LLM を用いた関係抽出のデータ拡張における データ選定と利用方法の検討

小島大世　三輪誠
豊田工業大学
{sd21043,makoto-miwa}@toyota-ti.ac.jp

概要

関係抽出における LLM を用いたデータ拡張において，生成されたデータを評価するための指標と拡張データを利用する学習手法について提案する．具体的には，生成されたテキストの正しい関係トリプルの内容の保持と多様性の評価のため，元の訓練データのテキストと生成されたテキストの比較方法について検討・比較を行う．また，拡張データと訓練データの分布の違いによる悪影響を抑えるために拡張データと訓練データを順に学習に利用する手法について提案する．DrugProt データセットを利用した実験では，拡張データの違いが抽出性能に大きく影響することと提案した学習手法によって拡張データによる悪影響を抑えられることを確認した．

1　はじめに

テキストデータから人物や場所などの特定の用語ペアとその間の関係を抽出する関係抽出が広く研究されている [1, 2]．高性能な関係抽出の実現には，ラベル付きデータを用いたファインチューニングが有効である．しかし，ラベル付きデータの作成にはコストと時間がかかるため，大量のデータの収集は難しい．

この問題に対処するため，データ拡張 [3, 4] が広く研究されている．この中でも，近年は大規模言語モデル (Large Language Models; LLM) の高い言語生成能力を利用したデータ拡張手法が注目されている [5, 6]．これらの手法では，訓練データ内のテキストとそれに含まれる用語とその関係を元に，その関係を含む新たなテキストを作成する．LLM は指示された用語ペアとその関係で表される**関係トリプル**を含む多様な文構造や表現のテキストを生成できる．生成されたテキストは指示された関係トリプルを正確に保持することと多様性のあるデータであることが必要であるが，この評価を行う手法は確立されていない．また，生成されたテキスト集合が訓練データの分布を反映しているとは限らないため，単純に拡張データを追加してしまうと元の訓練データの分布を壊してしまう可能性がある．そこで，本研究では，関係抽出における LLM を用いたデータ拡張において，生成されたテキストが正しい関係トリプルの内容を保持できているかと多様であるかを評価するための指標と作成された拡張データの分布の違いによる影響を抑えるための学習への利用方法の検討を目的とする．指標については，元とするテキストと生成されたテキストそれぞれについて，直接の比較と用語ペアに注目した比較を行う指標について検討・比較する．用語ペアにフォーマスした比較では，LLM を用いて用語ペアに注目した要約を作成する手法を提案する．また，学習への利用においては作成した拡張データで事前にファインチューニングし，さらに元の訓練データでファインチューニングすることで拡張データの分布の違いによる悪影響を抑える手法について提案・評価する．

本研究の貢献は以下のとおりである．

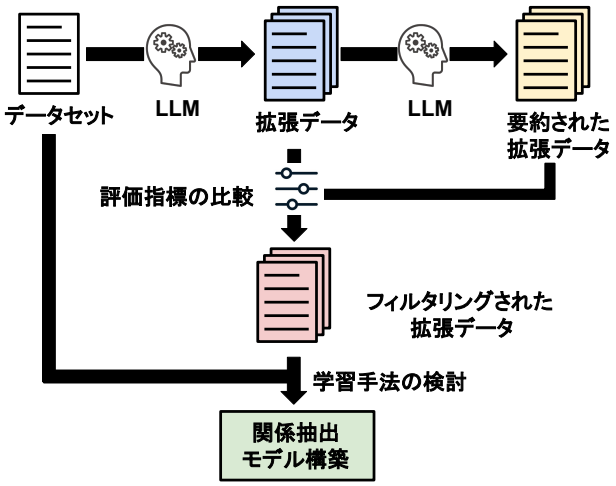

図 1　提案手法の概要図

- 関係抽出における LLM を用いたデータ拡張において，生成した拡張データを評価するための指標を比較し，DrugProt データセット [7] を利用した実験において，拡張データの違いが関係抽出の性能に大きく影響することを示した．
- 拡張データを用いた学習において，拡張データと訓練データの分布の違いによる悪影響を抑えるための手法について提案し，性能の低下を抑えることができることを確認した．

