医学生物学文献からのオントロジー構築のためのメンション非依存型情報抽出

西田典起 [1*] Oumaima El Khettari[1,2*] Shanshan Liu[1] Rumana Ferdous Munne[1]
山縣友紀 [1] 古崎晃司 [3] Solen Quiniou[2] Samuel Chaffron[2] 松本裕治 [1]
[1] 理化学研究所 [2]Nantes Université - LS2N [3] 大阪電気通信大学
noriki.nishida@riken.jp

概要

生物学的プロセスなどのエンティティは、医学生物学文献内で常に明示的に言及されているとは限らず、しばしば専門知識の共有を仮定して暗黙的に言及される。本研究では、医学生物学文献からのオントロジー自動構築を目指し、プロセス抽出と文書レベル関係抽出の二段階からなるメンション非依存型情報抽出システムを提案する。提案システムでは、一貫してメンション表現に依存せず、入力文書に対して対象オントロジーに紐づいたプロセスエンティティ集合の抽出とエンティティ間関係の同定を可能にする。新たなベンチマークデータセットを構築し、BERT ベースまたは LLM ベースのメンション非依存型アプローチを提案し、評価と比較を行う。

1 はじめに

医学生物学の領域では、疾病メカニズムの解明が治療や予防に大きく寄与する。しかし、その情報は多くの論文の中に散在し、多様な語彙で記述されるため、それらを体系化する方法が必要である。オントロジーは、エンティティとその関係性に基づいて知識を構造化するための枠組みである。Homeostasis Imbalance Process Ontology (HOIP) では、細胞老化や COVID-19 の感染メカニズムといったプロセスに関する知識を構造化するため、手作業による注釈が行われてきた [1, 2]。しかし、手作業での注釈は、高コストで時間がかかるという大きな課題がある。オントロジーの網羅性と質を向上させるために、情報抽出技術による自動注釈アプローチの開発が期待されている。

一方で、従来の情報抽出では、エンティティがテキスト内で**明示的**に表現されることを仮定している (これらのテキスト表現をメンションと呼ぶ) [3, 4, 5]。メンションは、エンティティがテキスト内でどのように記述されているかを直接的に示すため、エンティティと関係性の抽出において非常に強力な指標となる。しかし、現実世界の医学文献では、重要なエンティティはしばしば医学生物学的な背景知識の共有を仮定に**暗黙的**に語られる。

そこで本研究では、HOIP 等のオントロジーの完全自動注釈を目指し、二段階からなるメンション非依存型情報抽出システムを提案する。図 1 に、提案システムの概要と、暗黙的に言及されるエンティティの具体例を示す。入力として与えられた文書に対して、第一段階の「プロセス抽出 (Process Identification; PI)」では、その文書に記述されている、あるいは専門知識に基づいて推測可能なエンティティ (生物学的プロセス) の集合を抽出する[1]。エンティティは、オントロジー内の一意の ID として表現される。第一段階で抽出されたエンティティ集合は、第二段階の「文書レベル関係抽出 (Document-level Relation Extraction; DocRE)」 [5, 6, 7] に渡され、エンティティ間の関係性が同定される。情報抽出システム全体の出力は、(主語エンティティ ID, 関係ラベル, 目的語エンティティ ID) の形式の三つ組として表される。本研究では、PI と DocRE の両タスクにおいて、メンション非依存な複数のアプローチを提案し、評価する。具体的には、両タスクにおいて、BERT [8] ベースの教師あり識別アプローチと、大規模言語モデル (LLM) と In-Context Learning (ICL) [9, 10, 11] に基づく生成アプローチを提案する。さらに、メンション非依存な情報抽出シ

＊ Equal contribution.

