情報抽出パイプラインにおけるエラー伝播抑制手法の提案：オーバーサンプリング、フィルタリング指向学習、概念対応

西田典起 [1]　Shanshan Liu[1]　Rumana Ferdous Munne[1] 徳永なるみ [1]　山縣友紀 [1]
古崎晃司 [2]　松本裕治 [1]
[1] 理化学研究所　[2] 大阪電気通信大学
noriki.nishida@riken.jp

概要

本研究では、情報抽出パイプラインにおけるエラー伝播問題を軽減するための方法を提案する。エンティティの偽陰性問題 (抽出漏れ) に対して、「オーバーサンプリング」を導入する。また、エンティティの偽陽性問題 (過剰抽出) に対しては、「フィルタリング指向学習」を提案する。さらに、エンティティ曖昧性解消による間違った概念とメンションの紐づけの文書レベル関係抽出に対する影響を軽減するために、「概念対応」を導入する。複数ドメインのデータセットを用いた実験の結果、提案手法によって、情報抽出パイプラインのエラー伝播に対するロバスト性が向上することを確認した。

1　はじめに

これまで、様々な情報抽出システムが提案されてきた [1, 2, 3, 4, 5]。それらの多くは、**パイプライン型**の枠組みを採用している。本研究では、情報抽出パイプラインを以下の 3 タスクのカスケードとして定義する。最初の固有表現抽出（Named Entity Recognition; NER）では、入力テキスト中のエンティティのメンション (言及) を検出する [6, 7]。次のエンティティ曖昧性解消（Entity Disambiguation; ED）では、検出されたメンションを知識グラフ内の一意の概念 ID にリンクし、概念 ID に基づいてメンションをエンティティ単位にグループ化する [8, 9, 10, 11]。最後の文書レベル関係抽出（Document-level Relation Extraction; DocRE）では、これらのエンティティ間の関係を識別し、知識グラフの拡張に必要な関係構造を構築する [12, 13, 14, 15]。

パイプライン型には、特に柔軟性の面でいくつかの利点がある。各モジュールが独立して動作するため、システム全体を大幅に変更することなく、個々のコンポーネントを更新または置き換えることが可能である。また、このモジュール性により、メンテナンスが容易になり、技術の進歩やユーザー要件の変化に迅速に対応ができる。

一方で、パイプライン型には、**エラー伝播**に対して脆弱であるという題題がある [16]。情報抽出パイプラインではタスクが連鎖的に進行するため、NER や ED 段階で発生したエラー (誤陰性、誤陽性) が DocRE タスクまで伝播し、システム全体の結果の質を損なうことが知られている [17]。エラー伝播への主なアプローチとしては、複数のタスクを単一のアーキテクチャで処理する Joint Framework が挙げられる [18, 3, 5]。しかし、Joint Framework はモデル間の依存性を高めるため、前述したモジュール性や柔軟性が大幅に損なわれるという仕様的な欠点がある。

そこで本研究では、情報抽出パイプラインの柔軟性を享受しつつエラー伝播問題を軽減するための三つの方法を提案し、それの効果を調査する。(1) 偽陰性 (false negatives) エラー (NER や ED 段階で検出漏れしたエンティティ) に対しては、**オーバーサンプリング**戦略を導入する。特に NER 段階において抽出メンションをオーバーサンプリングすることにより、正しいエンティティが ED や DocRE に渡される可能性を高め、システム全体の Recall を向上させることを狙う。(2) オーバーサンプリングは Recall を向上させる一方で、無関係 (ノイジー) なメンションやエンティティを増加させてしまうリスクがある。この偽陽性 (false positives) エラーに対しては、**フィルタリング指向学習**を導入する。具体的には、DocRE モデルを手動アノテーションされたメンションやエンティティだけでなく、前段階モジュールのオーバーサンプリング結果 (自動アノテーション) と組み合わせて訓練することで (言い換えると訓練時

