研究活動で産出された論文と研究データの対応付け

岡田怜真 [1]　茂木光志 [2]　伊藤滉一朗 [2]　松原茂樹 [2,3]
[1] 名古屋大学情報学部　[2] 名古屋大学大学院情報学研究科
[3] 名古屋大学情報基盤センター
{okada.ryoma.p0, motegi.koshi.h9}@s.mail.nagoya-u.ac.jp
{ito.koichiro.z5, matsubara.shigeki.z8}@f.mail.nagoya-u.ac.jp

概要

オープンサイエンスの潮流のもと，論文と研究データの公開が進んでいる．同一の研究活動で産出された論文と研究データが相互に参照可能であることは，双方の理解を深めるうえで重要となる．本論文では，同一の研究活動で産出された論文と研究データを対応付ける手法について述べる．本手法では，論文中の URL を，その周辺テキストを利用して，対応する研究データの公開先であるか否かに分類する．BERT ベースのモデルを fine-tuning することで本手法を実装し，実験でその性能を評価した．

1　はじめに

近年，オープンサイエンスの世界的な潮流のもと[1, 2]，研究成果の公開が進んでいる．研究成果は論文と研究データに大別され，研究データにはデータセットやソフトウェアなどが含まれる．研究成果の利活用を促進するために，研究成果を他の成果と関連付けることが有効である．関連付けの観点は様々存在するが，本研究では，同一の研究活動で産出された論文と研究データの対応付けに着目する．

同一の研究活動で産出された論文と研究データが対応付けられて，相互に参照可能であることは，双方の理解を深めるうえで重要となる．例えば，論文で報告された研究を再現する際には，その研究で産出された研究データが手がかりとなる．また，研究データを利用する際は，その作成方法も把握する必要があり，その研究で産出された論文が手がかりとなる．しかし，膨大な量の論文と研究データを人手で対応付けることは，即時性，網羅性，持続可能性の点で現実的ではない．

本論文では，同一の研究活動で産出された論文と研究データを自動で対応付ける手法について述べる．論文には，対応する研究データの公開先 URL が記載されることがある（図 1 上）．この種の URL は，研究データの識別子とみなせるため，それらを論文から獲得することで，同一の研究で生み出された論文と研究データが対応付けられることになり，相互に参照可能となる．しかし，論文に記載される URL は，論文に対応する研究データの公開先であるとは限らない（図 1 下）．そこで，論文中の URL を，その周辺のテキストを利用して，対応する研究データの公開先であるか否かに分類する．

本手法の分類性能を評価するために実験を行った．まず，論文中の URL が，当該論文に対応する研究データの公開先であるか否かが付与されたデータを作成した．次に，作成したデータを用いて，URL の分類手法を実装した．具体的には，事前学習済みの BERT[5] ベースのモデルを fine-tuning することで実装した．最後に，実装した手法の分類性能を作成したデータで評価した．

> … , and adding cross-sentence relations using coreference links. Our dataset is publicly available at: http://nlp.cs.washington.edu/sciIE/. Table 1 shows the statistics of SciERC.

論文と対応する研究データの公開先URL

> … . We leave it as future work to augment our dataset with these structured fields. To extract tokens and sentences, we use the SpaCy (https://spacy.io/) library.

論文と対応する研究データの公開先でないURL

図 1　論文に記載される URL の例（出典：[3, 4]）

2　先行研究及び関連動向

2.1　研究データのメタ情報の獲得

これまでに，研究データのメタデータ生成に有用な情報（メタ情報）を論文から獲得する手法が提案されている．データセットやソフトウェアの名称などがアノテーションされた論文コーパスとして，SciERC [3], SciREX [4], TDMci [6], SoMeSci [7], GSAP-NER [8], SciDMT [9] などがあり，これらを用いて手法が開発されている．また，研究データの種別や作成者を論文から獲得する研究も存在する [10, 11, 12]．抽出手法としては，fine-tuning された BERT [5] や SciBERT [13] に基づく手法や，LLM を用いた手法などが採用されている．

本研究では，同一の研究活動で産出された論文と研究データの対応関係に着目する．研究データに対応する論文は，その研究データのメタ情報の獲得に有用な情報源である．特に，研究データの作成方法に関わる情報は，対応する論文に含まれやすいと考えられる．本論文の対応付け技術と論文からメタ情報を獲得する技術を組み合わせることは，研究データのメタデータ生成に向けた有効な方略といえる．