1 プロセス抽出はエンティティ・リンキングと類似しているが、エンティティ・リンキングがメンションの抽出とオントロジー中のコンセプト ID への紐づけを目的とするのに対して、プロセス抽出はエンティティがメンションとして明示的に現れない場合でも、文書に対してコンセプト ID を紐づけることを目的とする。

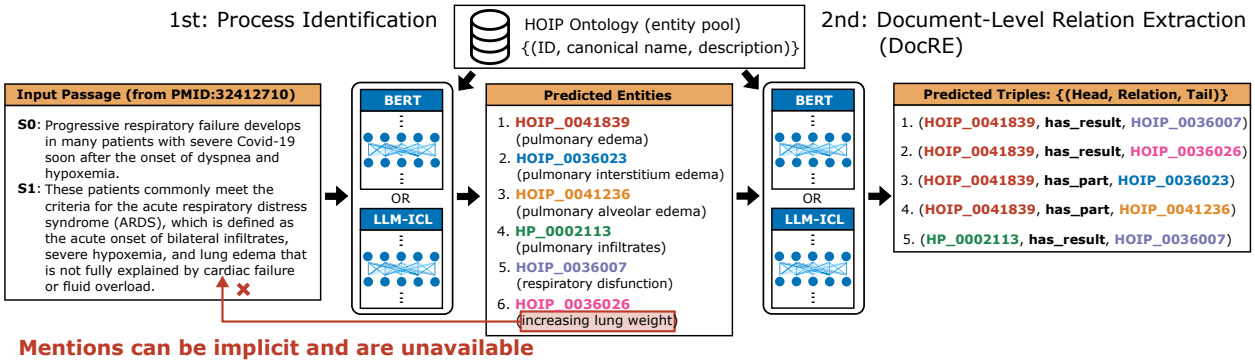

図 1　プロセス抽出 (Process Identification) と文書レベル関係抽出 (Document-Level Relation Extraction; DocRE) の二段階からなるメンション非依存型情報抽出システムの概要図。図中の例は HOIP データセットに収録されている実際の事例。

ステムの開発とベンチマークを促進するため、HOIP データセットを構築し、公開した (A 節を参照)。

実験の結果、プロセス抽出では、LLM による生成アプローチが BERT ベースの識別アプローチを上回ることを観測した。一方で、文書レベル関係抽出 (DocRE) では、提案したメンション非依存型 ATLOP モデルが LLM ベースの方法を上回り、メンションのヒントなしでも F1 スコアで 56-59 ポイントに到達可能であることを示した。しかし、パイプラインレベルの評価の結果、現段階ではエラー伝播の影響が大きく、プロセス抽出と DocRE の両方の改善が重要であることが示された。本実験で用いたソースコードおよびデータセットは公開している。[2]

2　メンション非依存型情報抽出

本節では、図 1 の二段階情報抽出システムを構成するプロセス抽出タスクおよび文書レベル関係抽出 (DocRE) タスクのそれぞれに対して、メンション非依存型アプローチを提案する。

2.1　プロセス抽出

プロセス抽出タスクでは、入力文書に対して、そこで明示的または暗黙的に言及されているエンティティの集合を同定する。各エンティティは、対象オントロジーに登録されているエンティティ (またはコンセプト) の集合 (一意の ID が付与されている) を候補として、そこから選ばれる。

BERT によるプロセス抽出:　プロセス抽出タスクは、対象オントロジー中の各エンティティをクラスと見なすことで、多クラス・マルチラベル分類問題として定式化することができる。そのような多クラス・マルチラベル分類問題を、BERT ベースの二値分類器によって解く。具体的には、入力文書 d と各エンティティ候補 e の組 (d, e) に対して、BERT ベース二値分類器によって、エンティティ e が文書 d で語られているか否かを二値分類する。BERT への入力では、各エンティティはオントロジーで定義される正式名によって表現され、文書と [SEP] トークンを挟んで結合される。