に敢えてノイズを混入させることで)、ノイズの多い入力に対する DocRE モデルのロバスト性を向上させる。(3) エンティティ曖昧性解消のエラー (メンションとエンティティの誤リンク) に対しては、**概念対応**という方法を提案する。概念対応では、従来のようにエンティティの表現をメンションの表現のみから求めるのではなく、知識グラフに登録される正式名や定義文を用いることで、エンティティ表現を補完し、DocRE の精度を高める。

医学生物学ドメイン (CDR) と一般ドメイン (Linked-DocRED) におけるベンチマークデータセットを用いて提案手法の評価を行い、従来のパイプラインと比較してエラー伝播に対するロバスト性が大幅に向上することを示した。オーバーサンプリングによるカバレッジの向上とフィルタリング指向学習および概念対応によるノイズ軽減を組み合わせることで、実世界の情報抽出においてより信頼性が高く、ロバストな解決策が提供できることを示す。

2 手法

本節では、実験で構築した情報抽出パイプラインについて説明し、次にパイプライン内のエラー伝播の軽減手法としてオーバーサンプリング、フィルタリング指向学習、概念対応をそれぞれ説明する。図 1 に、情報抽出パイプラインと各手法の概要を示す。

2.1 情報抽出パイプラインの形式化

本研究のパイプラインは、入力文書に対して関係トリプルの集合を出力する。関係トリプル (h, r, t) は、主語エンティティ $h \in \mathscr{C}$、目的語エンティティ $t \in \mathscr{C}$、および関係ラベル $r \in \mathscr{R}$ で構成される。ここで、$\mathscr{C}$ は知識グラフ内の概念ノード ID の集合を、$\mathscr{R}$ は関係ラベルの集合を表す。

固有表現認識（NER）では、テキスト中のメンションの集合 M を検出する。各メンション $m_i \in M$ は、テキストスパンと対応するエンティティタイプからなるタプルとして表される。本研究では、スパンベースの NER 手法として広く利用されている Biaffine-NER モデル [6] を採用した。このモデルは、まずテキスト内の全ての可能なスパンを列挙し、それぞれのスパンに対してエンティティタイプ（*Non-Entity* クラスを含む）のスコアを予測する。次に、Non-Entity に分類されなかったスパンをスコアと選択ルールに基づいてフィルタリングする。

次に、エンティティ曖昧性解消（ED）では、各メンション $m_i \in M$ を特定の概念 ID にリンクし、タプル (m_i, id_i) を作成する。これらのメンションは概念 ID に基づいてエンティティレベルにグループ化され、エンティティの集合 E を形成する。各エンティティ $e_k \in E$ は以下のタプルとして表される。本研究では、BLINK [8] を採用した。BLINK の詳細については紙面の都合上割愛する。

