デコーダモデルを用いた生物医学イベント抽出

金児一矢　三輪誠
豊田工業大学
{sd24416,makoto-miwa}@toyota-ti.ac.jp

概要

生物医学イベント抽出は階層的な構造のイベントが多く現れるため，イベント間の関係も考慮した抽出が必要となる．従来研究ではエンコーダモデルを利用した手法が多く提案されているが，モデルサイズなどの制約から，複雑な関係を捉える言語能力に限界があった．そこで，本研究では生物医学分野におけるデコーダモデルを用いた階層的な構造を持ったイベント抽出の実現を目指し，質問応答を繰り返して，階層的なイベントを抽出できるようにデコーダモデルをファインチューニングする手法を提案する．実験ではベースモデルに LLaMA-3.2，データセットに GENIA2011 を用いて，性能を評価する．

1　はじめに

生物医学イベント抽出はゲノム解析や新薬の開発の補助などに幅広い応用が期待される重要なタスクである [1]．生物医学イベントは，生体内の反応経路内の反応などを含むタンパク質のリン酸化や遺伝子の発現制御などの生体内エンティティの状態変化を表す．このため，イベントの自動抽出は反応経路の整理に利用でき，生物医学研究の効率化に貢献する．イベント抽出は，イベントの発生を最も表す表現であるトリガーとイベントの種類を表すイベントタイプ，イベントの要素である引数となるエンティティもしくはイベントとその役割を特定する．

生体内ではある反応が別の反応を引き起こす複雑な反応経路があり，反応に対応する生物医学イベントの記述においてもあるイベントが別のイベントの引数となる階層的な構造が多く現れる [2]．図 1 に階層的なイベントの例を示す．*Phosphorylation* (リン酸化) タイプのイベントは *Positive_regulation (+Regulation)* (正の制御．イベントの活性化を表す) タイプのイベントの引数となる．そのため，生物医学イベント抽出では個々のイベントだけでなく，イベント間の関係も考慮した抽出が必要となる．

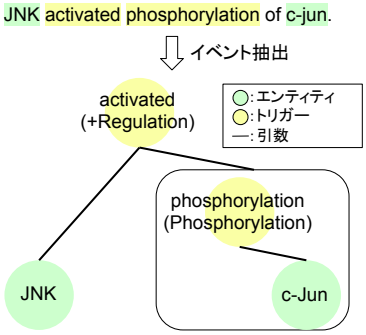

図 1　階層的なイベント抽出の例

近年，情報抽出のタスクにおいて，事前に大量のテキストで学習された Large Language Model (LLM) を用いた生成による抽出手法が注目されている [3, 4]．情報抽出手法は，BERT [5] などのエンコーダモデルによる分類ベースのアプローチが主流であったが，その後エンコーダ-デコーダモデルの登場により生成による手法が提案された．近年では，LLaMA [6] などのデコーダモデルの台頭により，(1) 事前学習で獲得した豊富な言語知識の活用，(2) 少量データでの効果的な学習，といった利点を活かした生成による手法が提案されている．一般分野での固有表現抽出や関係抽出のタスクにおいては，デコーダモデルをファインチューニングした手法が優れた性能を示すことが報告されている [7, 8]．

一方で，イベント抽出においてはエンコーダ-デコーダモデルをファインチューニングした研究 [9, 10, 11] は存在するものの，デコーダモデルは zero-shot や few-shot での利用に留まっている．特に生物医学イベントは複雑な階層構造をとるため，ファインチューニングを利用した手法が有効であると考えられるが [12, 13]，デコーダモデルをファインチューニングし，生成問題として定式化した研究は存在しない．そのため，生物医学イベント抽出タスクにおけるさらなる性能向上のためには，デコーダモデルをファインチューニングしながら活用できる手法が必要である．