LLM によるプロセス抽出:　大規模言語モデル (LLM) を用いるプロセス抽出では、入力文書と少数の入出力事例 (demonstrations) をモデルに与え、入力文書中で明示的または暗黙的に言及される生物学的プロセスを列挙するように指示する。生成されたプロセスエンティティを対象オントロジー中のエンティティと紐づけるため、生成されたエンティティとオントロジー中のエンティティの埋め込み表現を Sap-BERT [12] によって求め、各生成エンティティに対して、コサイン類似度が高い上位 K 個のエンティティを紐づけ、すべての生成エンティティについてまとめたものを、入力文書に対する出力エンティティ集合とする。

2.2　文書レベル関係抽出 (DocRE)

入力文書 d、プロセス抽出によって得られたプロセスエンティティの集合 $\{e_1, \cdots, e_k\}$、および事前定義された関係ラベルの集合 $\mathcal{R}$ を入力として、文書レベル関係抽出 (DocRE) では、エンティティペア (e_i, e_j) に対して、関係ラベル $r \in \mathcal{R} \cup \{\text{NA}\}$ を予測することを目的とする [5, 6, 7]。ここで、e_i および e_j はそれぞれ主語エンティティおよび目的語エンティティを表し、NA ラベルはエンティティペアに関係が存在しないことを示す。従来の DocRE では、

2　HOIP Dataset: https://github.com/norikinishida/hoip-dataset; Process Identification code: https://github.com/sl-633/bio-process-identifier; DocRE code: https://github.com/norikinishida/kapipe.

各エンティティは文書 d 中の具体的なメンションに紐づき、メンションに関する情報も与えられるが、本研究の DocRE では、そのようなメンション情報は与えられない。

QA 型 DocRE モデル (QA-Model) メンションに依存しないエンティティ間の関係分類は、Yes-No 型の質問応答 (QA) タスクとして定式化することができる。すなわち、入力として与えられたすべての可能なエンティティペアと関係ラベルからなるトリプル (e_i, r, e_j) について、関係ラベルごとに事前に定義したテンプレートを用いて質問文 q を生成する。ここで、各エンティティはオントロジーで定義された正式名によって表現される。入力文書 d と質問文 q を結合し、BERT エンコーダ [8, 13] と 2 層フィードフォワード型分類器によって質問に対する Yes-No 型の回答出力 (二値分類) を行う。

メンション非依存型 ATLOP (Mention-Agnostic ATLOP; MA-ATLOP): 上記の QA 型モデルは、膨大な数のトリプル候補について推論を行う必要があるため、計算効率が低い。そこで、一回のフォワードパスで全ての可能なトリプルに対して関係分類を行うために、従来の一般的な DocRE 手法である ATLOP [5] を、明示的なメンションに依存しない方式に拡張する。ATLOP では、入力文書 d を BERT ベースのエンコーダによってエンコードし、各エンティティのメンション (i.e., 位置とスパン) 情報をもとに、エンティティの埋め込み表現を求める。しかし、本研究ではメンションの情報に頼ることができない。そこで、入力文書 d とエンティティ e_i の組 (d, e_i) を各エンティティについて独立にエンコードし、エンティティ e_i が文書 d 内でどのように記述されているかを考慮した密なベクトルを生成する。ここで、エンティティはオントロジーに登録された正式名と定義文によって表される。その後、ATLOP と同様に、Group Bilinear 分類器 [14] によってエンティティ間の関係分類を効率的に行う。

LLM による DocRE: LLM による DocRE では、入力文書とエンティティ集合、および少数の入出力事例をモデルに与え、関係トリプルを箇条書きで生成するように指示する [9, 10, 11]。ここで、各エンティティはその正式名で表現する。生成された箇条書きの各行から、正規表現を用いて主語エンティティ ID e_i、関係ラベル r、目的語エンティティ ID e_j を抽出し、出力トリプルとする。

表 1 プロセス抽出の結果。K の値は、LLM によって生成された各プロセスエンティティについて、埋め込み表現の類似度が高い上位 K 個のオントロジー中エンティティを紐づけたことを表す。

