GPT-4 による診療文書からのオントロジー自動構築の初期検討

小林 和馬 [1,2] 山本 和英 [3] 浜本 隆二 [1,2]
[1] 国立がん研究センター研究所 [2] 理化学研究所革新知能統合研究センター [3] 言語商会
kazumkob@ncc.go.jp, yamamoto.kazuhide@gmail.com, rhamamot@ncc.go.jp

概要

読影レポートからなるテキスト・コーパスを用いたオントロジーの自動構築を試みた。まず、オントロジーは医学的エンティティが属する三段階の意味的階層からなると定義し、これを出現形、標準表記、ラベルとした。続いて、各階層のスキーマをデータ駆動的に発見する過程を、ラベル発見タスクと標準表記発見タスクに分解し、GPT-4 による Proposal-Synthesis 型の段階的推論を行った。これにより構築したオントロジー・スキーマに基づき、それぞれの医学的エンティティに対する識別、標準化、ラベリングの過程からなるマッピングを行い、オントロジーを自動構築し、その妥当性を検証した。

1 はじめに

大規模言語モデル (LLM: Large language model) は、汎用性の高い推論能力を示す一方で、内部動作が Black-box であり、生成する情報の事実性を保証しきれないという課題がある [1]。医学のような専門分野では、最新の知識を正確に運用することが要請されるため、LLM が生成する情報の信頼性や正確性を担保し、誤った情報が個人や社会に対して深刻な被害を引き起こさないようにするための技術的アプローチが求められている。

こうした技術的アプローチの 1 つに、**外部知識ベース**の活用が挙げられる。外部知識ベースとは、オントロジーや知識グラフのような構造化情報や、ドキュメント・データベースなどの非構造化情報の集合を指し、対象ドメインにおける概念やその関係性が格納されたものをいう。例えば、検索拡張生成 (RAG: Retrieval Augmented Generation) と呼ばれるフレームワークでは、LLM が未学習であった知識を必要とする場合に、外部知識ベースにアクセスすることにより参照情報を取得し、これを元に回答を生成することができる [2]。すなわち、外部知識ベースを補助的に活用することで、LLM が生成する情報の事実性を保証できる可能性がある。

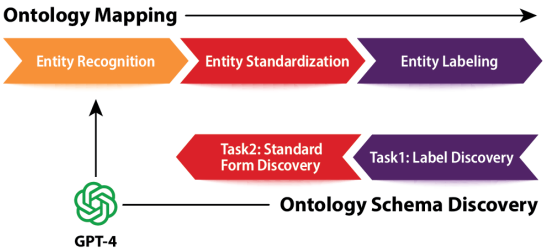

図 1 オントロジー・スキーマをデータ駆動的に発見する過程 (Ontology Schema Discovery) を、ラベル発見タスク (Label Discovery) と標準表記発見タスク (Standard Form Discovery) に分け、GPT-4 にて推論した。次に、構築したオントロジー・スキーマに対して、医学的エンティティの識別 (Entity Recognition)、標準化 (Entity Standardization)、ラベリング (Entity Labeling) からなるマッピング・タスク (Ontology Mapping) を行い、オントロジーを自動構築した。

本稿では、外部知識ベースのうち、オントロジーや知識グラフのような**構造化情報**に注目する。構造化された外部知識ベースは知識の追加や削除、更新を行うことが容易である。更に、LLM とアーキテクチャ・レベルで統合することで [3]、推論時に依拠する知識を White-box 化できる可能性も示唆される。一方で、大規模なテキスト・コーパスから構造化情報を抽出し、これを外部知識ベースとして構築するためには、エキスパートによる手作業が必要となり、大きなコストを要することが課題であった。

そこで、本研究では、医療分野における大規模なテキスト・コーパスより、GPT-4[4] を用いて構造化情報を自動的に抽出するための初期検討を行う。具体的には、胸部 X 線写真に関する読影レポートからなるコーパスを用いて、医学的に意味を持つ語や句（以下、医学的エンティティ）の階層的関係を表現する**オントロジーの自動構築**を目指す（**図 1** 参照）。