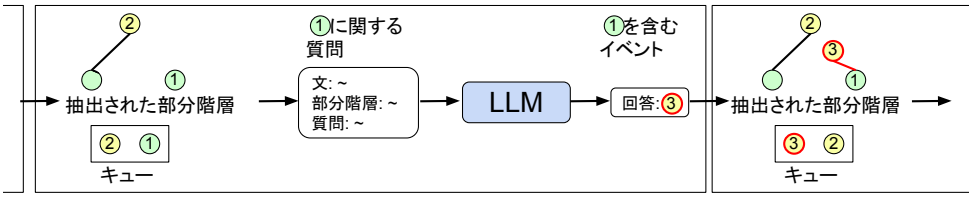

図 2　提案手法の概要図．外枠で囲った質問応答を繰り返す．

そこで，本研究では生物医学分野におけるデコーダモデルを用いた階層的な構造を持ったイベント抽出の実現を目指す．このために，質問応答を繰り返し，階層的なイベントを抽出できるようにデコーダモデルをファインチューニングする手法を提案する．本研究の貢献は次の通りである．

- 生物医学イベント抽出においてデコーダモデルを利用するためのマルチターン質問応答による手法を提案
- 生物医学分野のイベント抽出のデータセットである GENIA2011 [14] を用いて，デコーダモデルを用いた手法の有効性を確認

2　関連研究

2.1　生物医学イベント抽出

従来，生物医学イベント抽出は分類問題として定式化されてきたが [15, 16, 13]，近年では質問応答タスクとしての定式化 [12] が提案され，高い性能を示している．しかし，いずれの手法も BERT [5] などのエンコーダモデルを基盤としており，モデルサイズや事前学習データの制約から，複雑な関係を捉える言語能力に限界があった．そこで，より最近の研究では BART [17] などのエンコーダ-デコーダモデルを利用した生成による手法 [18] も提案されているが，より大規模な言語モデルであるデコーダモデルをファインチューニングする手法は未だ提案されていない．

2.2　エンコーダ-デコーダモデルをファインチューニングしたイベント抽出

ニュースなどの一般ドメインではイベント抽出を生成問題として定式化し，エンコーダ-デコーダモデルをファインチューニングする手法が提案されている [9, 10]．Hsu ら [9] や Duan ら [10] はプロンプトにイベントに関連するキーワードを入れることで，ラベルの意味的活用を促す手法を提案した．このようなモデルにヒントとなる情報を与えることで，事前学習された知識を引き出し，少ないデータでも高い性能を実現できることを示している．しかし，これらの手法は階層的な構造を考慮していないため，生物医学イベント抽出への直接的な応用は難しい．

2.3　デコーダモデルをファインチューニングした情報抽出

固有表現抽出や関係抽出では LLaMA [6] などの LLM をベースとしたデコーダモデルをファインチューニングする手法が提案されている [7, 8]．Sainz ら [8] はデコーダモデルをファインチューニングすることで，一部のデータにおいてエンコーダ-デコーダモデルより高い性能を示している．一方で，生物医学イベント抽出における階層的な構造の抽出においては，デコーダモデルをファインチューニングする手法は提案されておらず，その有効性はわかっていない．

3　提案手法

本研究では生物医学分野において，デコーダモデルを用いた階層的な構造の抽出の実現のために，マルチターン質問応答による幅優先探索に基づく階層的イベント抽出手法を提案する．提案手法はエンコーダモデルを用いた手法 [13] を拡張した手法となっている．本手法では，入力文とそれに含まれるエンティティが与えられた状態からイベント構造を抽出するタスクを対象とする．

従来手法 [13] では，階層構造に対して疑似的な深さ優先探索を利用した抽出を行っていたが，以下の二つの課題が存在した：(1) ある部分階層の探索を完了した後，新たな部分階層の探索を開始する際に，既に探索で抽出された部分階層の情報を活用できない点，(2) 深さ優先探索の性質上，複数のイベントを部分階層に持つような複雑な上位階層の構造の探索をすぐに行わなければならない点である．

これらの課題に対処するため，本研究では階層構造に対して幅優先探索に基づく抽出を行う．幅優先探索を採用することで，部分階層を持たない，もしくは，部分階層が少ない下位の階層から段階的に情報を収集し，それらの情報をプロンプトに追加する

ことで，後続の下位階層を部分階層に持つ深い上位階層のイベントの抽出に活用する．提案手法の概要を図 2 に示す．

3.1 タグ付けによる同一表層の区別

デコーダモデルを用いた生成ベースの手法において，同一文中に同一の表層を持つトリガーやエンティティのメンションが複数存在し，その表層を持つテキストが生成された場合，生成されたものがどのメンションを指しているかがわからないことが問題となる．この問題に対処するため，本研究では文を単語単位に分割し，同一の表層を持つ単語に対して，出現順を表す番号を付与する．例えば，以下の文が入力された場合：「Transient^1 tranfection^1 of^1 NB4^1 with^1 a^2 **C/EBP-epsilon^1** increased^1 cell^2 growth^1, while^1 antisense^1 **C/EBP-epsilon^2** decrease^1 growth^2.」このタグ付けにより，「**C/EBP-epsilon^1**」は「tranfection^1」のイベントの引数，「**C/EBP-epsilon^2**」は「antisense^1」のイベントの引数などのように，同じ表層の言及を明確に区別することが可能となる．