手法	Precision	Recall	F1
BERT ベース (PubMedBERT [15])	22.2	42.6	29.2
LLM ベース (Llama3 8B) K=1	**34.3**	28.4	**31.1**
LLM ベース (Llama3 8B) K=5	15.6	55.9	24.4
LLM ベース (Llama3 8B) K=10	9.5	**62.1**	16.5

3 実験

3.1 プロセス抽出の結果と考察

プロセス抽出における提案手法の有効性を HOIP データセットを用いて評価した。訓練データ、検証データ、テストデータはそれぞれ 255 件、37 件、35 件の文書事例から構成される。また、評価尺度としては、モデルが出力するエンティティの集合と、人手で付与されたエンティティ集合を比較し、Precision/Recall/F1 スコアを用いた。

結果を表 1 に示す。実験の結果から、LLM ベースの生成アプローチが、BERT ベースの識別アプローチ (多クラスマルチラベル分類) を F1 スコアおよび Precision で上回ることが観測された。これは、候補エンティティの数と多様性に対して、HOIP データセットが含む教師情報のカバレッジが十分ではなく、BERT ベースの教師ありモデルが十分に汎化していないことを示唆する。一方で、LLM は医学生物学に関する知識を事前学習で獲得している可能性があり、それと少数入出力事例を活用できていると考えられる。特に、K=10 の場合では、Precision は低いものの、Recall は 62.1%に達成している。

3.2 DocRE の結果と考察

次に、DocRE における提案手法の有効性を評価した。データセットとしては、HOIP データセットと CDR データセット [16] を用いた。CDR は、1,500 件の PubMed 論文の要旨 (英語) から構成され、Chemical または Disease タイプに属するエンティティメンションと、それらの MeSH ID、エンティティの間の Chemical-Disease-Relation 関係が人手で付与されている。尚、CDR は HOIP オントロジーではなく MeSH を対象にしているため、CDR を用いた実験では、オントロジーとして HOIP ではなく MeSH を用いた。モデルが出力する関係トリプルの

表 2　CDR テストセットにおける DocRE 実験の結果。P, R, F1 はそれぞれ Precision, Recall, F1 スコアに対応。

モデル	メンション活用	P	R	F1
ATLOP	全メンション	**64.61**	**75.92**	**69.74**
Llama3 8B	全メンション	42.26	48.69	45.25
QA-Model	初出メンションのみ	56.40	67.39	61.36
MA-ATLOP	初出メンションのみ	57.54	**68.11**	**62.34**
Llama3 8B	初出メンションのみ	43.62	49.34	46.30
QA-Model	メンション非依存	53.37	64.01	58.12
MA-ATLOP	メンション非依存	53.72	65.92	59.18
Llama3 8B	メンション非依存	44.75	51.97	48.09

集合と、人手で付与された関係トリプルの集合を比較し、Precision/Recall/F1 スコアによって、モデルの性能を評価した。

CDR における結果を表 2 に示す。表の上段、中段、下段のブロックはそれぞれ、メンションの情報をフルに活用できる場合、オントロジーから取得した正式名ではなく入力文書で最初に出現するメンション表現をエンティティ名として使用できる場合、一切のメンション情報を活用できない場合の結果に対応する。すべてのメンション情報を活用可能なとき (上段)、ATLOP は F1 スコアで 69.74 を達成している。これに対し、メンション非依存型モデルの中で最良の結果を示した MA-ATLOP（初出メンションのみ) は、F1 スコアが 62.34 であり、ATLOP (全メンション) より 7.4 ポイント低い結果となった。メンションのヒントが全くない場合 (下段)、MA-ATLOP と QA-Model はそれぞれ F1 スコアが 59.18 と 58.12 を記録した。これらの結果は、メンションのヒントがなくても、提案手法が関係トリプルを期待以上に正確に識別できることを示唆するが、一方で、それでもなおメンションが関係性の認識に重要な役割を果たしていることを示している。MA-ATLOP は QA-Model を一貫して上回る結果を示し、計算効率の面でも QA-Model より優れていることから、実応用では MA-ATLOP がより適していると言える。また、プロセス抽出の結果とは反対に、本タスクでは、BERT ベースの教師ありモデルが LLM ベースのモデルを上回る傾向にあることを観測した。

