大規模言語モデルによる症例報告の構造的要約

八幡早紀子 [1]　清丸寛一 [1]　Cheng Fei[1]　黒橋禎夫 [1]　佐藤寿彦 [2]　永井良三 [3]
[1] 京都大学　[2] 株式会社プレシジョン　[3] 自治医科大学
{yahata,kiyomaru,feicheng,kuro}@nlp.ist.i.kyoto-u.ac.jp
satoh@premedi.co.jp　rnagai@jichi.ac.jp

概要

われわれは医学的知見の共有を促進する基盤システムの構築を目的として、症例報告の中から重要な情報を抽出し、それらの関係を構造化するタスク、構造的要約に取り組んでいる。本稿では、大規模言語モデル（Large Language Model; LLM）が関係抽出タスクにおいて高い性能を達成していることを背景に、症例報告の構造的要約における LLM の有効性を検証する。実験の結果、ファインチューニングした LLM が既存手法と同程度の性能を達成することを確認した。

1　はじめに

医学分野は専門化・細分化が進んでおり、一人の医師が包括的な観点から診断を行うことが困難な状況にある。一方、医療現場では、経験頻度の少ない症例や他の診断に役立つ示唆を含む症例について要点をまとめた**症例報告**が各診療科に蓄積されている。症例報告の知見の共有を促進することはこの状況を改善する上で極めて重要であると考えられる。

この目的のもと、症例報告検索システム J-CaseMap[1)] が開発されている。J-CaseMap は、検索対象である各症例報告に症例報告の要点となる病態や所見の因果関係、修飾関係などを構造的に表した**構造的要約**がアノテーションされている点に特徴がある。構造的要約の例を図 1 に示す。J-CaseMap は構造的要約を活用することで、単に特定の病態や所見を含む症例を検索するだけでなく、病態や所見の因果関係や修飾関係を考慮した症例の検索を可能にしている。しかし、構造的要約の作成には医学の専門知識と仕様の理解が必要であり、人手で大量に作成することが難しい。

この問題を解決するため、関係抽出に基づき症例報告から構造的要約を自動生成する手法が提案されている [1]。本手法は、構造的要約中のエンティティと関係を対応する症例報告に発見的に対応付けることで関係抽出の弱教師データを作成し、それを用いて BERT [2] のようなエンコーダ型の事前学習モデルを訓練するものである。

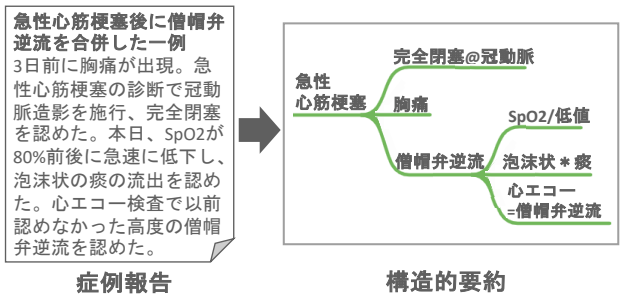

図 1　構造的要約タスク。症例報告を入力としてその構造的要約を生成する。本例は国家試験問題を元にした作例。

一方、近年の言語モデルの飛躍的な性能向上により、テキスト生成の枠組みで関係抽出を解くことが可能になってきた [3, 4]。このアプローチは、先行研究 [1] の弱教師データ構築のステップを経ず、症例報告からグラフ構造要約を直接予測することを学習できることから、構造的要約生成の性能改善に有効であると考えられる。

本稿では、大規模言語モデル（LLM）を用いた症例報告からの構造的要約生成の有効性を検証する。実験の結果、ファインチューニングした LLM が先行研究と同程度の性能を達成することを確認した。

2　症例報告の構造的要約

本研究では、症例検索システム J-CaseMap において使用されている症例報告の構造的要約の自動生成に取り組む。構造的要約の例を図 1 に例を示す。構造的要約は木構造をなす。ノードは症例報告中の主な病態や所見などのエンティティを表す。親子関係は、病態や所見同士の因果関係などのつながりを表す。ノードは内部にさらに構造を持ち、病態や所見

1）https://www.naika.or.jp/j-casemap/

のエンティティ（以下**主辞**と呼ぶ）とそれらを修飾するエンティティの関係を表現することができる。修飾関係は以下のように記述される。

- **解剖部位（記号：@）**：病態や所見の存在する部位を表す（例：「完全閉塞@冠動脈）。主辞は前方のエンティティ。
- **極性情報（記号：／）**：検査結果の高値・低値、薬剤投与の有効・無効などを表す（例：「SpO2／低値」）。主辞は前方のエンティティ。
- **検体名・検査名（記号：＝）**：所見が得られた検査項目を表す（例：「心エコー＝僧帽弁逆流」）。主辞は後方のエンティティ。
- **補足情報（記号：＊）**：部位の左右や病態の持つ特徴などを表す（例：「泡沫状＊痰」）。主辞は後方のエンティティ。

