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

楢木悠士 * ♯ [1]　山木良輔 ♯ [2,6]　池田愛和 [3,6]　堀江孝文 [2]　長沼大樹 [4,5,6]
[1]LegalOn Technologies　[2] 立命館大学　[3] 大阪大学
[4]Université de Montréal　[5]Mila　[6]ProPlace Inc.
yuji.naraki@legalontech.jp　{yamaki.ryosuke, horie.takafumi}@em.ci.ritsumei.ac.jp
{yoshikazu.ikeda, hiroki.naganuma}@proplace.jp

概要

固有表現抽出（NER）は自然言語処理の重要なタスクであり，幅広く応用されている．しかし，人手によるアノテーションは，コストの高さ・倫理的問題・品質の一貫性・機密情報保護等の問題が喫緊の課題である．本研究では，大規模言語モデル（LLMs）から生成したアノテーションを人手によるものと統合することで，この課題に取り組む．また，LLMs によるアノテーションではクラス分布の不均衡が引き起こるという副次的な問題を特定し，対処策として Label Mixing を提案する．提案手法により NER タスクの性能を強化するだけでなく，コストの大幅な効率化を図る．複数のデータセットで比較した結果，提案手法は，同一のアノテーションコスト下で，人手のアノテーションに比べ高い性能を示しただけでなく，学習データセットの品質が低い場合に顕著な性能の安定性向上に寄与した．

1　はじめに

自然言語処理の分野において，固有表現抽出（Named Entity Recognition; NER）は文章中の固有表現を特定し，“person”，“organization”，“location” などのカテゴリに分類する重要なタスクである [1]．NER は基本的な情報検索やコンテンツの分類から，質問応答などのより複雑なタスクまで，幅広い応用へ必要不可欠の技術である [2, 3]．NER の重要性は，これらの応用に留まらず，テキストデータに対する我々の理解と関わりを強化する役割にまで及ぶ [4, 5]．

このような有用性が存在する一方で，NER システムは，利用するデータセットのアノテーションの質に大きく依存する [6]．現在，人手によるアノテーションが標準的であるが，それにかかる人・時間・金銭的コストや倫理的な問題の面で課題が存在している [7]．また，人手のアノテーションは質や一貫性の担保が困難であり、NER モデルの信頼性を損ねる可能性がある [8]．さらに，ドメイン固有の分野においてはアノテーションを施したデータセットが存在していないため，教師データの確保が課題として存在する．[9]

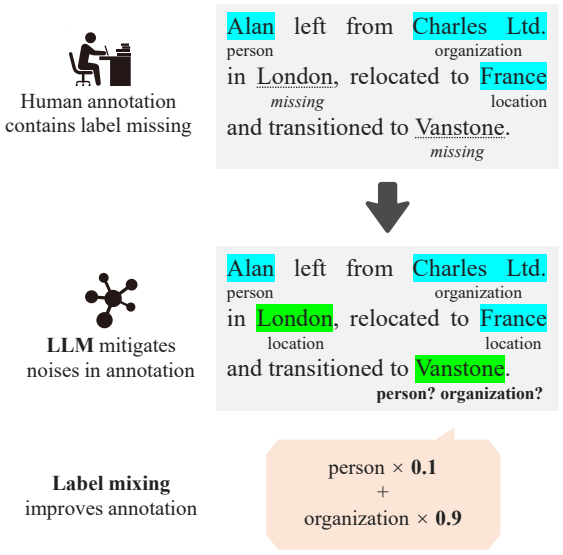

図 1　LLM によるアノテーションと Label Mixing.

上述の問題に取り組むために，我々は NER タスクの質・精度・コスト・倫理的問題などの側面において強化する新たなアプローチを提案する．図 1 に示されるように，我々の手法は，質や一貫性に問題のある人手のアノテーションを大規模言語モデル（LLMs）を活用し，その品質を改善する．このハイブリッドなアノテーション手法は，主にデータセット内のノイズを減少させること目的としている．さらに，我々の提案手法はクラス不均衡の改善にも