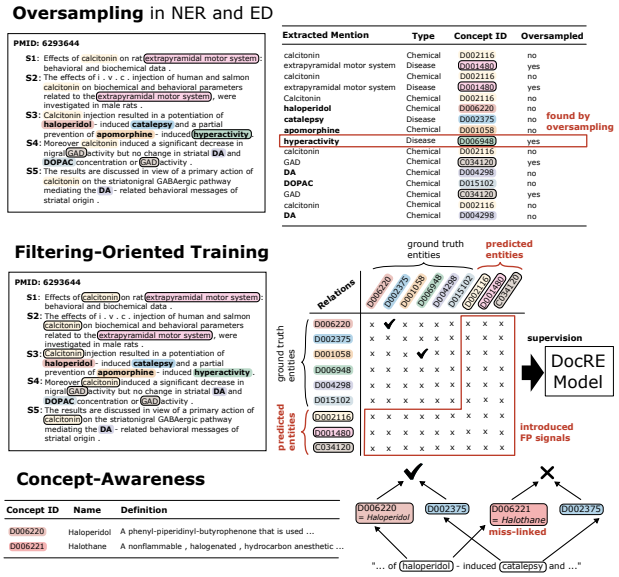

図 1 提案手法の概要図。ハイライトの色はそれぞれのエンティティ（概念）に対応する。太字は正しいメンションを表す。

文書レベル関係抽出（DocRE）では、各エンティティペア (h, t) に対して関係ラベル $r \in \mathscr{R}$ を予測する。ここで、$h \in E$、$t \in E$ である。最終的なパイプラインの出力は、トリプル (h, r, t) の集合である。本研究では、DocRE モデルとして広く使用されている ATLOP [12] を採用した。ATLOP は、メンション集合 M とエンティティ集合 E を入力として、テキストをエンコードし、スパン表現に基づいてメンションの埋め込み表現を取得する。その後、各エンティティ $e_k \in E$ に関連付けられたメンションの埋め込み表現をプーリングすることで、エンティティの埋め込み表現 $\boldsymbol{e}_k$ を獲得する。さらに、ATLOP は各可能なエンティティペアに対して文脈的な埋め込みを計算する。これらの埋め込みを用いてエンティティ間の関係を予測する。

2.2 オーバーサンプリング

偽陰性は、情報抽出パイプラインにおけるエラー伝播の主な原因の一つであり、NER および ED モジュールでエンティティが検出されず、見過ごされ

た場合に発生する。前節で述べたように、DocRE モジュールは NER および ED モジュールで特定されたエンティティ間の関係を予測する。そのため、検出できなかった正しいエンティティを含むトリプルは予測されることがなく、パイプライン全体の Recall の低下を招く。この問題に対処するには、DocRE モジュールではなく、NER および ED モジュールに対する介入が必要である。

偽陰性問題を軽減するために、NER および ED モデルの出力を追加的にサンプリングすることで、DocRE の入力に正しいエンティティが含まれる可能性を高めるオーバーサンプリングを導入する。具体的には、Biaffine-NER に対して以下のような変更を加える: 1. すでに出力結果として選ばれたメンションと重ならないスパンのうち、Non-Entity に分類されていない全てのスパンを出力に追加する。2. 予備実験の結果、上記の修正だけでは Recall を大幅に向上させるには不十分であることが判明した。これは、モデルの偽陰性の多くは、選択ルールによってフィルタリングされたのではなく、誤って Non-Entity に分類されていることを示唆する。これに対応するため、Non-Entity と予測されたスパンから追加的にサンプリングを行う。具体的には、Non-Entity と予測されながらも対応する確率が一定の閾値未満のスパンを出力に追加する。この閾値は、開発セットでの評価スコアに基づいて調整した。

2.3 フィルタリング指向学習

DocRE の入力には、NER および ED によって出力される無関係なメンションやエンティティが含まれる。さらに、オーバーサンプリングはメンション・エンティティの Recall を高める一方で、DocRE の入力にノイズを不可避的に追加してしまう。DocRE モデルの入力ノイズに対するロバスト性を向上させるために、フィルタリング指向学習を提案する。具体的には、学習時に、手動アノテーションされたメンションやエンティティだけでなく、オーバーサンプリングによって追加されるメンションやエンティティも含めて学習を行う。オーバーサンプリングによって追加されるメンション・エンティティは、訓練時では常に偽陽性であるため、提案手法ではこれらの偽陽性を積極的に訓練に利用する。

2.4 概念対応

従来の DocRE モデルは、エンティティ埋め込みをエンティティに関連付けられたメンションのみを基に計算する。つまり、エンティティ間の関係を予測する際に、それらのエンティティが表す意味的概念は考慮されず、メンションのみが情報として用いられる。例えば、ATLOP では、メンションスパンの埋め込み表現に対して Log-Sum-Exp プーリングを適用することで、エンティティの埋め込み表現を求める。

このような手法は、メンションとエンティティの対応が正確であるベンチマーク環境では効果的に機能するが、現実のシナリオではこの対応がしばしば不正確である。そのため、正確な関係予測を行うには、メンションだけでなく、エンティティに関連付けられた意味的概念も考慮することが不可欠である。