修飾関係は「MRI ＝ DWI 高信号@右＊大脳半球」のように複数組み合わせて表される場合もある。

例えば、図 1 の構造的要約は、急性心筋梗塞という病態が胸痛、冠動脈の完全閉塞および僧帽弁逆流を引き起こしたこと、さらに僧帽弁逆流が SpO2 の低値という検査結果および泡沫状の痰を引き起こしたこと、僧帽弁逆流が心エコー検査によって観察されたことを表す。

構造的要約は二つのエンティティとその関係からなる**関係三つ組**の集合として表すことができる [1]。ノード内のエンティティの修飾関係は、主辞のエンティティ、それを修飾するエンティティ、修飾関係を要素とする関係三つ組で表される。ノード間の親子関係は、親ノードの主辞のエンティティ、子ノードの主辞のエンティティ、親子関係を要素とする関係三つ組で表される。

3 関連研究

症例報告から構造的要約を自動生成する手法として、尾崎ら [1] が関係抽出に基づく手法を提案している。尾崎らの手法は、構造的要約を関係三つ組の集合とみなし、関係抽出モデルを用いて関係三つ組を予測することで構造的要約を生成する。関係抽出モデルは、情報抽出モデルとエンティティ間の関係予測モデルからなる。情報抽出モデルが症例報告からエンティティを抽出する。関係予測モデルは抽出された二つのエンティティ間の関係を予測し、親子関係や修飾関係を同定する。

関係抽出モデル [5, 6] は BERT のようなエンコーダ型の事前学習モデル [2, 7] を基盤として構築される。情報抽出モデルの訓練には、症例報告に構造的要約中のエンティティに言及しているスパン（メンション）をアノテーションしたデータが必要である。しかし、構造的要約は症例報告とは独立した形でアノテーションされており、エンティティとメンションの対応はアノテーションされていない。この問題を解決するため、尾崎らは構造的要約中のエンティティを対応する症例報告に発見的に対応付けることで、情報抽出の弱教師データ [8, 9, 10] を構築している。

尾崎らの手法は、弱教師データの品質が情報抽出モデルの性能のボトルネックとなる点に課題がある。また、関係抽出モデルがエンティティ間の関係を独立に判定するため、抽出される関係三つ組間の一貫性を陽に考慮できないという問題もある。

一方、近年の言語モデルの飛躍的な性能向上により、テキスト生成の枠組みで関係抽出タスクを高精度に解くことが可能になりつつある [11, 12, 3, 4]。テキスト生成による関係抽出では、テキストを入力として、その中で言及されている関係三つ組を特定のフォーマットで生成する。この方法では、エンティティが入力テキスト中のどのスパンに対応しているかのアノテーションは必ずしも必要ない。また、予測した関係三つ組を文脈として参照できるため、関係三つ組間の一貫性も考慮できる。加えて、大規模言語モデル（LLM）になると医学的知識をかなりの程度保持しており、医師免許国家試験に合格するほどの能力を持つことが示されている [13]。こうした背景から、LLM を症例報告の構造的要約の生成に適用することは有望であると考えられる。

4 提案手法

本研究では、LLM を利用した構造的要約生成モデルを提案する。本稿ではこれを**生成モデル**と呼ぶ。

生成モデルの学習の概要を図 2 に示す。生成モデルは入力された症例報告から構造的要約を直接出力することを学習する。LLM はテキストを入出力とするため、LLM によって構造的要約を生成する際は構造的要約を何らかの形式のテキスト表現に変換する必要がある。

本研究では、構造的要約を図 2 のように変換した。この表現では、一行に一つのノードが記述され、字下げによって親子関係が表現される。近年の LLM は事前学習データには一般に Python 等のプロ

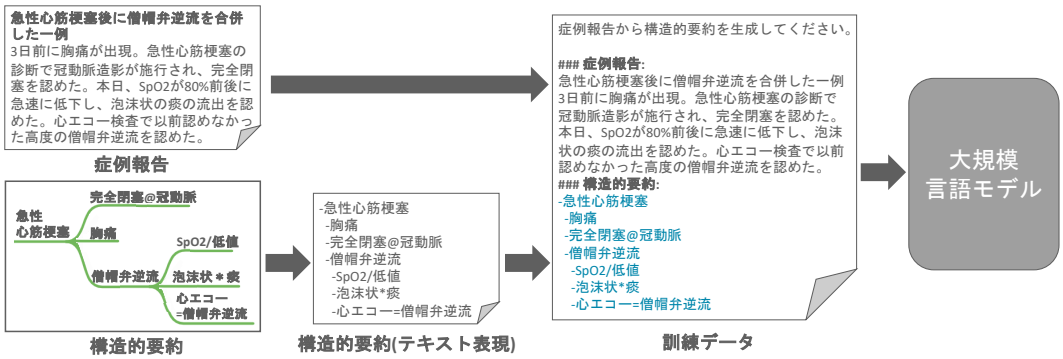

図 2　学習のワークフロー。構造的要約をテキスト表現に変換し、症例報告と構造的要約のペアから訓練データを作成する。作成した訓練データを利用し、LLM をファインチューニングする。訓練では訓練データの黒字部分を入力として青字部分を予測することを学習する。

