を集めている．LoRA はコストを抑えつつ，自然言語理解や自然言語生成などのタスクで Full Fine-tuning と同等の性能を発揮することが報告されている．

本研究では生命医学分野の NER を対象とし，LoRA による LLM のファインチューニングに有効な学習設定を明らかにするため，図 1 のように，プロンプト，NER タスクの解法，LoRA の設定の観点

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

から，それぞれが性能に及ぼす影響を調査する．

本研究の貢献は次の通りである．

- NER に対する LLM のファインチューニングにおいて，プロンプトが性能に与える影響は無視できることを確認した．
- LLM による NER タスクのタグ付けと抽出による解法を比較し，タグ付けを用いた方が高い性能を得られることを確認した．
- LoRA を適用する層の種類を多く設定することで，少ない学習パラメタ数で Full Fine-tuning と同等の性能を達成することを確認した．

2 関連研究

2.1 大規模言語モデルによる固有表現抽出

LLM は学習データ量やモデルのパラメタ数が膨大に増加したことで，プロンプトの文脈からタスクを学習する In-Context Learning (ICL) の能力を獲得し，プロンプトエンジニアリングと少数事例を用いた ICL に基づく Few-shot 学習により，幅広いタスクにファインチューニングなしに適応することが可能となってきた [9]．NER においても LLM を利用する研究が進んでおり，LLM の ICL に基づいた NER の研究 [4] や，LLM をファインチューニングする研究 [6, 7] が行われている．NER は系列ラベリング問題として取り組まれることが多い [10] が，LLM はテキスト生成を行うモデルであるため，系列ラベリングとは異なるタスク設定が必要である．テキスト生成による NER は様々な手法が提案されており，“I have a headache” という入力文に対して，“I have a @@headache##” のように固有表現にタグ付けした文を出力させる手法 [3] や，“Disease: headache” のように固有表現のみを出力させる手法 [11] などが提案されている．一方で，LLM による NER の性能，特に生命医学分野のような専門的な文章に対する性能は，BERT を始めとするエンコーダモデルを用いた最先端モデルに未だ劣っている．

2.2 LoRA

LoRA はモデルが持つパラメタを凍結し，新たに導入した低ランク行列のパラメタのみを更新することで，性能を保ちつつ低コストなファインチューニングを実現する手法である [8]．LoRA では，モデルの初期パラメタ Φ_0 に対してパラメタ数が $|\Theta| \ll |\Phi_0|$ となるパラメタ Θ を導入し，パラメタの差分 $\Delta\Phi(\Theta)$ を以下のように最適化する．

$$\max_{\Theta} \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log(P_{\Phi_0+\Delta\Phi(\Theta)}(y_t|x, y_{<t})) \quad (1)$$

具体的には，モデルのある線形層 $W_0 \in \mathbb{R}^{d\times k}$ に対して，$r \ll \min(d, k)$ の低ランク行列 $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ を用いて $W_0 + \Delta W = W_0 + BA$ とし，B, A のみを最適化する．LoRA は単純な線形変換であるため，BA を W_0 にマージすることができ，推論時の遅延が発生しないという利点もある．入力 x に対して出力 h は以下のように計算される．

$$h = W_0x + \Delta Wx = W_0x + BAx = (W_0 + BA)x \quad (2)$$

3 実験

本研究では，生命医学分野の NER を対象とした LLM のファインチューニングを対象に，LLM の学習設定を，プロンプトの必要性，NER タスクの解法，LoRA の設定に細分化し，段階的に評価を行い，最終的なモデルを他のモデルと比較した．まず，ファインチューニングにおけるプロンプトの必要性を調査するために，学習データに既存研究のプロンプトを用いる場合と，入力文のみを用いる場合の性能を比較した (3.2 節)．次に，テキスト生成を行う LLM に対して，NER タスクをどう解くように学習させればよいかを調査するために，LLM による NER の既存研究におけるタスクの解法を「タグ付け」と「抽出」の 2 種類に分類し，性能の比較を行った (3.3 節)．さらに，NER において LoRA の学習設定が性能に与える影響を調査するために，LoRA を適用する層とランクについて，複数の組み合わせを設定し，性能の比較を行った (3.4 節)．最後に，既存モデル，Full Fine-tuning との比較を行った (3.5 節)．