2 関連研究

オントロジーや知識グラフのような構造化情報と LLM を組み合わせる研究は、2 つの主要な視点に分

けられる [5]。1 つ目は、構造化情報をどのようにして LLM の推論プロセスに統合し、活用するかという視点であり、その代表的なアプローチの 1 つが RAG となる。2 つ目は、テキスト・コーパスから構造化情報をデータ駆動的に構築する過程に対して、どのように LLM を活用するかという視点である。本研究では後者のうち、オントロジーの自動構築を目的とした LLM の活用に焦点を当てる。

2.1 LLM による構造化情報の自動構築

語彙意味論的なパターンに基づいて、WordNet のような既存のオントロジーを自動的に拡張する研究が従来より多くなされてきた [6, 7, 8]。一方、LLM の言語理解能力をオントロジーの自動構築に応用した研究は限られている。Giglou らは、オントロジーの自動構築に関する過程を、いくつかのサブタスクに分解し、既存のオントロジーから得たプリミティブな語彙エンティティ対してプロンプト・テンプレートを当てはめることで、LLM が用語のタイピングや関係性を識別できるかを検討した [9]。結果、LLM はオントロジーの自動構築に一定の性能を示すものの、高度なドメイン知識が必要な場合には、エキスパートによる補助が必要であるとされた。

2.2 プロンプト・エンジニアリング

LLM に所望の情報を生成させるためには、プロンプト・エンジニアリングが重要となる。プロンプト・エンジニアリングとは、LLM に与える指示や命令を最適に設計することを指す [10]。特に、高度で複雑なタスクに対しては、それを一連のサブタスクに分解した上で、LLM に対して step-by-step による段階的な推論を促すことが有効である [11]。プロンプト・エンジニアリングには様々な経験則が報告されており、タスクを直線的に分解する Chain-of-Thought[12]、複数の生成内容から最適な回答を探索する Tree-of-Thought[13]、生成内容に対する自己フィードバックを与える Self-Refine[14] などが含まれ、固有表現抽出などの自然言語処理タスクにおいても有用性が見出されている [15]。

2.3 医療分野のオントロジー

医学領域における従来のオントロジーとして、Disease Ontology (DO)[16] 等が知られているが、人手によるキュレーションを前提としているものが大部分であった。

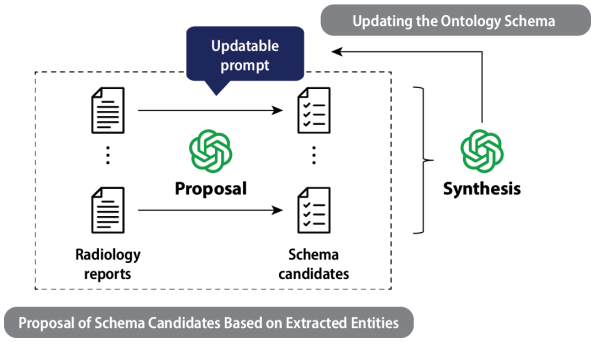

図 2 Proposal-Synthesis 型の段階的推論は、個々のサンプルから医学的エンティティを抽出し、これをマッピングするべき上位階層（標準表記及びラベル）のスキーマの候補を提案する Proposal の過程と、その結果が複数のサンプル間で整合するように統合的に解析し、新たなオントロジー・スキーマを出力する Synthesis の過程からなる。特に、Synthesis の過程で得られた新たなオントロジー・スキーマが、次の Proposal の過程におけるプロンプトに補助情報として渡されることで、データセットにおける反復過程を通してオントロジー・スキーマを逐次的に更新することができる。

3 実験と結果

急速に進歩発展する医学知識を体系化するためにも、テキスト・コーパスを用いたオントロジーの自動構築は有用である。そこで、プロンプト・エンジニアリングを活用することで、実際の診療文書からオントロジーを自動構築できるかの初期検討を行った。

3.1 データセットの構築

公開データセットである MIMIC-CXR[17] を取得した（**図 A.1** 参照）。MIMIC-CXR には、胸部 X 線写真の画像情報と、その所見が英語で記載された 20 万件以上の読影レポートが含まれている。ランダムに 100 サンプルの読影レポートを選択し、データセットを構築した。