* 本研究は，所属の LegalOn Technologies とは無関係である．
♯ Equal contribution

寄与する．このクラス不均衡は特に “Miscellaneous” カテゴリにおいて顕著である．本稿における実験では CoNLL03[10]，WikiGold[11] データセットを用い，NER モデルには BERT[12] を用いた．実験結果から予算の制約がある場合においても我々のアプローチが顕著な改善を示している．

本研究では LLMs を用いた NER タスクへの以下の 3 つの項目を実証する．

- データセットにノイズが存在する場合でも，LLMs を活用することで NER タスクの性能を向上させられる (図 2).
- 予算を固定した実験において，人手と LLMs によるアノテーションのアプローチがどちらか一方のみを用いた際のアプローチを上回る (表 1).
- LLMs を用いたアノテーションの際に，特に悪化する “Miscelaneous” カテゴリのクラス不均衡問題のための解決策として，Label Mixing を提案 (図 4).

2 関連研究

NER タスクで使用されるデータセットの信頼性は精査の対象である．Han ら [13] の研究では，CoNLL03[10] などのデータセットに固有の誤りが存在することを強調している．これらの不正確さは，エンティティへの誤ったラベル付けだけでなく，一貫性のない分類など，NER システムの全体的な性能と信頼性に影響を与える [14, 6].

人手によるアノテーションにおける限界への潜在的な解決策として，自動アノテーションが注目を集めている．Wang ら [15] は，GPT-3.5 [16] のような大規模言語モデルをアノテーションタスクに使用し，正確で信頼性の高いアノテーションを生成することの有効性を探求している．

本研究は，自動アノテーションを活用するだけでなく，NER データセットの複雑な課題に対処するための人手と LLM によるアノテーションの両方を統合する混合アプローチを提案する．この点で，本研究は Wang ら [15] とは異なる．また，実用的な設定として，少数の人手のアノテーションが存在する設定での実験を行なっている．

3 手法

LLM によるアノテーション: LLM を用いた言語データのアノテーション自動化において，Wang ら [15] が示した GPT-NER の手法は，LLM を用いて自動的にアノテーションを行った．LLM がテキスト内の特定のエンティティを識別し，これらのセグメントを特定の文字列（“@@” と “##’）で囲む能力を活用することが GPT-NER 戦略の中心的な要素である．具体的には，入力として *“I live in Tokyo”* が与えられた場合，目的は `LOC`（location）エンティティを特定することである．このとき，LLM の出力は *“I live in @@Tokyo##”* となる．

さらに，本手法は，各エンティティに固有の特性とそれらの間の相互作用を few-shot の入出力例によって LLM に理解させることができる．上記の内容を踏まえ，以下のプロンプトを構築した．

大規模言語モデルへ入力するプロンプト

You are an excellent linguist. Identify [*target entity type*] entities in given text. If the text contains no [*target entity type*] entities, replicate the input text without any changes. Strictly adhere to the definition provided below.
Definition of [*target entity type*]:
[*definition target entity type*]
[*Example* 1 *Input*]
[*Example* 1 *Output*]
· · ·
[*Example K Input*]
[*Example K Output*]

このプロンプト設計では，[*target entity type*] スロットには，`PER`，`ORG`，`LOC`，`MISC` などの様々なエンティティタイプが含まれる．同時に，[*definition target entity type*] スロットは，エンティティタイプの定義が割り当てられる．さらに，[*Example K Input*] と [*Example K Output*] スロットは，それぞれ入力文とそれに対応する出力例が割り当てられ，few-shot 例として機能する．

このプロンプトにより，アノテーション対象のテキスト内の各対象エンティティタイプに対して，固有表現抽出が行われる．この戦略では，各入力文に対して複数回の推論を必要とするため，1 つの単語に複数の異なるエンティティタイプが割り当てられる可能性がある．我々は，最も単純な方法として，エンティティの重複が起きた場合に複数の候補から 1 つのエンティティをランダムに選択する方法を採用する．

few-shot の入出力例を選択する基準は，GPT-NER の手法に触発されたものであり，入力テキストに存在するトークンの埋め込み表現のコサイン類似度に基づいて，検証データから例を選択する．