2.2　研究データの URL 引用とその利用

論文中の URL を手がかりに，研究データに関する情報の獲得を試みた研究も存在する．論文で研究データに言及する際に，その公開先 URL が記載されることがある．一方で，研究データへの言及以外の場面でも URL が記載されることがあり，論文中の URL を研究データの公開先であるか否かに分類する手法 [14] が提案されている．

URL で引用された研究データについて，その種別を推定する研究も存在する [15, 16, 17]．種別として，既存の研究データの利用（Use）や拡張（Extend）のほか，新たな研究データの作成（Produce）などがある．このうち，Produce タイプの引用を特定することは，同一の研究活動で産出された論文と研究データを対応付けることに相当する．

2.3　論文と研究データの対応付け

同一の研究活動で産出された論文と研究データが対応付けられ，相互に参照可能であることは，双方の理解を深めるうえで重要となる．これまでに，論文と研究データの対応関係が登録されたプラットフォームが開発されており，その 1 つが，Papers With Code (PWC) [1)2)] である．PWC には，主に機械学習分野の論文と研究データ（コードやデータセットなど）のメタデータが人手で登録され，その中には，論文と研究データの対応関係も含まれている．論文と研究データを自動で対応付けることは，このようなプラットフォームの構築や拡張に寄与すると考えられる．

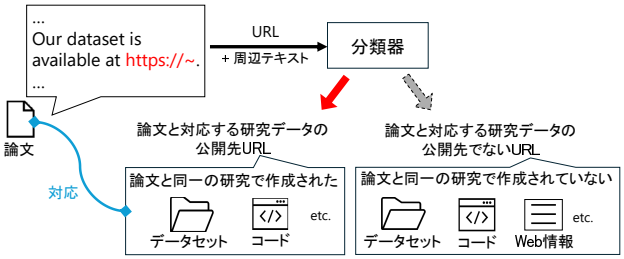

図 2　論文と研究データの対応付け手法の概要

3　論文と研究データの対応付け手法

同一の研究活動で産出された論文と研究データを対応付けるために，論文に対応する研究データの公開先 URL を論文から獲得することが考えられる．しかし，論文に記載される URL は，対応する研究データの公開先とは限らず，Web 上の情報や利用した研究データの URL が記載されることがある．

本手法では，論文中の URL を，その周辺テキストを利用して，対応する研究データの公開先であるか否かに分類する．図 2 に，本手法の概要を示す．論文中の URL の出現位置は，本文（Body），脚注（Footnote），参考文献リスト（Reference）の 3 つに大別される [16]．本手法では，脚注あるいは参考文献リスト中の URL に対しては，本文中の脚注タグあるいは参考文献の引用タグ周辺の文章も，URL の周辺テキストとして利用する．それぞれの周辺テキストの例を表 1 に示す．

4　実験

3 章で述べた対応付け手法，すなわち，論文中の URL 分類手法の性能を実験的に評価した．

1)　https://github.com/paperswithcode/paperswithcode-data

2)　PWC は現在は稼働を終了しており，その後継として Hugging Face Trending Papers（https://huggingface.co/papers/trending）が稼働中である．

表 1　本文，脚注，参考文献リストに出現する URL の周辺テキストの例

		URLの出現位置		
		Body	Footnote	Reference
URLの周辺テキスト	節タイトル	Dataset	Abstract	Related work
	URL周辺	SCIERC extends previous datasets in scientific articles SemEval 2017 Task 10 (SemEval 17) (Augenstein et al., 2017) and SemEval 2018 Task 7 (SemEval 18) (Gabor et al., 2018) by extending entity types, relation types, relation coverage, and adding cross-sentence relations using coreference links. Our dataset is publicly available at: http://nlp.cs.washington.edu/sciIE/. Table 1 shows the statistics of SciERC.	Experiments show that our multi-task model outperforms previous models in scientific information extraction without using any domain-specific features. We further show that the framework supports construction of a scientific knowledge graph, which we use to analyze information in scientific literature.1	This method samples a diverse set of reasoning paths from a language model via reasoning traces prompting and returns the most consistent final answer in the set. Other work evaluates the diversity of a reasoning path (Li et al., 2022), or the consistency of an inference step (Creswell et al., 2022) or finetune LLMs (Zelikman et al., 2022) to improve on difficult NLP tasks. In contrast to these works, we present a suit of metrics that focus on determining the type of the error (e.g., commonsense or logical inconsistency) in a reasoning path, if one exists.
	脚注文または参考文献の書誌情報		1 Data and code are publicly available at: http://nlp.cs.washington.edu/sciIE/	Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. arXiv, 2022. URL https://arxiv.org/abs/2205.09712