3.1 設定

LLM には，Llama-2-7B [12] と，対話に特化させるインストラクションチューニングが施された Llama-2-chat-7B を使用した．各実験における学習・評価には，生命医学ドメインの NER データセット BC5CDR を使用した．固有表現ラベルは Disease と Chemical の 2 種類である．BC5CDR のデータの統計を付録 A の表 5 に示す．なお，3.2 節と 3.3 節では，LoRA を学習可能な全ての層に適用し，ランクを 8 に固定してファインチューニングを行った．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

プロンプト

The task is to identify {entity type} entities in the input.
Please rewrite the input text and surround the start and end of {entity type} entities with @@ and ##, respectively.

<few-shot demonstration>

入力文

Input: {input sentence}
Answer: {model output}

図 2　プロンプトの概要

表 1　プロンプトの有無に対する F1 値 (%)

モデル	プロンプト	F1
Llama-2-chat-7B	あり	93.1 ± 0.3
	なし	93.4 ± 0.2
Llama-2-7B	なし	93.6 ± 0.4

3.2　プロンプトの必要性

ICL による NER の Sun ら [3] が使用したプロンプトの概要を図 2 に示す．ファインチューニングを行わない場合，LLM は事前学習で得た知識と ICL に基づいてタスクを解くため，プロンプトは NER の性能に大きく影響する重要な要素となる．一方でファインチューニングを行う場合，プロンプトが性能に影響するかは不明である．

そこで，ファインチューニングを行う場合のプロンプトの必要性を調査するために，図 2 のプロンプトを用いた場合と，入力文のみを用いた場合で，Llama-2-chat-7B のファインチューニングを行った．入力文のみの設定では Llama-2-7B のファインチューニングも行い，対話用のインストラクションチューニングの必要性を調査した．固有表現ラベルは Chemical のみとし，損失は “Answer:” 以降を対象に計算した．ファインチューニング後の開発データに対する評価を表 1 に示す．

Llama-2-chat-7B の結果より，ファインチューニングにおいてプロンプトの有無が性能に与える影響は無視できることが確認できた．また，Llama-2-7B が Llama-2-chat-7B と同等の性能であることから，インストラクションチューニングされたモデルである必要もないことが確認できた．以上の結果を踏まえて，以降の実験では入力文のみのプロンプトと Llama-2-7B を用いる．

3.3　NER タスクの解法

LLM による NER の既存研究では，テキスト生成によって NER タスクを解く様々な方法が考案されているが，その方法は大きくタグ付けと抽出の 2 種類に分類できる．タグ付けによる NER では，入力文に対して固有表現と認識した部分にタグを付けた文を出力させ，抽出による NER では，入力文に対して固有表現と認識した単語のみを出力させる．さらに，両者において各固有表現ラベル毎に独立してモデルを学習・推論させる方法と，全固有表現ラベルを一度に学習・推論する方法に分けられる．

表 2　NER の各解法に対する F1 値 (%)

NER 解法	学習・推論	F1
タグ付け	ラベル毎	88.6 ± 0.6
	全ラベル	88.4 ± 0.6
抽出	ラベル毎	85.8 ± 0.4
	全ラベル	86.5 ± 0.1

ファインチューニングにおいて NER タスクの解法が性能に及ぼす影響を調査するために，NER タスクの 2 種類の解法と，学習・推論ラベルの 2 種類の組み合わせでそれぞれファインチューニングを行い，開発データに対する抽出性能を評価した．各設定の出力フォーマットは付録 B の表 6 の例のようにした．なお，各設定の性能を同一条件で比較するため，各固有表現ラベルごとに独立して学習・推論を行う設定については，1 つのトークンに 1 つのラベルのみが割り当てられるように，Disease > Chemical の優先度でラベル付けする後処理をした．