グラムコードが含まれ、字下げによって論理構造を表すことは LLM にとって自然な形式の一つであると考えられる。構造的要約のテキスト表現の違いが性能に与える影響については今後検証を行う予定である。

5　実験

LLM の構造的要約の生成における有効性を検証するため実験を行った。

5.1　実装の詳細

症例報告と構造的要約のペアを用いてモデルのファインチューニングを行った。ファインチューニングでは LoRA [14] を使用した。訓練データとして 15,100 件の症例報告と構造的要約のペアを利用した。15,100 件の症例報告は、訓練データ 14,400 件, 開発データ 200 件, テストデータ 500 件に分割して使用した。事前学習済みの日本語 LLM として、Youri-7B-instruction[2], Swallow-13B-instruction[3], LLM-jp-13B-instruction[4] の三つを検証した。

また比較のため、尾崎らの手法 [1] も検証した。本稿ではこのモデルを**関係抽出モデル**と呼ぶ。実験では基盤モデルとして DeBERTa [15] を使用し、関係抽出モデルを学習した。

5.2　評価指標

尾崎ら [1] にならい、出力の構造的要約を関係三つ組の集合に分解し、正解の正解三つ組の集合と比較することで評価を行った。出力と正解の間でエンティティのペアとその関係性が全て一致している関係三つ組が存在する場合を正解とし、Precision、Recall、F 値を算出した。

エンティティの同一判定では、表記揺れおよび誤字脱字などを一部許容した。まず、同義語辞書によってエンティティの代表語への変換を行った。次に、出力エンティティと正解エンティティの間の編集距離を正解の文字数で割ったものを閾値と比較し、閾値以下であれば同一であるとみなした。本実験では、経験的に閾値を 0.5 とした。ただし、修飾関係のうち極性情報は語彙が閉じているため、表記揺れを許容せず、完全一致の場合のみ同一と見なした。また、同じ親を持ち子を持たない同一のノードが複数回出力された場合は、一つに統合した上で評価を行った。

5.3　結果

表 1 に結果を示す。生成モデルの中で、最も高いスコアを達成したモデルは LLM-jp-13B-instruction であった。パラメータ数が 7B のモデルと 13B のモデルを比較すると、パラメータ数のより大きいモデルの方が高いスコアを達成していた。このことから、さらに大規模な LLM をファインチューニングすることでスコアの向上が見込めると考えられる。関係抽出モデルと比較すると、生成モデルは Recall においてわずかに優れているものの、Precision と F 値はやや下回る結果となった。

5.4　分析

生成モデルは原理的に構造的要約としてパースできない不適当な出力を生成する可能性がある。し

2)　https://huggingface.co/rinna/youri-7b-instruction
3)　https://huggingface.co/tokyotech-llm/Swallow-13b-instruct-hf
4)　https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0

表 1　実験結果。親子関係は病態や所見の因果関係を表す。修飾関係は、エンティティの解剖部位、極性情報、検体名・検査名、補足情報などの関係を表す。**太字**は最高スコアを示す。P,R,F1 はそれぞれ Precision,Recall,F 値を表す。

		親子関係			修飾関係			全体		
		P	R	F1	P	R	F1	P	R	F1
関係抽出モデル	DeBERTa	**0.528**	0.440	**0.480**	**0.554**	0.491	**0.520**	**0.570**	0.489	**0.526**
生成モデル	Youri-7b-instruction	0.382	0.374	0.378	0.440	0.421	0.430	0.439	0.424	0.431
	Swallow-13b-instruction	0.434	0.441	0.437	0.495	0.493	0.494	0.484	0.486	0.485
	LLM-jp-13B-instruction	0.444	**0.443**	0.444	0.512	**0.517**	0.514	0.499	**0.499**	0.499