4.1　実験データ

Papers With Code (PWC) の稼働終了時点のアーカイブ[3]を利用して実験データを作成した．作成方法の概要を図 3 に示す．PWC には機械学習分野の学術論文のメタデータなどが登録されており，その中には同一の研究活動で産み出された研究データの URL も含まれる．研究データには様々な種別が存在するが，本研究ではデータセットを対象とした．

実験データを以下の手順で作成した．

1. PWC に登録されている論文 PDF の取得
2. PDF ファイルのテキスト化
3. URL とその周辺テキストの抽出
4. URL の一致判定

手順 4. では，抽出された URL が，当該論文の PWC に登録されている研究データの公開先 URL と一致する場合を正例とし，一致しない場合を負例とした．実験データの作成方法の詳細を付録 A に示す．

実験データには，2,638 件の論文から抽出した，URL とその周辺テキスト 17,977 件が含まれている．そのうち，論文と同一の研究活動で産出された研究データの公開先である URL は，17.72% (3,185/17,977) であった．

4.2　分類手法の実装

分類手法として，事前学習済みの BERT [5] ベースのエンコーダモデルに，分類のための出力層を追加するアプローチを採用した．入力に用いた URL 周辺テキストとして，その URL が出現した節のタイトル，および，その URL が含まれる文とその前後 1 文（計 3 文）を用いた[4]．脚注あるいは参考文献に出現した URL は，本文上の脚注タグあるいは引用タグの位置を URL の出現位置とし，脚注文あるいは参考文献の書誌情報も入力として利用した．

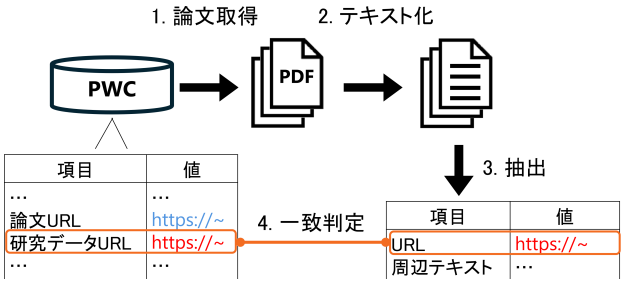

図 3　実験データの作成方法の概要

モデルへの具体的な入力として，節タイトル，URL 周辺の 3 文，脚注文あるいは参考文献の書誌情報の計 3 種の素性を，[SEP] トークンで連結したものを与えた．分類対象となる URL の出現位置を明示するため，本文中の URL の直後と参考文献の引用タグの直後に [CITE] という文字列を追加した．また，分類対象の URL を含む脚注の脚注タグを [CITE] に置換した．

エンコーダには BERT[5] と SciBERT [13] [6] を採用し，損失関数として cross entropy loss を用いて fine-tuning した．モデルの実装には，主に Pytorch[7] と Hugging Face の Transformers[8] を用いた．fine-tuning

3）https://huggingface.co/pwc-archive

4）前後 1 文の取得は同一段落内に制限した．

5）https://huggingface.co/google-bert/bert-base-uncased

6）https://huggingface.co/allenai/scibert_scivocab_uncased

7）https://pytorch.org/

8）https://github.com/huggingface/transformers

表 2　実験結果

	再現率	適合率	F 値
Random	0.210	0.127	0.158
BERT	0.780	0.683	0.729
SciBERT	0.801	0.690	0.740

表 3　正しく抽出できた研究成果の公開先 URL とその周辺テキストの事例

節タイトル	Conclusion
URL周辺	We hope that JEEBench guides future research in reasoning using LLMs. Our code and dataset are available at https://github.com/dair-iitd/jeebench[CITE].

では，Patience を 5 に設定し，early stopping を適用した．最適化手法には AdamW [18] を採用し，学習率とバッチサイズはそれぞれ 1e-5 と 32 とした．その他のハイパーパラメータは，実装に使用したライブラリのデフォルト値を利用した．

4.3　評価方法

実験データは，論文の発行年に基づいて，8:1:1 で学習，検証，テストに分割した．具体的には，発行年が最新のものから順に 10%をテスト用に割り当て，残りをランダムに学習用と検証用に割り当てた．その結果，学習，検証，テストに含まれる URL とその周辺テキストのペアは，それぞれ，13,715 件，1,635 件，2,627 件となった．

