URL 引用の要否判定において 学習データの品質とドメインが与える影響の分析

和田和浩 [1]　角掛正弥 [2]　松原茂樹 [1,2]
[1] 名古屋大学大学院情報学研究科　[2](株) 日立製作所研究開発グループ
[3] 名古屋大学情報基盤センター
wada.kazuhiro.s8@s.mail.nagoya-u.ac.jp

概要

研究活動において研究データの適切な引用は論文の信頼性のための重要な要素である．URL を参照識別子として研究データを引用すること（**URL 引用**）が多いが，これを対象とする引用要否判定にはデータセットの品質が低いことや分野が限定されているといった課題がある．本研究では，より高精度な OCR を用いて高品質な複数の分野からなるデータセットを作成し，学習データの品質とドメインの違いが URL 引用の要否判定の性能に与える影響を分析した．実験の結果，高品質なデータセットを用いることでモデルの F 値が 14.1%向上することを示した．また，異なるドメイン間でも高い性能を維持できることを明らかにした．

1　はじめに

研究活動において研究データは重要な役割を果たしている．論文における研究データの適切な引用は，研究データ[1)]のアクセス性の向上に加え，論文の信頼性も向上させる．そのため，論文執筆時や査読時の引用漏れ防止のために，これらの引用が適切に行われているかを確認する必要がある．このために論文内の文に対して引用が必要か否か判定する引用要否判定が取り組まれてきた [1, 2, 3, 4, 5, 6]．

これらの研究は主に文献タグ（e.g., (Michael et al., 2025)，[1]）を用いた引用が対象である．しかし，研究データの引用には文献タグだけでなく URL も用いられる（**URL 引用**）．論文では研究データを表す識別子として URL が最も使用されており [7, 8]，研究データについても引用の適切さを検証するため URL 引用を対象とした引用要否判定を行う必要がある．URL 引用を対象とした要否判定を行った研究はあるものの [9]，以下の課題が存在する．

1. 論文 PDF のテキスト化に精度の低いツールを使用しておりデータセットの品質が低い
2. 自然言語処理の分野に限定されており，他のドメインでの性能，ドメイン間の関係性が不明

そこで本論文では，データセットの品質とドメインの違いが URL 引用の要否判定の性能に与える影響を分析する．検証のために先行研究と比較してより高精度な OCR を採用し，自然言語処理，天体物理学，電気工学・システム科学の 3 分野からなる URL 引用要否判定向けのデータセットを作成した．URL 引用の要否判定において重要である URL の抽出精度と脚注文と本文中の脚注番号の対応付けの正確さを先行研究のデータセットと比較した結果，作成したデータセットの品質がより高いことを確認した．

次に，データセットの品質が性能に与える影響を調べるために，先行研究と本研究のデータをそれぞれ学習データとした際の判定性能を比較した．その結果，本研究で作成した質の高いデータセットで学習したモデルの方が 14.1%高い F 値を示した．また，ドメインの違いが性能へ与える影響を明らかにするために，学習と評価データのドメインを変えたところ，自然言語処理分野以外のドメインにおいても高い性能で要否を判定することができることに加え，学習と評価データのドメインが異なる場合でも，同じドメインを使用した場合に匹敵する性能を発揮可能であることがわかった．

2　関連研究

文献タグを用いた引用に関する要否判定は Sugiyama が提案し [2]，多くの研究が行われている [1, 3, 4, 5, 6, 10]．引用要否判定は判定したい文を入力とし，その文に引用が必要か否かを判定する 2 値

1) データセットやソフトウェア，モデルなど研究の過程，結果として収集，生成される情報：https://www.nii.ac.jp/service/upload/docs_rdm_week1-3_2017.pdf