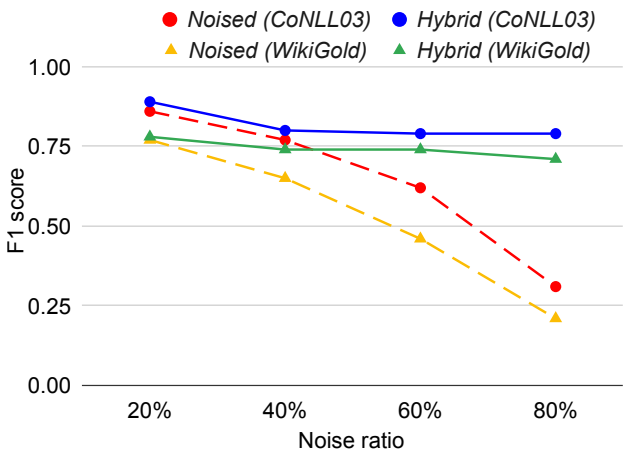

図 2 ノイズを含むデータセットと Hybrid アノテーションを用いた場合のモデルの性能の比較.

Hybrid: LLM ベースと人手のアノテーション: データセットの品質を向上させるために，LLM によるアノテーションと人手のアノテーションをマージして Hybrid アノテーションを作成した．マージされたデータは，人手によるすべてのエンティティラベルと，それに重ならない LLM によるエンティティラベルを組み合わせたものである．人手によるエンティティラベルの方が LLM によるものよりも高品質であることが期待されるため，人手によるものを優先してマージする．これにより各ソースに存在しうるバイアスやエラーを軽減し，より堅牢で多様なアノテーションを提供する．

LLM ベースのアノテーションにおける Label Mixing: 前述した 1 つのトークンに複数のラベルが割り当てられてしまう問題に対処するために，画像処理で伝統的に使用されている Mixup [17] を NER タスクに拡張した．我々のアプローチは，1 つのトークンに対して 2 つの異なるエンティティラベルを混合して新しいアノテーションを作成することである．この方法は以下のように形式的に表現される．

$$y_{\text{mix}} = \lambda \times y_i + (1 - \lambda) \times y_j$$

ここで，y_i と y_j は 2 つのラベルが付けられた NER タグであり，$\lambda \in [0, 1]$ は混合比を決定するパラメータである．Mixup [17] に従い，λ はベータ分布（$\alpha = \beta = 0.2$）から引き出される．直感的には，複数のエンティティタイプに属する可能性のあるトークンに対し，不確実性を持った推定に貢献することが期待される．

表 1 固定の予算の下で，手動と LLM ベースのデータの異なるデータ数のバランスと性能の比較.

Budget	#Manual	#LLM	F1 Score
$38	87	0	0.02
	70	702	0.81
	35	2106	0.84
	0	3510	**0.85**
$152	351	0	0.63
	280	2808	0.85
	140	8424	**0.87**
	0	14041	0.86
$608	1404	0	0.86
	1263	5616	**0.87**
	1123	11232	**0.87**

4 実験

4.1 実験設定

我々の研究では，2 つの異なるデータセットを用いて実験を行った．まず，NER の分野で標準的なベンチマークとして認識されている CoNLL03 データセット [10] を用いた．また，CoNLL03 データセットにおける性能を考慮し，より難易度の高い WikiGold データセット [11] についても実験を行った．WikiGold データセットは，CoNLL03 と比べると，アノテーションの品質やデータ数が少ないなどの要因から高い性能を達成することが難しい．WikiGold データセットは，実験のために 7:1:2 の比率で訓練，検証，テストセットに分割した．

自動アノテーションでは，LLM として ChatGPT (`gpt-3.5-turbo-0613`) API[1)] を利用し，プロンプトに含める few-shot の例の数を 32 とした．

最後に，我々の提案するアノテーション手法の有効性を検証するために，提案手法によってアノテーションされたデータセットを用いて，事前学習済み BERT [12] に基づく分類モデルをファインチューニングした．NER タスクの評価は Sang らの研究 [10] に準拠し，本論文では重み付き平均 F1 スコアを示す．ハードウェアおよびソフトウェアの構成については A.1 で詳細に説明する．

4.2 実験結果

LLM によるアノテーションによるノイズ回復: まず初めに我々は，LLM のアノテーションによるノイズ回復能力を評価した．Han ら [18] は，CoNLL03 [10] において，アノテーションの欠落やラベルの入れ替えなどのエラーを指摘した．我々は，一部のアノ