表 2 の結果より，NER の解法については，抽出よりもタグ付けの方が高い性能が得られた．学習・推論については，各固有表現ラベル毎に行う場合と全固有表現ラベルを同時に行う場合で性能差は見られなかった．以上の結果を踏まえて，以降の実験では NER 解法をタグ付けとし，学習と推論のコストが低い全固有表現ラベルを同時に学習する設定でファインチューニングする．

3.4　LoRA の設定

LoRA を適用する層とランク r の大きさが性能に与える影響を調査するため，それぞれを変化させた場合の比較実験を行った．LoRA はモデルの任意の層に適用でき，その層のみを学習対象にできる．また，ランク r の大きさによって，学習可能なパラメタ数を変更できる．Llama-2 における学習可能な層には，注意機構 (query/key/value/output projection) と全結合層 (up/down/gate projection, embedding, lm head) があり，それぞれ W_q, W_k, W_v, W_o, W_u, W_d, W_g, W_e,

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 3　LoRA を適用する層とランク r の組み合わせに対する F1 値 (%)

LoRA 適用層	$r=2$	$r=4$	$r=8$	$r=64$	$r=128$
W_q	68.3	69.1	65.3	66.9	68.0
W_q, W_k	64.3	61.5	59.6	60.3	60.6
W_q, W_k, W_v	88.3	88.4	89.1	88.2	88.2
W_q, W_k, W_v, W_o	88.4	89.0	88.8	88.3	88.6
$W_q, W_k, W_v, W_o, W_u, W_d, W_g, W_e, W_h$	89.2	89.4	89.0	89.4	**89.5**

表 4　LoRA, Full Fine-tuning, SOTA モデルの性能 (%)

手法	モデル	F1
LoRA	Llama-2-7B	88.9
Full Fine-tuning	Llama-2-7B	89.3
BINDER (SOTA)	PubMedBERT-base	**91.9**

W_h とする．各設定における学習パラメタ数を付録 C の表 7 に示す．

表 3 のファインチューニング後の開発データに対する結果より，LoRA を適用する層が 2 種類以下のとき，性能は大幅に低く，ランクの変動によって性能も大きく変化した．層の種類の増加により，性能は向上し，学習可能な全ての層に適用したときに，ランクの影響を受けずに高い性能が得られた．学習パラメタ数で比較すると，例えば W_q のみの $r=128$ と W_q, W_k, W_v, W_o の $r=2$ では，後者の方が学習パラメタ数が少ないにも関わらず性能が高く，学習パラメタ数が多いほど性能が高くなるとは限らないことが確認できた．以上より，同じパラメタ数を学習させる場合にはランクよりも LoRA を適用する層を多く設定する方が効果的であると考えられる．

3.5　NER の性能評価

3.4 節の結果より，学習可能な全ての層に LoRA を $r=128$ で適用し，テストデータで評価した．また，比較のために Full Fine-tuning を行い，同様に評価した．各手法の学習コストを付録 D の表 8 に示す．それぞれの評価値と BC5CDR において SOTA を達成した Zhang ら [13] の性能を比較する．

表 4 の結果より，LoRA を適用することで，少ない学習パラメタ数で Full Fine-tuning と同等の性能を達成した．しかし，SOTA モデルの性能を上回ることはできず，更なる性能向上を目指す必要がある．

4　考察

NER に対する LLM のファインチューニングにおいて，プロンプトの有無は性能に影響を与えなかった．これは，大量のタスクデータを学習したことで，推論の手助けとなるタスクの説明や Few-shot の例示の必要性がなくなったからだと考えられる．プロンプトを除いた最低限のデータで学習できるため，コストの削減につながる．

NER の解法については，抽出よりもタグ付けの方が性能が高くなった．入力文が与えられたときに，単に固有表現のみを出力させるよりも，再度同じ文を出力しつつ途中でタグを付ける方が文脈を捉えることができると考えられる．

LoRA については，ランクよりも適用する層の種類の方が NER における性能への影響が大きいという結果が得られた．LoRA を適用する層が幅広いほど，各層が担う NER に必要な潜在能力を幅広く刺激できたと考えられる．