HOIP データセットにおける結果を表 3 に示す。表の上段は、人手で付与された gold エンティティを使用した場合の結果を示している。HOIP データセットにはメンションが注釈されていないため、メンション非依存なモデルのみを評価した。

表 3　HOIP テストセットにおける DocRE 実験の結果。

モデル	エンティティ	P	R	F1
QA-Model	Gold	51.5	**63.1**	56.7
MA-ATLOP	Gold	67.2	52.6	**58.9**
Llama3 8B	Gold	18.5	16.7	17.6
Upper-bound	予測	100.0	26.8	42.3
MA-ATLOP	予測	7.7	14.9	10.2

BERT ベースのアプローチ (MA-ATLOP, QA-Model) は、LLM ベースのアプローチ（F1 スコア 17.6）と比較して、はるかに高い F1 スコア（56.5〜59.0）を達成した。また、MA-ATLOP は QA-Model を 2.2 ポイント上回る F1 スコアを記録した。これらの結果は、CDR データセットでの結果と一致しており、MA-ATLOP が本タスクにおいて、QA-Model および LLM ベースの方法に比べて、精度と計算効率の両面で有効であることを示している。

次に、より現実的な設定でのシステム全体の性能を評価するため、LLM (Llama3 8B) によるプロセス抽出器によって予測されたエンティティ集合に対し、MA-ATLOP を適用し、その質を評価した。表 3 の下段にその結果を示す。まず、予測されたエンティティに基づいて作成可能な正解トリプルのサブセットを作成し、上限スコアを計算した結果、Precision、Recall、F1 スコアはそれぞれ 100.0、26.8、42.3 であった。低い Recall は、プロセス抽出の False Negative の多さが、現状のボトルネックの一つであることを示唆している。MA-ATLOP の予測結果は、Precision が 7.7、Recall が 14.9、F1 スコアが 10.2 であった。gold エンティティを用いた場合の高い Precision（67.2）と比較すると、この結果は、現在の DocRE モデルが無関係なエンティティ (プロセス抽出の False Positives) を出力トリプルから除外するのに苦戦していることを示唆している。まとめると、現在の状況では、プロセス識別と DocRE における Recall（カバレッジ）と Precision（ノイズの少なさ）の両方を改善する必要があり、このタスクの難しさを示している。

4　おわりに

本研究は、オントロジーに基づく医学生物学的知識の自動体系化を支援するため、新しいベンチマークデータセットと、エンティティおよび関係性の抽出のためのメンション非依存型アプローチを提案した。

謝辞

本研究は JSPS 科研費 JP22K17959，JP21K17815，JP22H05015，および ANR AIBy4 Project (ANR-20-THIA-0011) の助成を受けたものです.

参考文献

[1] Yuki Yamagata, T. Kushida, Shuichi Onami, and Hiroshi Masuya. Ontology development for building a knowledge base in the life science and structuring knowledge for elucidating the covid-19 mechanism. In **Proceedings of the Annual Conference of JSAI**, pp. 3H1GS3d01–03H1GS03d01, 2021.

[2] Yuki Yamagata, Tsubasa Fukuyama, Shuichi Onami, and Hiroshi Masuya. Prototyping an ontological framework for cellular senescence mechanisms: A homeostasis imbalance perspective. **Sci Data**, Vol. 11, p. 485, 2024.

[3] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6470–6476, Online, July 2020. Association for Computational Linguistics.

[4] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. 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. 6397–6407, Online, November 2020. Association for Computational Linguistics.

[5] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 35, pp. 14612–14620, 2021.