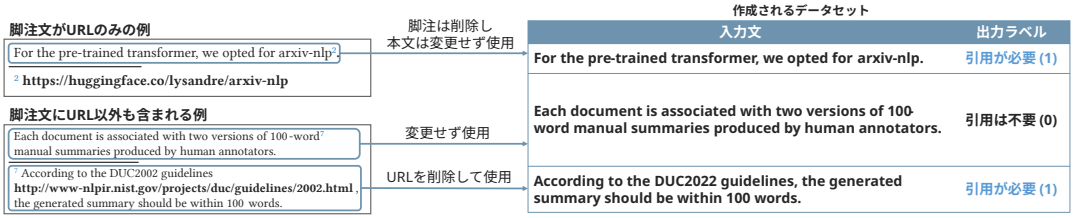

図 1　脚注文の内容によるデータセットの作成時の出力ラベルの違い

分類タスクである．先行研究の多くは学術論文を使用し，文献タグを含む文を引用が必要な文，含まない文を引用が不要な文として自動でデータセットを作成している．その中でも，Wright らは文献タグの削除時に発生する文法上誤った文に対処し，データセットの質を向上させた [11]．また，Buscaldi らは小規模だが人手で要否のラベルを付与している [1]．多くの研究は自然言語処理分野を対象としているが [4, 2, 3]，他のドメインでも広く取り組まれている [5, 6, 1, 10]．Zeng らや Vajdeˇcka らは PubMed から収集した論文から医学分野を対象としたデータセットを作成している [5, 6]．また，工学や物理学など多様なドメインの論文が含まれる Arxiv を利用したデータセットも作成されている [1, 6]．

一方，URL 引用の要否判定も行われており文献タグに加えて URL 引用も判定対象とした要否判定が取り組まれている [9]．その際，データセットの作成には URL 引用の性質上，脚注文と本文中の脚注番号の対応付けが必要であるが（図 1），使用されている論文のテキスト化ツールの性能が低く，これがデータセットの質に悪影響を与えている可能性がある．ドメインも自然言語処理分野に限定されており，その他の分野についての研究はされていない．

3　引用要否判定

本研究では，多くの先行研究で採用されている，判定対象の 1 文に対してその文に引用が必要か否かを判定するタスク設定を採用する．具体的には，判定対象の文 $\boldsymbol{s}$ に対し，モデル f の出力 $\hat{y}$ は

$$\hat{y} = f(\boldsymbol{s}) \ (\hat{y} \in \{0, 1\})$$

で計算される．ここで，$\hat{y} = 1$ は引用が必要であることを，$\hat{y} = 0$ は不要であることを示す．本研究で検証する手法ではモデル f は事前学習済みのエンコーダと分類層を組み合わせた一般的なテキスト分類モデルを採用する（図 2）．まず，判定する文 $\boldsymbol{s}$ をエンコーダ（e.g., BERT [12]，RoBERTa [13]）に入力し，最初のトークン（[CLS]）に対応する出力ベクトル $\boldsymbol{h}_{\mathrm{CLS}} \in \mathbb{R}^d$ を取得する．これは次式で表される：

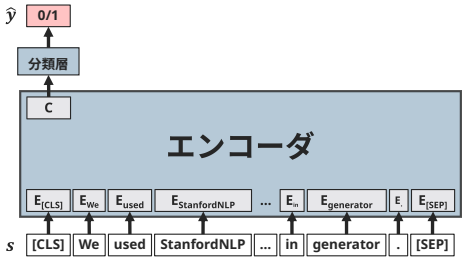

図 2　モデル図

$$\boldsymbol{h}_{\mathrm{CLS}} = \mathrm{Encoder}(\boldsymbol{s})$$

次に，得られたベクトル $\boldsymbol{h}_{\mathrm{CLS}}$ を分類層に入力し，引用の要否を表す 2 値のスコア $\boldsymbol{o} \in \mathbb{R}^2$ を計算する：

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{h}_{\mathrm{CLS}} + \boldsymbol{b}$$