DocRE モデルにおけるエンティティ埋め込みを改善するため、エンティティのテキスト上のメンションだけでなく、知識グラフから取得した概念の意味情報を活用する。この概念対応を DocRE モデルに導入することで、エンティティを正確に表現する能力を向上させ、特にエンティティの曖昧性解消が困難なシナリオにおいて、関係抽出タスクの性能を強化することを目指す。具体的には、ATLOP の従来のメンションに基づくエンティティ埋め込みプロセスに加えて、入力文書と概念の正式名、定義文を用いてエンティティ埋め込みを行い、それら 2 種類の結果を組み合わせる。

3 実験結果と考察

図 2 に、オーバーサンプリングが NER および ED の結果に与える影響を示す。紙面の都合上、Linked-DocRED における実験結果は表示しないが、同様の結果であった。CDR [19] (医学生物学ドメイン) および Linked-DocRED データセット [20] (一般ドメイン) において、オーバーサンプリングにより、NER タスクの Recall が平均 4.4 ポイント、ED タスクの Recall が平均 3.0 ポイント向上することがわかった。この結果は、オーバーサンプリングが見逃されたメンションやエンティティに関する偽陰性問題を効果的に解消することを示している。一方で、この手法は若干のノイズを追加しており、Precision の低下が確認されました。しかし、後述するように、こ

表 1　CDR および Linked-DocRED における DocRE 評価の結果。

手法	メンション・エンティティ	CDR			Linked-DocRED		
		P	R	F1	P	R	F1
ベースパイプライン	予測	40.20	56.38	46.93	18.92	20.52	19.69
+ オーバーサンプリング	予測	35.52	**58.35**	44.16	15.94	**20.87**	18.08
+ フィルタリング指向学習	予測	**53.68**	47.84	50.60	27.39	14.15	18.66
+ 概念対応	予測	51.52	51.03	**51.27**	**33.10**	16.91	**22.38**
アッパーバウンド	予測	100.0	77.49	87.31	100.0	41.29	58.45
ATLOP	ゴールド	65.04	75.23	69.77	66.09	59.93	62.86

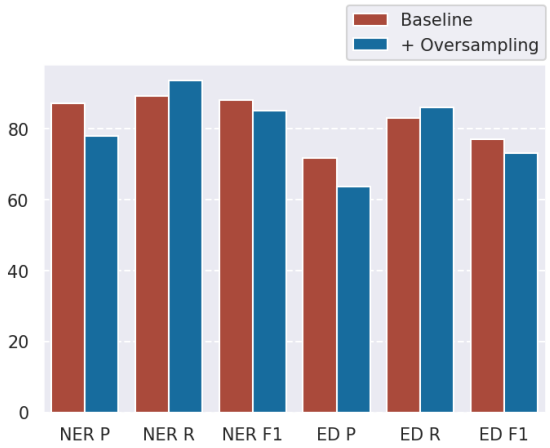

図 2　CDR における NER および ED の Precision (P), Recall (R), F1 スコア。

の低下した F1 スコアは、フィルタリング指向学習と概念対応を利用することで改善される。

表 1 は、オーバーサンプリング、フィルタリング指向学習、および概念対応が DocRE の結果に与える影響を示している。オーバーサンプリングにより、両データセットにおける DocRE 予測の Recall の向上が確認された。これは、図 2 で示されるように、DocRE モデルが利用可能なメンションおよびエンティティの Recall が向上したためである。しかし、オーバーサンプリングは Recall を向上させる一方で、ノイズを大幅に増加させ、Precision の顕著な低下を引き起こすことが確認された。