3.2 オントロジーの定義

胸部 X 線写真の所見が記載された読影レポートには、患者に関する背景情報（患者の年齢、性別等）、臨床情報（患者の症状等）、撮影技術の詳細（正面、側面撮影等）、所見（画像情報に認められた診断の参考となる特徴等）、インプレッション（臨床情報や所見等に基づく診断の推定）、推奨事項（追加の CT 検査等）等が一般的に記載される。

本研究では、こうした各種の情報を臨床的に意味のあるものとして構成し、医学的に定義可能である

語や句を**医学的エンティティ**と呼ぶ。これは、固有表現とほぼ同義である。そして、個々の医学的エンティティは、**出現形**、**標準表記**、**ラベル**の 3 つの階層からなるオントロジーにマッピング可能であると定義する。すなわち、出現形とは、読影レポートに実際に記載された語や句であり、表記ゆれを伴う。標準表記とは、出現形の表記ゆれを同一の医学的概念を指し示すように標準化した表記である。ラベルとは、読影レポートに含まれる情報の種類（患者に関する背景情報や、臨床情報、撮影技術の詳細等）に応じて区分され、複数の標準表記を包含する。

3.3 オントロジーの自動構築

本研究では、オントロジーの自動構築を 2 つの過程に大別する（**図 1** 参照）。すなわち、オントロジーの各階層のスキーマ（オントロジー・スキーマ）をデータ駆動的に発見する過程と、構築したオントロジー・スキーマに対して、医学的エンティティを抽出してマッピングする過程である。より詳細には、前者の過程をラベル発見タスク (Task1) と標準表記発見タスク (Task2) に分解し、GPT-4 による Proposal-Synthesis 型の段階的推論を行う（**図 2** 参照）。これにより構築したオントロジー・スキーマに基づき、それぞれの医学的エンティティに対する識別、標準化、ラベリングからなるマッピングを行うことで、オントロジーを自動構築できる。

3.3.1 Proposal-Synthesis 型の段階的推論

読影レポートなどの診療文書には、個々のサンプルの記載内容は多様である一方で、背後に一定の知識体系が存在することが多い。そのため、個々のサンプルから医学的エンティティを抽出し、これをマッピングするべき上位階層（標準表記及びラベル）のスキーマの候補を提案する **Proposal** の過程と、その結果が複数のサンプル間で整合するように統合的に解析し、新たなオントロジー・スキーマを出力する **Synthesis** の過程からなる Proposal-Synthesis 型の段階的推論を考案した（**図 2** 参照）。

この推論の枠組みを、データセットにおける i 番目の反復過程として説明する。まず、K 個のサンプルからなるバッチが取得される。そして、それぞれのサンプルの読影レポートから医学的エンティティを抽出し、これをマッピングするべき上位階層（標準表記及びラベル）のスキーマの候補を提案する Proposal が行われる。このとき、i-1 番目までの反復過程で構築されたオントロジー・スキーマを補助情報としてプロンプトに入力することで、それまでの反復過程の結果から大きく逸脱しない提案となるようにする。ここまでの過程はサンプルごとに独立して処理される。続いて、それぞれの Proposal の結果がサンプル間で整合するように統合的に解析し、新たなオントロジー・スキーマを出力する Synthesis の過程が行われる。この結果が、次の i+1 番目の反復過程にて Proposal の補助情報として用いられる。以上の反復過程を通して、最終的にデータセットに含まれる全てのサンプルに基づくオントロジー・スキーマを構築することができる。

こうした反復過程を安定的かつ効率的に実行するために、読影レポートのサンプル単位での埋め込み表現に基づいた類似度計算を行い、各バッチ内での多様性を担保しつつ、連続する反復過程の間では文書間の類似性が維持されるようなサンプリングを行った（**Appendix B 参照**）。

3.3.2 Task-1: ラベル発見タスク

オントロジー・スキーマをデータ駆動的に発見する過程では、上位階層であるラベル発見タスクより取り組んだ。Proposal-Synthesis 型の段階的推論を行う際に、初回 (i=1) の反復過程においては、補助情報として参照するべき既存のオントロジー・スキーマが与えられない。そこで、ラベル発見タスクにおいては、読影レポートの一般的性質を GPT-4 に伝えることによって、オントロジー・スキーマの初期値を得た（具体的には、Anatomical Terms、Radiological Findings、Patient Information、Technical Details、Device Information、Diagnosis の 6 種類のラベル）。以後、**表 C.1** に掲載するプロンプトを用いて、ラベル階層のオントロジー・スキーマの逐次的な更新を行った。