分類手法の性能については，テストデータにおける再現率，適合率，F 値で評価した．本実験では，シード値を変更して fine-tuning を 3 回実施し，各試行における分類性能の平均値を，最終的な分類性能とした．また，URL 周辺のテキストを入力として分類を行う手法の比較対象として，学習データの正例と負例の比率に基づきランダムに分類を行う手法（Random）を実装し，その性能を評価した．

4.4　実験結果

実験結果を表 2 に，BERT と SciBERT の両者が正しく抽出できた研究データの公開先の URL とその周辺のテキストの事例を表 3 に，それぞれ示す．BERT と SciBERT の F 値はそれぞれ 0.729 と 0.740 であり，本手法によって一定の水準で論文との対応付けが可能であることが示唆された．また，BERT と SciBERT の F 値は共に，ランダムに分類を行った際の F 値を有意に上回った[9]．以上より，本手法は論文中の URL と対応する研究データ公開先 URL の対応付けに有効であり，ランダム分類と比較しても優れていることが確認された。

表 4　URL の出現位置ごとの分類性能

モデル	出現位置	再現率	適合率	F 値
BERT	Body	0.870	0.686	0.766
	Footnote	0.728	0.591	0.651
	Reference	0.273	0.280	0.273
SciBERT	Body	0.866	0.714	0.780
	Footnote	0.846	0.571	0.681
	Reference	0.273	0.316	0.291

4.5　エラー分析

本節では，本手法の分類性能に関するエラー分析として，URL の各出現位置（本文，脚注，参考文献リスト）における分類性能を評価する．各出現位置における再現率，適合率，F 値を 4.3 節と同様に算出した．BERT，SciBERT による結果を表 4 に示す．

本文中および脚注に含まれる URL については，両方のモデルにおいて，適合率が再現率よりも低い傾向にあった．本文および脚注に出現する URL の分類性能改善のためには，研究データの公開先 URL の誤抽出を減らすことが重要であるといえる．

両方のモデルで，参考文献リストに含まれる URL の F 値は 0.3 以下にとどまり，低い値となっている．これは，参考文献リストの URL に占める，対応する研究データの公開先を示したものが少なく，十分に学習できていないためであると考えられる．

5　まとめ

本論文では，同一の研究活動で産出された論文と研究データの自動対応付けについて述べた．本研究では，論文中の URL を，その周辺のテキストを利用して，対応する研究データの公開先であるか否かに分類する手法を採用した．本手法の分類性能を評価するために実験を行ったところ，BERT ベースのモデルを fine-tuning することで実装された本手法の分類性能を確認した．

本手法では，対応する研究データの公開先 URL が論文に記載されている必要がある．しかし，必ず URL が記載されるとは限らず，例えば，研究データの名称のみが記載されることもある．今後は，このような場合にも適用可能な手法を検討したい．

9)　ボンフェローニ補正を適用したマクネマー検定を行い，3 回の試行の全てにおいて有意差が認められた（$p < 0.05$）．

謝辞

本研究は，JSPS 科研費 JP25K03418 の助成，文部科学省「AI 等の活用を推進する研究データエコシステム構築事業」の支援を受けたものです．

参考文献

[1] UNESCO. Open science, （2025-12-15 閲覧）. https://www.unesco.org/en/open-science.

[2] Annex 1: G7 open science working group (OSWG), （2025-12-10 閲覧）. https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/annex1_os.pdf.

[3] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities and relations and and coreference for scientific knowledge graph construction. **In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3219–3232, 2018.

[4] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. **In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7506–7516, 2020.

[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**, pp. 4171–4186, 2019.

[6] Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. **In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 707–714, 2021.

[7] David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. SoMeSci- A 5 star open data gold standard knowledge graph of software mentions in scientific articles. **In Proceedings of the 30th ACM International Conference on Information & Knowledge Management**, pp. 4574–4583, 2021.

[8] Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar, and Stefan Dietze. GSAP-NER: A novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. **In Proceedings of the findings of the Association for Computational Linguistics**, pp. 8166–8176, 2023.

[9] Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. SciDMT: A large-scale corpus for detecting scientific mentions. **In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, pp. 14407–14417, 2024.

[10] Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. Capabilities and challenges of LLMs in metadata extraction from scholarly papers. **In Proceedings of the 26th International Conference on Asia-Pacific Digital Libraries**, Vol. 15493, pp. 280–287, 2024.

[11] Zaid Alyafeai, Maged S. Al-shaibani, and Bernard Ghanem. MOLE: Metadata extraction and validation in scientific papers using LLMs. **In Findings of the Association for Computational Linguistics**, pp. 12236–12264, 2025.