3.2 マルチターン質問応答による幅優先探索

グラフ探索の概念を応用したマルチターン質問応答により，階層的なイベント構造を効率的に抽出する．具体的には，階層的なイベント構造をグラフとみなし，幅優先探索によってイベントとその引数を順次抽出する．この手法により，下位階層のイベントを一通り特定した後に，それらを引数として持つ上位階層のイベントを特定することが可能となる．提案手法では，入力文とエンティティのみが与えられ，用意した探索対象を格納するキューにエンティティを格納してから，以下の 2 つのステップを上位階層のイベントが見つからなくなるまで繰り返す．

1. 下位階層を起点としたイベント検出：キューから取り出した入力文中の各エンティティもしくはすでに検出されたイベントに対して「このエンティティ (もしくはイベント）を引数として持つイベントはありますか？」という質問を生成し，イベントの有無を特定する．イベントが存在する場合，そのトリガーをノードとしてグラフに追加し，イベントタイプを属性として付与する．例えば，Trigger A に対して Cause B と Theme C の引数を持つイベントの場合，この段階で Trigger A がノードとして追加され，イベントタイプが属性として付与される．見つかったイベントはキューに追加され，更なる上位階層のイベントの発見に用いられる．
2. 引数の役割特定：検出されたイベントのトリガーノードと各エンティティノード（もしくはイベント）間の関係について，「このエンティティ（もしくはイベント）はイベントにおいてどのような役割を持ちますか？」という質問を生成し，Theme，Cause などの引数の役割を特定する．これにより，トリガーノードとエンティティノードもしくはイベント間にラベル付きエッジが追加される．上記の例では，Trigger A とエンティティ B 間に Cause ラベルのエッジ，Trigger A とエンティティ C 間に Theme ラベルのエッジが追加される．

3.3 プロンプト設計

入力のプロンプトには入力文，過去に抽出したイベント，質問文を与える．さらに，会話履歴と生成を制御するためにデータセットの説明をプロンプトに含める．会話履歴を含めることで，モデルが幅優先探索の過程で行った判断を参照でき，特に階層的なイベントの抽出において，下位階層のイベントの抽出過程を考慮した上位階層の探索が可能になると期待される．データセット説明の条件では，イベントタイプの階層的な分類情報の有無を比較する．具体的には，イベントタイプを以下の 3 グループに分類した説明と，各イベントタイプが取りうる引数の情報を含める条件と含めない条件を設定する：

- 基本イベント（*Gene_expression* 等）
- 複合イベント（*Binding* 等）
- 制御イベント（*Regulation* 等）

プロンプトのテンプレートの詳細を付録 C に示す．

3.4 学習と評価

学習時の訓練データは，3.2 節で示した質問それぞれに対して，個別に正解のイベント構造から適切な回答を対応付けることで作成する．入力は文章とこれまでのイベント抽出状況および質問で構成され，出力はその質問に対する正解のイベント構造から作成した回答となる．学習では，この入力・回答ペアをランダムに学習する．

評価時は入力文とエンティティの情報を与え，3.2 節の手順に従って質問応答を繰り返すことに

表 1 イベント抽出の結果 (%)．最も高いスコアは**太字**，次に高いスコアを下線 で示す

設定	階層的なイベント			全イベント		
	P	R	F	P	R	F
ベースライン	36.64	23.40	28.56	52.72	35.71	42.58
+ データセットの説明	42.06	26.16	**32.26**	58.82	40.85	**48.22**
+ 会話履歴	33.78	24.60	28.47	48.88	36.52	41.80
+ データセットの説明, 会話履歴	46.27	20.29	28.21	60.81	34.07	43.67