3.3.3 Task-2: 標準表記発見タスク

ラベル階層のオントロジー・スキーマを固定した後、同様にデータセットを用いて標準表記発見タスクを行った。比較的に種類が限られることが期待されていたラベルとは異なり、標準表記の数は膨大となり得る。そのため、標準表記発見タスクにおいては、オントロジー・スキーマの初期値を与えることなく、完全にデータ駆動的に推論した（プロンプトは**表 C.1** を参照）。尚、この推論の過程において、それぞれの医学的エンティティに対して、対応する標準表記及びラベル階層が得られる。この関係性

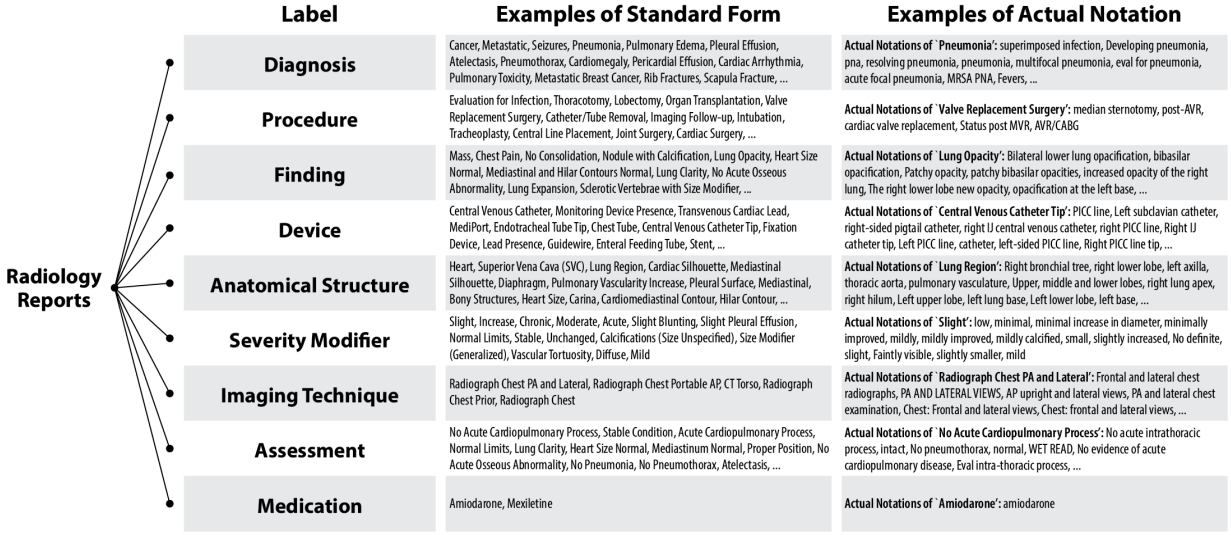

図 3　自動構築されたオントロジー。合計 9 種類のラベル、226 種類の標準表記、931 種類の出現形を含んでいた。紙面の関係から、標準表記及び出現形については一部を例示している。

は、それぞれの医学的エンティティに対する識別、標準化、ラベリングからなるマッピングに対応するため、結果として自動構築されたオントロジーを得た。

3.4　自動構築されたオントロジーの評価

自動構築されたオントロジーは、合計 9 種類のラベル、226 種類の標準表記、931 種類の出現形を含んでいた（**図 3** 参照）。このうち、ラベル階層のオントロジー・スキーマは概ね妥当であるが、検査目的や患者の症状等を含む ‘臨床情報’ に相当するラベルがあっても良かったと専門医により評価された（例: Chest Pain という標準表記が Finding ラベルの 1 つとして得られたが、これは胸部 X 線写真の画像情報のみからは読み取れない患者の臨床情報に相当する）。続いて、以下の基準に基づくエラー分析を行った。