2 関連研究

従来から，関係抽出タスクのデータ不足を補うために，データ拡張が広く研究されてきた．Mintz ら [8] は，既存の知識ベースを活用し，対象となる用語間の関係を含む文を収集することで擬似的なラベル付きデータを生成する遠距離教師学習を提案した．この手法により，大規模な未ラベルデータから多様な関係データを効率的に抽出することを可能とした．また，Yu ら [9] は既存のラベル付きデータを基に，バックトランスレーションを用いることで多様な表現のパラフレーズ文を生成することでモデルの性能を向上させた

近年では，LLM の進歩により，関係抽出のデータ拡張手法がさらに発展している．Hu ら [5] は，LLM を用いて，文の意味的一貫性と構文構造を維持しつつ，多様性のある文を生成する手法を提案している．また，Zhou ら [6] は，LLM を用いて，元の文と意味的には類似しているが異なる表現の文を生成するパラフレーズ手法と，関係トリプルを基に新たな文を生成する手法を提案している．これらの研究では生成した擬似的なデータによる関係抽出モデルの性能向上が報告されている．

3 提案手法

本研究では，関係抽出における LLM を用いたデータ拡張において，正しい関係トリプルを含んでいるかと多様であるかを評価するために生成されたテキストを評価する指標と作成された拡張データを有効に利用するための学習手法について提案する．本節ではまず，データ拡張手法について説明し，次に生成されたテキストを判定する指標について説明する．最後に，作成した拡張データを利用した学習方法について説明する．

図 2 データ拡張の例

3.1 LLM を用いたデータ拡張

LLM を用いたデータ拡張については，以下の 5 つのセクションで構成されたプロンプトを LLM に入力することでデータを生成する．図 2 にプロンプトの例を示した．

Instruction: LLM の役割とデータセットの概要を説明する．

Entities: データセットで使用される用語タイプの定義を説明する．

Relationship Definitions: データインスタンス内の関係タイプの定義を説明する．図 2 の例では，データインスタンスの関係トリプル [“DPHD:CHEMICAL”,“ACTIVATOR”, “MAP kinase:GENE”] の関係タイプ ACTIVATOR についての定義を説明している．

Data Instance: 生成すべきデータに関する具体的な指示を示す．関係トリプルと元の文を提示し，それを参考に条件を満たす文を生成するよう指示する．

Example Format: 生成する文の形式の指示を示す．具体的には，各文を改行で分け，番号や追加の装飾なしで記載する指示を行う．

多様な文を作成するために，図 2 のように元の文と意味的に類似するような文を生成するように指示をする場合と以下のようにプロンプトの Data

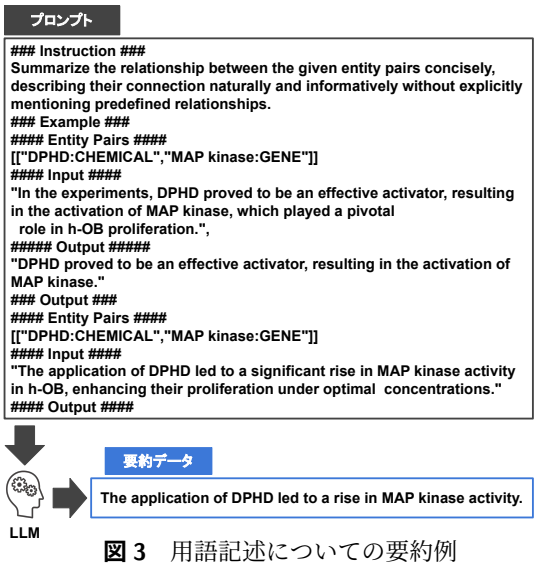

図 3　用語記述についての要約例

Instance の一部を変更して元の文と意味的に異なるように指示する場合の 2 種類でデータ拡張を行う．

2. The sentences must not be semantically similar to the following sentence

この 2 種類のプロンプトで 10 個の文を生成させることで，計 20 個の拡張データを生成する．

3.2　生成テキストを評価する指標

生成テキストの評価には，元の訓練データと拡張データのテキストを直接比較する指標と用語ペアに注目して比較する指標の 2 つの評価指標を利用する．テキストの比較には，Sentence-BERT（SBERT）[10] で文の埋め込みに変換し，コサイン類似度を用いて類似度スコアを計算し，拡張データの評価指標として利用する．本節では，以降，後者の用語ペアに注目して比較する指標について説明する．