よって，階層的なイベント構造を抽出する．ただし，質問応答の結果からイベント構造への変換には後処理が必要となる．本研究では Wang ら [12] の手法に従い，あるトリガーに対して複数の引数が抽出された場合，イベントタイプに合わせて．引数の組み合わせから個別のイベントを生成する．例えば，Cause と Theme をそれぞれ引数にとるイベントタイプのトリガーに対して，Theme A,B と Cause C,D が抽出された場合，全ての可能な組み合わせ（Theme:A-Cause:C，Theme:A-Cause:D，Theme:B-Cause:C，Theme:B-Cause:D）に対応する 4 つのイベントを生成する．

4 実験

4.1 実験設定

評価にはヒトの血液細胞内の転写因子に関する文献から作成された GENIA2011 [14] データセットを利用した．GENIA2011 は 2 つのエンティティタイプと 9 つのイベントタイプ，6 つの引数役割から構成される．評価には，公式の評価スクリプトを利用し，適合率，再現率，F 値を報告する．なお，GENIA2011 には文書レベルのアノテーションが付与されているが，本手法では 3 節で説明した通り，文単位でのイベント抽出に焦点を当てているため，文単位のイベントについてのみ抽出を行った．そのため，文を跨ぐイベントは False negative となり，理論上の F 値の最大値は 89.91%となる．付録 A に示した通り，表層の区別が必要であることがわかったため，全ての評価においてタグの区別を行うこととした．ベースモデルには，LLaMA-3.2-1B-Instruct を採用し，付録 B に示したハイパーパラメタを利用した．

実験では，入力文，過去に抽出したイベント，質問文のみをプロンプトに含める設定をベースラインとして採用し，イベント抽出の性能に影響を与える 2 つの要因（データセット説明，会話履歴）について検証を行った．データセット説明の条件では，イベントタイプの階層的な分類情報の有無を比較した．会話履歴の条件では，これまでの質問応答履歴を含める条件と含めない条件を比較し，両条件の違いが抽出性能に与える影響を分析した．

4.2 結果

表 1 にさまざまなプロンプトの設定を変えた場合のイベント抽出の結果を示す．実験より，ベースラインにデータセットの説明を付与した条件が最も高い性能を達成した．これは，データセットの説明を付与することで，モデルがイベントタイプや引数役割の制約を適切に理解し，不適切なイベントの生成を抑制できたためと考えられる．その結果，F 値の向上につながったと推測される．一方，会話履歴を付与した条件（ベースライン+会話履歴，ベースライン+データセットの説明+会話履歴）では，いずれも F 値の低下が観察された．この結果は，会話履歴がモデルの予測にとってノイズとして作用した可能性を示唆している．

文書単位でイベント抽出をしている従来手法 [12] や最先端の手法 [19] と比較すると，F 値が 20 %以上低い性能となった．この性能差の要因として，過去の抽出結果のマークダウン形式での表現やターンを考慮しない学習方式が考えられる．そのため，今後はより，デコーダモデルにより適した抽出結果の利用方法やマルチターン形式での学習の検討が必要である．

5 おわりに

本研究では生物医学分野におけるデコーダモデルを用いた階層的な構造を持ったイベント抽出の実現を目的に，質問応答を繰り返して，階層的なイベントを抽出できるようにデコーダモデルをファインチューニングする手法を提案した．GENIA2011 データセットで評価した結果，プロンプトにデータセットの説明を付与することが性能の向上に寄与することが確認できた．今後は，デコーダモデルに有効なプロンプトの設計や学習方法を検討する．

参考文献

[1] Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. **IEEE Access**, 9:160721–160757, 2021.

[2] Ioannis Korkontzelos Sophia Ananiadou Makoto Miwa, Paul Thompson. Comparable study of event extraction in newswire and biomedical domains. In **COLING**, 2014.

[3] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: a survey. **Frontiers of Computer Science**, 18:186357, 2024. Open access.

[4] Étienne Simon, Helene Olsen, Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. Generative approaches to event extraction: Survey and outlook. In **Proceedings of the Workshop on the Future of Event Detection (FuturED)**, pages 73–86, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **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)**, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[7] Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction, 2024.

[8] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction, 2024.

[9] I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. Degree: A data-efficient generation-based event extraction model. In **ACL**, 2022.

[10] Junwen Duan, Xincheng Liao, Ying An, and Jianxin Wang. Keyee: Enhancing low-resource generative event extraction with auxiliary keyword sub-prompt. In **ACL**, 2024.