1) https://chat.openai.com

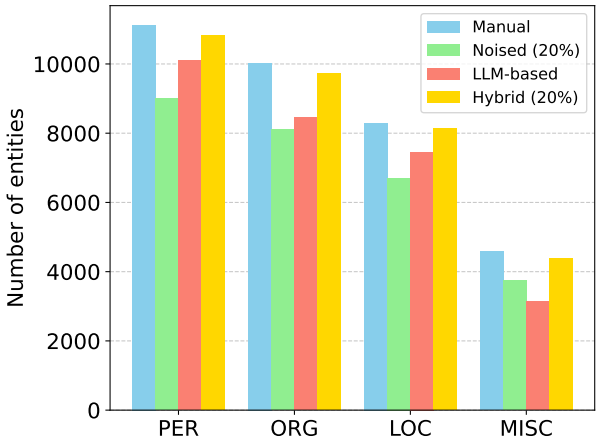

図 3 人手，ノイズ付き，LLM ベースのデータセットの各エンティティタイプのアノテーション数.

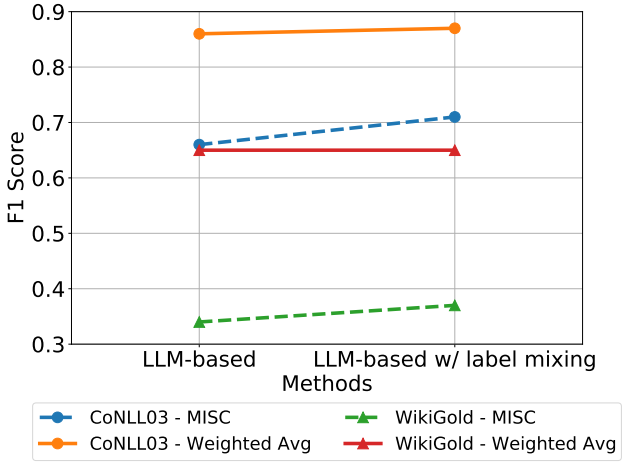

図 4 Label Mixing による最も性能の低いエンティティタイプ（MISC）の改善.

テーションをランダムに削除することで，ノイズ付きデータを用意した．前述の方法で LLM によるアノテーションを組み合わせることで，これを改善できることを示す．20%，40%，60%，80%のエンティティのアノテーションを削除したデータについて実験を行った．ノイズ付きデータと比較すると，提案の組み合わせ手法によって改善されたデータセットでは，性能の改善が確認された．図 2 は，これらのデータセットの比較分析からノイズ低減の効果を示している．特に，ノイズが深刻な場合でも，LLM によるアノテーションを組み合わせることで，大幅に改善できると示された．

同じ予算の下での比較: Ding ら [19] の研究と同じように，同一の予算制約の下で我々の手法を評価した．我々の実験は，3 つの異なる予算（$38, $152, $608）で行った．人手と LLM によるアノテーションから異なるデータ数を含む複合データセットを作成した．中間予算（$152）は，CoNLL03 のすべての訓練データの LLM でアノテーションしたときのコストであり，小規模（$38）と大規模（$608）の予算については，それぞれその 4 分の 1 と 4 倍のコストを想定した．各予算のデータ数のバランスと性能の関係を表 1 に示す．

予算が小さい場合，データ量が少なくなるため，人手のデータだけでは学習ができない．表 1 から，LLM によるアノテーションを用いて学習データを増やすことで，性能が向上したと考えられる．中間予算（$152）では，人手のデータだけを用いた場合，モデルの性能が低下した．LLM によるアノテーションを 2,808 個だけ用いた場合，有意な改善が見られた．また，Hybrid アノテーションの性能は，LLM によるのみの性能よりも高く，両方のアノテーションを用いることが性能に貢献することを示唆している．大規模予算（$608）では，人手のデータを 1,404 個用いるだけで十分な性能が得られるが，人手のデータを減らし，LLM によるアノテーションも用いることで，わずかな性能向上が確認された．これらの結果から，複合データセットは，低コストでデータ数を増やし，すべての予算下で下流タスクの性能を向上させることが確認できた．