DocRE モデルのフィルタリング指向学習により、Precision の顕著な向上が確認された。具体的には、オーバーサンプリングのみを導入したパイプラインと比較して、CDR では 18.2 ポイント、Linked-DocRED では 8.5 ポイントの Precision 向上が見られました。また、この手法を導入したパイプラインは、CDR データセットにおいてベースパイプラインの F1 スコアを 3.5 ポイント以上上回った。これらの結果は、フィルタリング指向学習が無関係 (ノイジー) なメンションやエンティティによる偽陽性問題の軽減に効果的であることを示している。さらに、DocRE モデルを、現実のエラー伝播を反映したメンションやエンティティを用いて学習することの重要性を示唆する。

概念対応によってエンティティエンコーディングに概念情報（正式名と定義）を組み込むことで、特に F1 スコアが両データセットでさらに向上することがわかった。具体的には、ベースパイプラインと比較して、CDR では 0.7 ポイント、Linked-DocRED では 3.7 ポイントの F1 スコアの向上が確認された。両データセットにおける F1 スコアの大幅な向上は、エンティティ埋め込みにメンション表現だけでなく概念情報を統合することが、正しいトリプルを識別する上で重要であることを示している。

これらの結果は、情報抽出パイプラインに提案した改良、すなわちオーバーサンプリング、フィルタリング指向学習、および概念対応の導入が、評価対象となったすべてのデータセットで性能を大幅に向上させることを示している。

参考のため、エンドツーエンドで予測されたエンティティ集合における DocRE 性能の上限値と、正解メンションおよびエンティティを使用した場合の結果を示す。上限値の Recall との大きな差は、モデルにさらなる改良の余地がまだ十分にあることを示している。また、正解設定との大きな差異は、情報抽出システムのエンドツーエンド評価の重要性を示唆している。

4　おわりに

本研究では、情報抽出パイプラインにおけるエラー伝播問題を軽減するために、オーバーサンプリング、フィルタリング指向学習、および概念対応エンティティ表現を提案した。実験の結果は、各手法およびその組み合わせの有効性を示唆している。

謝辞

本研究は JSPS 科研費 JP22K17959，JP21K17815，JP22H05015 の助成を受けたものです.

参考文献

[1] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. 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. 3036–3046, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7999–8009, Online, July 2020. Association for Computational Linguistics.

[3] Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. GenIE: Generative information 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. 4626–4643, Seattle, United States, July 2022. Association for Computational Linguistics.

[4] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4904–4917, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] John Giorgi, Gary D Bader, and Bo Wang. A sequence-to-sequence approach for document-level relation extraction. In **Workshop on Biomedical Natural Language Processing**, 2022.

[6] 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.

[7] Enwei Zhu and Jinpeng Li. Boundary smoothing for named entity recognition. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7096–7108, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] 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.

[9] Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. Global entity disambiguation with BERT. 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. 3264–3271, Seattle, United States, July 2022. Association for Computational Linguistics.

[10] Hassan Shavarani and Anoop Sarkar. SpEL: Structured prediction for entity linking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11123–11137, Singapore, December 2023. Association for Computational Linguistics.

[11] Qinyong Wang, Zhenxiang Gao, and Rong Xu. Exploring the in-context learning ability of large language model for biomedical concept linking, 2023.

[12] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. **CoRR**, Vol. abs/2010.11304, , 2020.

[13] 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.

[14] 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.

[15] Youmi Ma, An Wang, and Naoaki Okazaki. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1971–1983, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[16] Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. Neural relation extraction for knowledge base enrichment. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 229–240, Florence, Italy, July 2019. Association for Computational Linguistics.

[17] Shiao Meng, Xuming Hu, Aiwei Liu, Fukun Ma, Yawen Yang, Shuang Li, and Lijie Wen. On the robustness of document-level relation extraction models to entity name variations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 16362–16374, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[18] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. **CoRR**, Vol. abs/1909.03546, , 2019.

[19] 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 J. Biol. Databases Curation**, Vol. 2016, , 2016.

[20] Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Martino Lovisetto. Linked-docred - enhancing docred with entity-linking to evaluate end-to-end document-level information extraction pipelines. In **Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '23, p. 3064–3074, New York, NY, USA, 2023. Association for Computing Machinery.