- **標準表記の不統一:** 同一ラベルに所属する標準表記階層において、同じ医学的概念が複数の標準表記に分かれてしまっている場合（例: Diagnosis というラベルに所属する標準表記階層に、Pneumothorax と Loculated Pneumothorax という 2 つの標準表記が得られたが、これは ‘気胸’ を表現する 1 つの標準表記にまとめられるべきであった）。
- **不適切な標準表記:** 標準表記が不要な修飾語等を含み、十分に標準化されていない場合（例: Persistent Cough という咳嗽の 1 つの具体的な態様を現す標準表記ではなく、‘咳嗽’ に相当するより一般的な標準表記が定義されるべきであった）。
- **誤分類:** ラベル–標準表記間、標準表記–出現形間において、実際の割り当てよりも適切な上位階層が存在するべき場合（例: Fevers という出現形に対して Pneumonia を標準表記と割り当てたが、これは ‘発熱’ に相当する標準表記を定義し、割り当てるべきであった）。

専門医による評価の結果、標準表記階層において、標準表記の不統一は 11.9%、不適切な標準表記は 4.4%、誤分類は 7.5%、出現形階層において誤分類は 11.7% であった。

4　結論

本研究はあくまでも小規模な初期検討に留まったが、GPT-4 による Proposal-Synthesis 型の段階的推論によって、臨床的観点からも一定の妥当性を有する形で医学的エンティティの意味的階層を構築することが出来た。今後の課題として、より大規模なデータセットへの適用、医学的エンティティの間の関係性の抽出、Proposal-Synthesis 型の段階的推論の有用性に関する比較実験が必要である。また、本研究では、自動構築されたオントロジーの臨床的評価に留まり、読影レポートに含まれる医学的エンティティに対する識別性能の評価を行っていない点に留意されたい。更に、本研究では Zero-shot による推論を行ったが、エキスパートによる教示情報を活用することで、人間が理解する知識体系によりアライメントされた結果を得られる可能性もある。

謝辞

本研究は、JST ACT-X(JPMJAX23C9)、JSPS 科研費 (JP22K07681)、国立がん研究センター研究開発費 (2023-A-19) の支援を受けて行った。

参考文献

[1] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. **ACM Comput. Surv.**, Vol. 55, No. 12, 2023.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In **Neural Information Processing Systems (NeurIPS)**, 2020.

[3] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep Bidirectional Language-Knowledge Graph Pretraining. In **Neural Information Processing Systems (NeurIPS)**, 2022.

[4] OpenAI. GPT-4 Technical Report. **arXiv: 2303.08774**, 2023.

[5] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying Large Language Models and Knowledge Graphs: A Roadmap. **arXiv: 2306.08302**, 2023.

[6] George A. Miller. WordNet: A Lexical Database for English. **Commun. ACM**, Vol. 38, No. 11, 1995.

[7] Chung Hee Hwang. Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for Representing and Retrieving Information. In **Knowledge Representation Meets Databases**, 1999.

[8] Dan I. Moldovan and Roxana Girju. An Interactive Tool for the Rapid Development of Knowledge Bases. **Int. J. Artif. Intell. Tools**, Vol. 10, , 2001.

[9] Hamed Babaei Giglou, JenniferD' Souza, Sören Auer. LLMs4OL: Large Language Models for Ontology Learning. In **International Semantic Web Conference**, p. 408–427, 2023.

[10] 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 **Neural Information Processing Systems (NeurIPS)**, 2020.

[11] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In **Neural Information Processing Systems (NeurIPS)**, 2022.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In **Neural Information Processing Systems (NeurIPS)**, 2022.

[13] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. **arXiv: 2305.10601**, 2023.

[14] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. **arXiv: 2303.17651**, 2023.

[15] Mingchen Li and Rui Zhang. How far is Language Model from 100% Few-shot Named Entity Recognition in Medical Domain. **arXiv: 2307.00186**, 2023.

[16] J Allen Baron, Claudia Sanchez-Beato Johnson, Michael A Schor, Dustin Olley, Lance Nickel, Victor Felix, James B Munro, Susan M Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Rima Koka, Carol Greene, and Lynn M Schriml. The DO-KB Knowledgebase: a 20-year journey developing the disease open science ecosystem. **Nucleic Acids Res.**, Vol. 52, No. D1, pp. D1305–D1314, 11 2023.

[17] Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR Database (version 2.0.0). **PhysioNet**, 2019.