用語ペアに注目した比較では，訓練データのテキストと拡張データのテキストの両方について，LLM を用いて用語ペアに注目した要約を生成する (図 3)．ここで，用語ペアに関わる関係トリプルの関係以外の関係が記述されていたり，関係トリプルの関係が記述されていなかったりする可能性を考慮し，要約の生成においては関係の情報を与えないこととする．この要約により，テキストに含まれる用語ペア以外の情報を排除する．作成した要約は，SBERT を用いて比較を行い，評価指標として利用する．

3.3　拡張データを利用した学習

作成した拡張データを利用した学習では，拡張データを訓練データに追加してファインチューニングしたのちに，訓練データで再度ファインチューニングを行う．再度のファインチューニングにより，拡張データと訓練データにおける分布の違いによる悪影響を緩和することを期待する．

表 1　実験で利用したハイパーパラメタ

パラメータ	値
学習率	5e-5
ミニバッチサイズ	4
エポック数	10
ウォームアップ比率	0.1
重み減衰	0.1

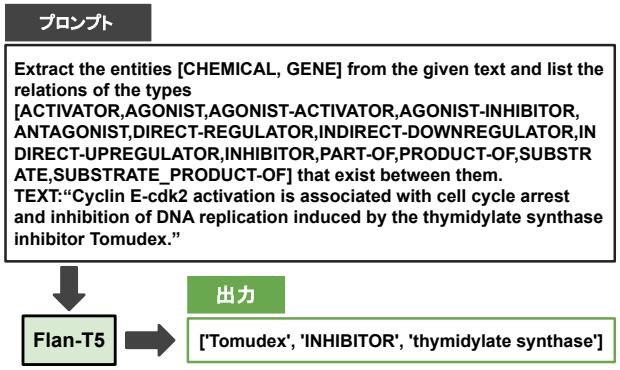

図 4　テキスト生成による関係抽出

4　実験

4.1　実験設定

関係抽出タスクのデータセットとしては，DrugProt データセット [7] を使用する．DrugProt は，生物医学領域における関係抽出タスクのために設定されたデータセットである．このデータセットには，薬物とタンパク質の間の 13 種類の関係が注釈された文が含まれる．本研究では，文単位での関係抽出を対象としているため，文を跨いだ用語ペアとその間の関係は評価には含めないこととした．結果として，文単位の関係は，訓練データが 6,594 件，検証データが 1,427 件となった．DrugProt にはテストデータが公開されていないため，検証データを 2 分割し，その半分をテストデータとして用いた．

評価には，高い性能が報告されている Flan-T5[11] を使用した関係抽出モデル [12] を利用した．既存研究に従って，Flan-T5 に文を入力し，文に含まれる用語ペアとその間の関係を出力するテキスト生成タスク (図 4) として，関係抽出を行うモデルを構築した．

ベースモデル及び拡張データによって構築するモデルは，表 1 に示すようなハイパーパラメタで学習を行った．モデルの性能評価は適合率（Precision）と再現率 (Recall) の調和平均である F 値 (Micro-F) を

表 2　拡張データの選択基準と異なる学習による関係抽出性能の比較

拡張データの選択基準	適合率 (%)	再現率 (%)	F 値 (%)
なし（訓練データのみ）	**63.95**	69.02	66.39
拡張データを訓練データに追加して学習			
ランダム	57.28	65.80	61.25
直接比較した類似度スコアが低い	62.39	70.82	66.30
直接比較した類似度スコアが高い	62.35	70.69	66.26
用語ペアに注目した類似度スコアが低い	57.47	67.41	62.05
用語ペアに注目した類似度スコアが高い	60.63	67.29	63.79
拡張データの後に訓練データを学習			
ランダムデータ	61.15	71.00	65.71
直接比較した類似度スコアが低い	62.49	**71.62**	**66.74**
直接比較した類似度スコアが高い	62.46	70.81	66.37
用語ペアに注目した類似度スコアが低い	61.22	70.63	65.54
用語ペアに注目した類似度スコアが高い	61.43	70.57	65.69