5　おわりに

本研究では生命医学分野の NER を対象として，LLM の LoRA を用いたファインチューニングにおいて，有効な学習設定を明らかにするために，プロンプトの必要性，NER タスクの解法，LoRA の設定について，それぞれが性能に及ぼす影響を調査した．ファインチューニングを行う場合，プロンプトを設計する必要はなく，タグ付けによる NER の解法を用いることで高い性能が得られることがわかった．LoRA については，ランクよりも適用する層の種類の方が重要で，学習可能な全ての層に LoRA を適用したときに極めて少ないパラメタ数で Full Fine-tuning と同等の性能を達成することが確認できた．

今後の課題として，LoRA を適用する層の組み合わせについて詳細に調査し，より低コストで高性能を発揮するための学習設定を明らかにすることが挙げられる．また，LoRA 以外にも低コストでファインチューニングをする手法が複数存在するため，それらの手法との比較も必要である．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです.

参考文献

[1] 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.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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, June 2019. Association for Computational Linguistics.

[3] Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, et al. Pushing the limits of chatgpt on nlp tasks. **arXiv preprint arXiv:2306.09719**, 2023.

[4] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gptner: Named entity recognition via large language models. **arXiv preprint arXiv:2304.10428**, 2023.

[5] Mingchen Li and Rui Zhang. How far is language model from 100% few-shot named entity recognition in medical domain. **arXiv preprint arXiv:2307.00186**, 2023.

[6] Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised llama finetuning. **arXiv preprint arXiv:2310.01208**, 2023.

[7] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction. **arXiv preprint arXiv:2310.03668**, 2023.

[8] 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.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[10] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. Instructuie: Multi-task instruction tuning for unified information extraction. **arXiv preprint arXiv:2304.08085**, 2023.

[12] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[13] Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. Optimizing bi-encoder for named entity recognition via contrastive learning. In **The Eleventh International Conference on Learning Representations**, 2023.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

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 の，固有表現ラベルごとの事例数を表 5 に示す．

表 5　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

B　使用したプロンプト

3.3 節で設定した，タグ付けによる NER と抽出による NER の出力フォーマットの例を表 6 に示す．

表 6　入力文 “Mesna significantly reduces IFO ’ s genotoxicity” に対する出力フォーマット

手法	学習ラベル	出力フォーマット
タグ付け	Disease	Mesna significantly reduces IFO ’ s @@genotoxicity##
	Chemical	A@@Mesna## significantly reduces @@IFO## ’ s genotoxicity
	全ラベル	An [Chemical]Mesna[/Chemical] significantly reduces [Chemical]IFO[/Chemical] ’ s [Disease]genotoxicity[/Disease]
抽出	Disease	genotoxicity
	Chemical	Mesna, IFO
	全ラベル	Disease: genotoxicity; Chemical: Mesna, IFO

C　LoRA のパラメタ数

3.4 節における，LoRA を適用する層とランクの組み合わせに対する学習パラメタ数を表 7 に示す．

表 7　LoRA を適用する層とランク r の組み合わせに対する学習パラメタ数

LoRA 適用層	$r=2$	$r=4$	$r=8$	$r=64$	$r=128$
W_q	524,288	1,048,576	2,097,152	16,777,216	33,554,432
W_q, W_k	1,048,576	2,097,152	4,194,304	50,331,648	67,108,864
W_q, W_k, W_v	1,572,864	3,145,728	6,291,456	6,291,456	100,663,296
W_q, W_k, W_v, W_o	2,097,152	4,194,304	8,388,608	67,108,864	134,217,728
$W_q, W_k, W_v, W_o, W_u, W_d, W_g, W_e, W_h$	5,141,504	10,283,008	20,566,016	164,528,128	329,056,256

D　学習コスト

3.5 節における，LoRA と Full Fine-tuning の学習コストを表 8 に示す．

表 8　LoRA と Full Fine-tuning の学習コスト

	LoRA	Full Fine-tuning
学習パラメタ数	329,056,256	6,738,415,616
GPU	V100 (16GB) × 4	A100 (40GB) × 8
ストレージ容量	3.7GB	26.3GB
学習時間	2.5 h/epoch	2.7 h/epoch

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).