クラスの不均衡問題の解決法: 図 3 は，各エンティティタイプの数に差があることを示している．また，LLM によるアノテーションで学習したモデルのエンティティタイプごとの性能の差は，人手のアノテーションよりも大きい．クラスの不均衡問題に対処するために，Label Mixing の使用を提案した．図 4 に LLM によるアノテーションと Label Mixing の比較を示す．Label Mixing により，タイプ間の性能差が減少した．したがって，one-hot のハードラベルではなく，複数のエンティティの可能性を考慮した LLM によるアノテーションのソフトラベルにより，モデルの学習がバイアスを持って行われることを防ぐことができると考えられる．

5 結論

本研究では，人手と LLM によるアノテーションを組み合わせることでノイズを含むデータや少量データに対して，NER タスクの性能を改善させることができると示した．LLM の性能は今後も向上していく [20, 21] と考えられ，現状の LLM で機能していることから，LLM による自動アノテーションは今後も有望だと考えられる．

謝辞

本研究は， GCP Startups Booster Program 及び Microsoft for Startups の助成を受けたものである．

参考文献

[1] Behrang Mohit. Named entity recognition. In **Natural language processing of semitic languages**, pp. 221–245. Springer, 2014.

[2] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In **Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '09, p. 267–274, New York, NY, USA, 2009. Association for Computing Machinery.

[3] Diego Mollá Aliod, Menno van Zaanen, and Daniel Smith. Named entity recognition for question answering. In **Australasian Language Technology Association Workshop**, 2006.

[4] Pengxiang Cheng and Katrin Erk. Attending to entities for better text understanding. **ArXiv**, Vol. abs/1911.04361, , 2019.

[5] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. **Artif. Intell.**, Vol. 165, pp. 91–134, 2005.

[6] Ruslan Mitkov, Constantin Orasan, and Richard J. Evans. The importance of annotated corpora for nlp: the cases of anaphora resolution and clause splitting. 2000.

[7] Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? **Computational Linguistics**, pp. 413–420, 2011.

[8] Arzoo Katiyar and Claire Cardie. Nested named entity recognition revisited. In **North American Chapter of the Association for Computational Linguistics**, 2018.

[9] Siliang Tang, Ning Zhang, Jinjiang Zhang, Fei Wu, and Yueting Zhuang. NITE: A neural inductive teaching framework for domain specific NER. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2652–2657, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[10] 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.

[11] Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. Named entity recognition in Wikipedia. In Iryna Gurevych and Torsten Zesch, editors, **Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)**, pp. 10–18, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[12] 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.

[13] Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. **ArXiv**, Vol. abs/2305.14450, , 2023.

[14] Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing demographic bias in named entity recognition. **arXiv preprint arXiv:2008.03415**, 2020.

[15] 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**, Vol. abs/2304.10428, , 2023.

[16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[17] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In **International Conference on Learning Representations**, 2018.

[18] Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors, 2023.

[19] 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.

[20] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

[21] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. **ACM Computing Surveys**, Vol. 56, No. 2, pp. 1–40, 2023.

A Appendix

A.1 実験設定の詳細

ハードウェアの設定:実験では，Google Cloud 上の仮想マシンを用いた．マシンの詳細は以下の通りである．

- マシンタイプ: `n1 standard 4`
- メモリ: 15 GB
- 仮想 CPU: `Intel Broadwell` x 4
- GPU: `NVIDIA T4` x 1

BERT の設定:NER タスクの性能評価に用いた BERT の設定は以下の通りである．

- 事前学習済みモデル: `bert-model-unbased` [2)]
- 学習率: $1e-05$
- 勾配ノルムの上限値: 10
- バッチサイズ: 学習時は 64，推論時は 32
- 最大トークン長: 128

学習は，3 エポック続けて Accuracy が変化しなくなったところで終了することとした．

NER では各単語単位にラベルを割り振る必要があるが，BERT のトークナイザーでは一つの単語を複数の単語に分割する場合がある．そこで本研究では，単語が複数のトークンに分割された場合，先頭のトークンに対するラベルを単語全体のトークンとして最小している．

2） https://huggingface.co/bert-base-uncased