[6] Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. SAIS: Supervising and augmenting intermediate steps for document-level relation extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2395–2409, Seattle, United States, July 2022. Association for Computational Linguistics.

[7] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In Zhi-Hua Zhou, editor, **Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21**, pp. 3999–4006. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[8] 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.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[10] 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.

[11] Yilmazcan Ozyurt, Stefan Feuerriegel, and Ce Zhang. Document-level in-context few-shot relation extraction via pre-trained language models, 2024.

[12] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4228–4238, June 2021.

[13] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[15] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **CoRR**, Vol. abs/2007.15779, , 2020.

[16] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. **Database : the journal of biological databases and curation**, 2016.

[17] M. Ashburner, et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. **Nat Genet**, Vol. 25, pp. 25–29, 2000.

A　HOIP データセット

本節では、本研究で構築した HOIP オントロジーの詳細について説明する。本研究では、自動構築の目標オントロジーとして、COVID-19 感染メカニズム（プロセス）の理解に焦点を当てる Homeostasis Imblance Process Ontology（HOIP）[1, 2] を選択した。メンション非依存型情報抽出システムの開発とベンチマークを促進するため、HOIP オントロジーに基づいて HOIP データセットを構築し、公開した。本データセットは、COVID-19 感染メカニズムに関連する医学生物学プロセスを記述した PubMed 論文から抽出された文章から構成される。各文章は、PubMed 論文の一部であり、少なくとも 2 つの具体的なプロセスを記述している。これらのプロセスは手作業で注釈され、(主語エンティティ, 関係, 目的語エンティティ) の形式のトリプルとして表現されている。

A.1　データ収集と改善

最初のステップとして、HOIP オントロジーのファイルを Apache Jena Fuseki[3]を使用して RDF ストアに保存し、SPARQL エンドポイントを構築した。データセットに必要な情報は SPARQL クエリを用いて取得し、その結果を CSV 形式に変換した。データセットを自動注釈に適するように、また明確性を向上させるため、HOIP オントロジーの階層構造に基づいていくつかの調整を行った。一部のエンティティには *course* の情報が含まれており、例えば “blood vessel damage *in severe COVID-19*” (course はイタリック体で表現) のように、course は特定の文脈を表現している。これらの course 情報はしばしば冗長であるため、エンティティから削除した。さらに、あまりにも細分化されたプロセスは自動抽出の評価には適さないと判断し、各プロセスの上位クラスを用いて、より一般化されたプロセスを優先した。ここでは Gene Ontology (GO) [17] に収録されている用語を割り当てるアプローチを採用した。この方法により、注釈が実用的で再利用性の高いものとなるようにしている。

A.2　データセットの整備

生成された CSV ファイルでは、各レコード (各行) が 1 つのトリプルに対応している。同じテキストおよび PubMed ID に関連付けられたトリプルを 1 つのグループにまとめ、このグループを最終データセットにおける 1 つの事例とした。文章間でテキストの重複が存在することが判明し、一方の文章 d_{src} が片方の文章 d_{dst} にテキスト的に包含され、かつ両者が同じ PubMed ID に関連付けられている場合、d_{src} に関連付けられたトリプル集合 T_{src} を d_{dst} のトリプル集合 T_{dst} に統合した。最終的に、同じ PubMed ID の論文から抽出された事例が、訓練セット、検証セット、テストセットに横断的に散在しないように配慮しながら、データセット全体を訓練セット、開発セット、テストセットに分割した。データセットの統計情報を表 4 に示す。

	Train	Dev	Test
文章数	255	35	37
エンティティ数	1988	143	211
トリプル数	1848	137	177
平均単語数 (per 文章)	75.5	70.4	61.8
平均エンティティ数 (per 文章)	7.8	4.1	5.7
平均トリプル数 (per 文章)	7.2	3.9	4.8

表 4　HOIP データセットの統計情報。

3　https://jena.apache.org/documentation/fuseki2/