表 C.1　オントロジーの自動構築で使用したプロンプト。

Label Discovery	Prompt
Initial Categories	You are a top researcher in natural language processing in the medical field. You conduct linguistic evaluations of radiological reports created by radiologists for chest X-rays. Extracts named entities and their categories from the radiological reports of chest X-rays. What kinds of named entities and their categories are expected in this context?
System	As a specialist in natural language processing within the medical field, your main responsibility is to analyze chest X-ray radiological reports prepared by radiologists.
Proposal	Your task is to identify and categorize named entities within the following radiology reports. Please proceed with the extraction of these entities and their respective categories from the provided chest X-ray radiological reports, especially focusing on the content that semantically corresponds to FINDINGS and IMPRESSION. When classifying categories, please base them on the given categories.
Synthesis	Your task is to integrate and standardize the categorization of named entities extracted from various radiology reports. Using the list of named entities and their categories in JSON format provided below, ensure that similar entities are uniformly categorized. Please propose any modifications to enhance the accuracy of categorization. Then, based on your earlier suggestions, please now revise each JSON file to reflect the updated categorization.
Standard Form Discovery	**Prompt**
System	As a specialist in natural language processing within the medical field, your main responsibility is to analyze chest X-ray radiological reports prepared by radiologists.
Proposal	The following is a dictionary provided with the category names as keys, and lists of medical entities included in each category. There are variations in terminology for the same medical entity across different categories. Please identify the unique entities in each category, standardize their terminology. When creating standardized terms, group together those that represent the same medical concept, and consider abstracting specific numbers and other details into generalized units. Please avoid creating categories like 'Others' as much as possible, and standardize each medical entity according to its medical meaning. While performing this standardization, refer to the following ontology as a guide. If you encounter any entities that do not align with the existing standard terms, please create a new category for them. Note that updating existing categories is not an option.
Synthesis	Following the recommended standardization guidelines, please standardize these terms for consistency and organize them into a structured format: category, standardized terms, and their actual notations.

A　読影レポートの例

読影レポートの例を**図 A.1** に示す。

FINAL REPORT

INDICATION: ___F with myasthenia ___, s/p fall with SOB and left sided back pain. // rib fx, pneumonia

TECHNIQUE: Single portable view of the chest.

COMPARISON: ___.

FINDINGS:

Bibasilar opacities silhouetting the hemidiaphragms again seen, suggestive of persistent pleural effusions and likely adjacent atelectasis. Please note that infection would be difficult to exclude. There is persistent pulmonary vascular congestion. Dual lead pacing device is again noted. No displaced fractures identified.

IMPRESSION:

Bibasilar opacities likely due to combination of persistent effusions and adjacent atelectasis noting that infection cannot be excluded. No evidence of displaced rib fracture on this portable chest x-ray.

図 A.1　読影レポートの例。胸部 X 線写真に関する適応 (INDICATION)、撮像方法 (TECHNIQUE)、ベースラインとなる過去検査との比較 (COMPARISON)、所見 (FINDINGS)、インプレッション (IMPRESSION) が記載されている。尚、患者の個人情報に関わる記載は削除されている。

B　文書サンプリングの方法

データセットのサイズ（N 個）をバッチに含まれるサンプル数（K 個）で割った商をサンプル単位間隔（L 個）とする。OpenAI 社の文書埋め込みモデル (text-embedding-ada-002) を用いて、データセットに含まれる読影レポートの埋め込みベクトルを得た。続いて、ランダムに選択した 1 つのサンプル（開始サンプル）に対して、残りの全ての読影レポートとのコサイン類似度を計算し、類似度順にソートした。その後、開始サンプルを起点にサンプル単位間隔を L×k (k=0, ..., K-1) 整数倍した位置に存在する合計 K 個のサンプルをサンプリングし、これを初回の反復過程に用いた。以後、i 番目の反復過程においては、開始サンプルから i 個離れたサンプルを起点として同様にサンプリングし、これを合計 N/K 回繰り返した。

C　プロンプト

オントロジーの自動構築で用いたプロンプトを**表 C.1** に示す。Proposal-Synthesis 型の段階的推論の実装に当たっては、OpenAI 社の対話モデル (gpt-4-1106-preview) を用いた。