[12] Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. Empowering knowledge discovery from scientific literature: A novel approach to research artifact analysis. **In Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software**, pp. 37–53, 2023.

[13] Kyle Lo Iz Beltagy and Arman Cohan. SciBERT: A pretrained language model for scientific text. **In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 3615–3620, 2019.

[14] Masaya Tsunokake and Shigeki Matsubara. Classification of URLs citing research artifacts in scholarly documents based on distributed representations. **In Proceedings of 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents**, pp. 20–25, 2021.

[15] He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. A context-based framework for modeling the role and function of on-line resource citations in scientific literature. **In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 5206–5215, 2019.

[16] Masaya Tsunokake and Shigeki Matsubara. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. **In Proceedings of the 1st Workshop on Information Extraction from Scientific Publications**, pp. 8–19, 2022.

[17] Kazuhiro Wada, Masaya Tsunokake, and Shigeki Matsubara. Classification of URL citations on scholarly papers using intermediate task training. **IEICE Transactions on Information and Systems**, Vol. E108-D, No. 10, pp. 1183–1193, 2025.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **In International Conference on Learning Representations**, 2019.

[19] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. **ArXiv:2308.13418**, 2023.

[20] Patrice Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. **In Proceedings of the 13th European Conference on Digital Libraries**, pp. 473–474, 2009.

A　実験データの作成方法の詳細

本文中に含まれる URL について，URL の周辺テキストとして，節タイトル，段落，その URL を含む文を取得した．まず，Nougat [19] によって論文の PDF ファイルをテキスト化した．次に，ルールベースの手法を用いて，URL が含まれる節タイトル・段落を抽出した．その後，spaCy [10] の en_core_web_lg [11] を用いて文分割を行い，URL を含む文を抽出した．また，URL については，正規表現[12] に基づいて抽出した．

脚注に含まれる URL については，URL の周辺テキストとして，脚注文に加え，その脚注タグが含まれる節タイトルと段落，URL を含む文を抽出した．そのために必要となる脚注と脚注タグの対応付けは，脚注番号とその参照位置を利用して行った．その後，URL とその周辺テキストを，本文中に含まれる URL の場合と同様の手法で抽出した．

参考文献リストに出現する URL については，URL の周辺テキストとして，参考文献の書誌情報に加え，その引用タグが含まれる節タイトルと段落，URL を含む文を抽出した．そのために必要となる参考文献と引用タグの対応付けには，Grobid [20] を利用した．周辺テキストは論文 PDF を Grobid で処理したものから，ルールベースの手法を用いて，URL が含まれる節タイトル・段落を抽出した．URL を含む文については，本文中に含まれる URL の場合と同様に spaCy を利用して抽出した．

B　学習データ量の効果

ここでは，本手法の分類性能に与える学習データ数の効果を検証する．本研究で作成した実験データの特徴の 1 つとして，既存のリソースを利用して自動作成された大規模なデータである点が挙げられる．そのため，本実験データは，本手法の分類性能に与える学習データ数の効果を検証することに適しているといえる．具体的な検証方法として，fine-tuning に用いる学習データ数を 2,000 件ずつ増加させて，テストデータにおける分類性能を検証した．Fine-tuning 時の設定は，4.2 節と同じものを使用した．

結果を図 4 に示す．BERT と SciBERT のいずれにおいても，fine-tuning に用いた学習データ数が増加するほど，分類性能は概ね向上した．このことから，本研究で作成した実験データの規模をさらに大きくできれば，分類性能のさらなる向上が見込まれる．

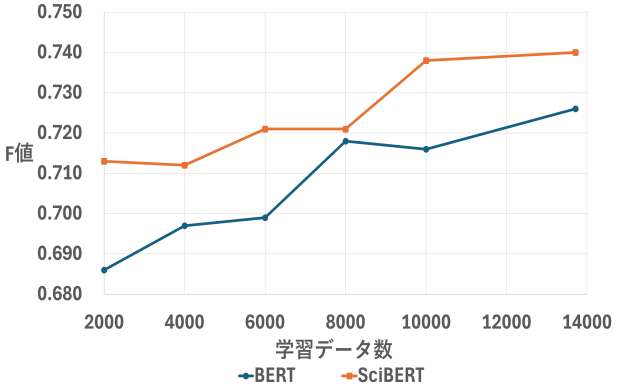

図 4　学習データ数と抽出性能の関係

10）https://spacy.io/
11）https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.7.1
12）http://，https://，ftp://から始まる文字列を URL とする．