[11] Wei Wang Nanyun Peng Mingyu Derek Ma, Alexander K. Taylor. Dice: Data-efficient clinical event extraction with generative models. In **ACL**, 2023.

[12] U.Leser X.Wang, L.Weber. Biomedical event extraction as multi-turn question answering. In **ACL**, 2020.

[13] Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. Biomedical event extraction as sequence labeling. In **ACL**, 2020.

[14] Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. Overview of genia event task in bionlp shared task 2011. In **Proceedings of BioNLP Shared Task 2011 Workshop**, pages 7–15, 2011.

[15] Kung-Hsiang Huang, Michael Yang, and Nanyun Peng. Biomedical event extraction with hierarchical knowledge graphs. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pages 1277–1285, 2020.

[16] Jari Björne and Tapio Salakoski. Biomedical event extraction using convolutional neural networks and dependency parsing. In **Proceedings of the BioNLP 2018 workshop**, pages 98–108, 2018.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[18] Haohan Yuan, Siu Cheung Hui, and Haopeng Zhang. A structure-aware generative model for biomedical event extraction. **arXiv preprint arXiv:2408.06583**, 2024.

[19] Zhizheng Wang Yuanyuan Sun Hongfei Lin Jinzhong Ning, Zhihao Yang. Proceedings of the thirty-second international joint conference on artificial intelligence. In **ODEE: A One-Stage Object Detection Framework for Overlapping and Nested Event Extraction**, 2023.

A　同一表層の区別の有無の影響の評価

3.1 節で説明した同一表層を区別するタグ付けによる影響を測るために，訓練データ上で作成した正解の出力を後処理によってイベント構造へ変換した場合の正解率を評価した．表 2 に結果を示す．評価では，同一テキストの区別に関する以下の手法を比較した[1)]．結果として，同一表層の区別をしないことで，イベント抽出において 7~13%以上の F 値の低下が見られており，同一表層の区別が重要であることがわかった．

- ランダム選択手法：同一テキストを持つトリガーとエンティティが複数存在する場合に，組み合わせをランダムに選択する手法
- 最近傍選択手法：トリガーとエンティティ間の文中での距離が最も近い組み合わせを選択する手法
- タグ付け手法：提案する位置情報に基づくタグ付けにより組み合わせを一意に特定する手法

表 2　訓練データ上での同一表層の区別の影響 (%)

	ランダム	最近傍	タグ付け
F 値	76.23	82.69	89.91

B　ハイパーパラメタ

表 3 にハイパーパラメタの設定をまとめる．

表 3　ハイパーパラメタ設定

パラメータ	値
エポック数	100
学習率	5e-5
ウォームアップ	0.1
重み減衰	0.01
バッチサイズ	8
最適化手法	Adam
LoRA ランク	8
量子化	4bit

1)　4.1 節で説明した通り，文を跨ぐイベントを無視しているため，100%の復元はできない

C　プロンプトテンプレート

利用したプロンプトのテンプレートをリスト 1 に示す．質問応答については [12] のものを利用した．

リスト 1　プロンプトテンプレート

```
Given the sentence:{sentence}

Current extracted events:{extracted_result}

Event types and Arugments Guide:
 1. Basic Events:
 - Gene_expression(Theme:Protein)
 - Transcription(Theme:Protein)
 - Protein_catabolism(Theme:Protein)
 2. Complex Events:
 - Phosphorylation(Theme:Protein, Site:Entity)
 - Localization(Theme:Protein, AtLoc/ToLoc:
     ↪ Entity)
 - Binding(Theme:Protein+, Site:Entity+)
 3. Regulatory Events:
 - Regulation(Theme:Protein/Event, Cause:
     ↪ Protein/Event, Site:Entity, CSite:
     ↪ Entity)
 - Positive_regulation(Theme:Protein/Event,
     ↪ Cause:Protein/Event, Site:Entity,
     ↪ CSite:Entity)
 - Negative_regulation(Theme:Protein/Event,
     ↪ Cause:Protein/Event, Site:Entity,
     ↪ CSite:Entity)

Conversation history:{history}

Question:
What are event of {entity}?
What are argument of {entity}?

Answer:
Event trigger is {trigger}. Event type is {
    ↪ event type}.
Answer:{argument role} is {entity}.
```