用いた．

3.1 節に従って，訓練データに含まれるそれぞれの関係について拡張データを生成した．3.2 節の指標を用いて，拡張元の訓練データのそれぞれのテキストと拡張したテキストをスコア付けし，以下の 5 通りの選択基準で 20 個のデータから 1 個ずつ拡張データの選択を行った．

- ランダム
- 直接比較した類似度スコアが低い
- 直接比較した類似度スコアが高い
- 用語ペアに注目した類似度スコアが高い
- 用語ペアに注目した類似度スコアが低い

学習では，訓練データに直接拡張データを追加した場合と 3.3 節で説明した拡張データで学習した後に訓練データで学習する方法について比較を行う．

4.2　関係抽出性能

DrugProt について，5 通りの異なる選択方法で作成した拡張データを 2 通りの学習方法で比較した結果を表 2 に示す．

結果より，まず，拡張データの違いが関係抽出の性能に大きく影響することが分かる．同じ拡張データを追加した学習の設定でも，ランダムに選んだ場合は大きく性能が低下し，用語ペアに着目した類似度スコアを用いた場合も一定の低下が見られるが，直接比較した類似度スコアで選んだ時は性能の低下が小さい．それぞれの指標で選んだデータが学習にどのような影響を及ぼしているかについては，より深い解析が必要である．

次に，全ての場合において拡張データを訓練データに追加して学習するより，拡張データの後に訓練データを学習する方が性能の低下を抑えることができている．これは拡張データの後に訓練データを学習することが，拡張データを用いた学習において，拡張データと訓練データの分布の違いによる悪影響を抑えることを示している．

最後に，直接比較した類似度スコアが低い拡張データを利用した際に，若干ではあるが，訓練データのみの抽出性能より高い性能が得られた．直接比較した類似度スコアに着目すると，スコアが高いデータでは学習方法の違いに影響が少ない一方で，スコアが低いデータでは学習方法によって性能の変化が見られる．この結果は，スコアが高い基準の場合は訓練データと似た分布のデータを選択し，スコアが低い場合は多様なデータを選択していることが原因ではないかと考えられる．

5　おわりに

本研究では，関係抽出における LLM を用いたデータ拡張において，生成されたテキストの正しい関係トリプルの内容保持と多様性を評価するための指標と作成された拡張データの分布の違いによる影響を抑えるための学習への利用方法の検討を目的に，元のテキストと生成されたテキストを比較する 2 つの指標と拡張データと訓練データを順に学習する学習手法について提案した．提案した指標を用いた 5 通りのデータ拡張を比較した結果，拡張データの違いが関係抽出の性能に大きく影響することを確認した．また，拡張データの後に訓練データを学習することで，拡張データによる関係抽出モデルの性能への悪影響を抑えることができることを示した．

今後は．提案した指標や学習手法についての解析・改善を進めるとともに，大規模なデータ拡張を行っていく予定である．

参考文献

[1] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In **Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP**, pp. 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[2] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In **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. 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Yannis Papanikolaou and Andrea Pierleoni. DARE: Data augmented relation extraction with GPT-2. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1–9, Online, November 2020. Association for Computational Linguistics.

[5] Xuming Hu, Aiwei Liu, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, and Philip S. Yu. GDA: Generative data augmentation techniques for relation extraction tasks. **arXiv preprint arXiv:2305.16663**, 2023.

[6] Yang Zhou, Shimin Shan, Hongkui Wei, Zhehuan Zhao, and Wenshuo Feng. PGA-SciRE: Harnessing llm on data augmentation for enhancing scientific relation extraction. **arXiv preprint arXiv:2405.20787**, 2024.

[7] Rafael Miranda, Martin Krallinger, Maria Perez, and Miguel Vazquez. DrugProt Task Overview: A Biomedical Relation Extraction Shared Task. In **Proceedings of the BioNLP Shared Task 2023**, Online, 2023. Association for Computational Linguistics.

[8] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In **Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 1003–1011, Suntec, Singapore, 2009. Association for Computational Linguistics.

[9] Author Name and Co author Name. Improving relation extraction with advanced deep learning techniques. **Journal of Computational Linguistics**, Vol. 46, No. 3, pp. 345–360, 2020.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In **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. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. **arXiv preprint arXiv:2210.11416**, 2022.

[12] Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Revisiting relation extraction in the era of large language models. **arXiv preprint arXiv:2305.05003**, 2024.