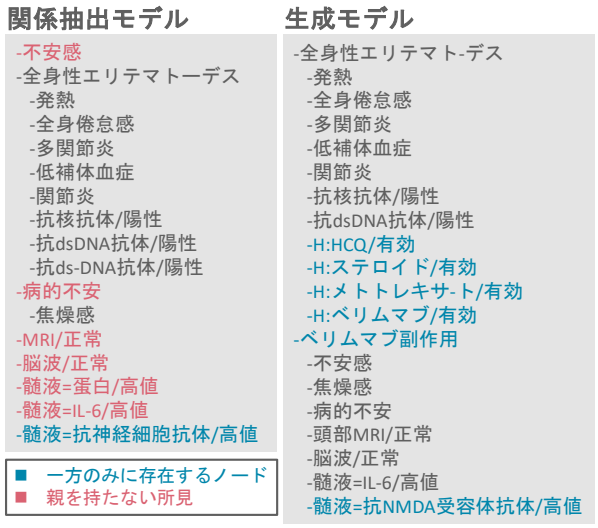

図 3　国家試験問題を元にした作例から生成された構造的要約の例。

かし、本実験において生成モデルの出力にそうしたフォーマットのエラーは確認されなかった。実験に使用した LLM は Python 等のコードを学習していることで字下げの深さを利用した構造的要約のテキスト表現を扱うことができ、また修飾関係を示す記号の意味も適切に学習できていると考えられる。

生成モデルおよび関係抽出モデルによる構造的要約の例を図 3 に示す。関係抽出モデルと比較して、生成モデルは出力したエンティティが多く、情報抽出の抜けが少なかった。また、関係抽出モデルよりも親を持たないエンティティが少なく、関係抽出モデルはより多くの関係を予測していた。これは、Precision の差が大きく Recall の差が小さいという実験結果に合致する。

生成モデルに特徴的な誤りとして、同一エンティティが繰り返し生成されるケースが確認された。同じ文字列が繰り返し生成されるのはテキスト生成の典型的な誤りであり、これはベースの LLM 自体の改善により一定程度緩和されると考えられる。

また、生成モデルの出力の一部に、症例報告中に存在しないエンティティを生成してしまうハルシネーションの問題が確認された。関係抽出モデルは原理的にこうしたハルシネーションは発生しない。今後、ベースとなる LLM 自体の改善や医学分野のテキストを用いた LLM の追加学習等により、この問題がどの程度緩和されるか検証する予定である。

5.5　医師による評価

生成モデル・関係抽出モデルの出力からランダムに選出された 10 件程度に対して著者の医師が評価を行った。結果、生成モデルは症例の主要部分の抽出抜けや親子関係のエラーが少なく、関係抽出モデルよりも人間の医師が作成したものに近い構造的要約を生成していることが確認された。

この結果は、自動評価の結果と合致していない。この原因は、現在の評価手法ではすべての関係三つ組を同等に扱っている一方、人間の医師は主要な病態の抽出とそれらの親子関係を重視している点にある。主要な病態に着目した、人間の医師の評価と整合する自動評価手法を開発することは今後の課題としたい。

6　結論

本稿では LLM を用いた症例報告の構造的要約生成の有効性を検証した。症例報告から構造的要約を生成するよう LLM をファインチューニングし、既存の関係抽出モデルに基づく手法と同等程度の性能が達成されることを確認した。実験では LLM のパラメータ数が性能に大きく影響することが示唆された。また、今回の実験で使った事前学習済みモデルは全て汎用的なドメインのモデルであった。医学分野のテキストを用いてドメイン適応することは性能の向上に有効であると考えられる。今後はより大規模な LLM の適用と LLM の医学分野へのドメイン適用について検証を進める予定である。

7 謝辞

本研究は戦略的イノベーション創造プログラム(SIP)「統合型ヘルスケアシステムの構築」JPJ012425の補助を受けて行った。

参考文献

[1] 尾崎立一, 清丸寛一, Cheng FEI, 黒橋禎夫, 永井良三, 佐藤寿彦. 弱教師学習に基づく症例報告の構造的要約. 第 26 回日本医療情報学会春季学術大会, 2022.

[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] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. 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. 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics.

[4] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. GPT-RE: In-context learning for relation extraction using large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3534–3547, Singapore, December 2023. Association for Computational Linguistics.

[5] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 50–61, Online, June 2021. Association for Computational Linguistics.

[6] Zhen Wan, Qianying Liu, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Jiwei Li. Rescue implicit and long-tail cases: Nearest neighbor relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 1731–1738, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[8] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, **Proceedings of the Joint Conference 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.

[9] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 3661–3672, Online, November 2020. Association for Computational Linguistics.

[10] Zhen Wan, Fei Cheng, Qianying Liu, Zhuoyuan Mao, Haiyue Song, and Sadao Kurohashi. Relation extraction with weighted contrastive pre-training on distant supervision. In Andreas Vlachos and Isabelle Augenstein, editors, **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 2580–2585, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[11] Daojian Zeng, Haoran Zhang, and Qianying Liu. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 34, pp. 9507–9514, 2020.

[12] Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 236–246, Online, November 2020. Association for Computational Linguistics.

[13] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations, 2023.

[14] 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.

[15] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing. In **The Eleventh International Conference on Learning Representations**, 2023.