ここで，$\boldsymbol{W} \in \mathbb{R}^{2\times d}$ は重み行列，$\boldsymbol{b} \in \mathbb{R}^2$ はバイアス項である．スコア $\boldsymbol{o}$ にソフトマックス関数を適用し，引用が必要である確率 $p(\hat{y} = 1|\boldsymbol{s})$ を計算する：

$$p(\hat{y} = 1|\boldsymbol{s}) = \frac{\exp(o_1)}{\exp(o_0) + \exp(o_1)}$$

最後に，確率 $p(\hat{y} = 1|\boldsymbol{s})$ が 0.5 以上であれば引用が必要であると判断し，$\hat{y} = 1$ とする．それ以外の場合は引用が不要であると判断し，$\hat{y} = 0$ とする．

$$\hat{y} = \begin{cases} 1 & \text{if } p(\hat{y} = 1|\boldsymbol{s}) \geq 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

4　データセットの作成

より高精度なテキスト化ツールを用いて複数のドメインからなるデータセットを作成した．先行研究の自然言語処理に加え，URL 引用がある程度行われており，互いに分野が離れている天体物理学と電気工学・システム科学の 3 分野を対象とした．自然言語処理は先行研究 [9] に従い 2000~22 年の ACL, NAACL, EMNLP の本会議の論文を ACL Anthology [2] から，天体物理学は 2023 年，電気工学・システム科学は 2022，23 年の論文を ArXiv [3] から収集した．

2）https://aclanthology.org/
3）https://arxiv.org/

表 1 各ドメインのデータセットの統計．括弧内の数値は全体に占める割合を表す．NLP は自然言語処理分野，AST は天体物理学，ENG は電気工学・システム科学を表す．

ドメイン	論文数	引用が必要な文	引用が不要な文	学習・検証データ	評価データ	合計
NLP	20,576	37,928 (0.89%)	4,245,504 (99.11%)	3,422,676	860,756	4,283,432
AST	17,912	29,681 (0.49%)	5,983,816 (99.51%)	4,997,191	1,016,306	6,013,497
ENG	33,032	25,560 (0.32%)	7,940,926 (99.68%)	6,856,554	1,109,932	7,966,486

4.1 作成手順

各ドメインについて先行研究 [9] に従い，以下の手順でデータセットを作成した．

1. 論文 PDF のテキスト化
2. 文分割
3. 脚注番号の対応付け
4. URL 引用の検出

手順 4 では PDF から作成した文集合を引用が必要な文（$\hat{y} = 1$）と不要な文（$\hat{y} = 0$）に区分する．URL 引用の要否判定では URL を含む文を引用が必要な文とする．基本，URL を含む文を引用が必要な文とするが，URL が脚注に書かれた場合は脚注文に応じて異なる扱いが必要である（図 1）．脚注文が URL のみである場合，その脚注に対応する脚注番号を含んだ本文中の文を引用が必要な文とする．これは URL のみの脚注が本文中で括弧書きで行われる URL 引用と本質的に同一であるためである．このために脚注と脚注番号の対応付けを手順 3 で行う．

先行研究は論文 PDF のテキスト化に PDFNLT-1.0 [14] を用いていたが，本研究では Nougat [15] を使用する．Nougat は高いテキスト化性能に加え，脚注に対応する本文中の段落を絞り込めるために脚注番号との対応付け性能の向上が期待できる．ツールの変更に伴い脚注を対応付けるためのルールも調整した．また，誤ったデータや判定する意義の無い短い定型文（e.g., “Available at ~”）を除くために 3 単語未満の文を除外する．

4.2 作成したデータセットの統計

表 1 に作成したデータセットの統計を示す．どの分野でも引用が必要な文は全体の 1%未満であり，正例が少ない偏った分布をしている．

4.3 データセットの品質

作成したデータセットの品質を確認するために，先行研究のデータセットと URL の抽出精度と脚注番号の対応付けの性能を比較した．

表 2 先行研究（従来）と作成したデータ（新規）の URL の抽出の正解率．太字は最も高い数値を表す．

データセット	正解率
従来データ [9]	0.75
新規データ	**0.83**

表 3 先行研究（従来）と作成したデータ（新規）の脚注番号の対応付けの性能．太字は最も高い数値を表す．

データセット	適合率	再現率	F 値
従来データ [9]	0.901	0.918	0.909
新規データ	**0.984**	**0.946**	**0.965**

URL の抽出精度 表 2 に無作為に抽出した 15 論文における URL の抽出の正解率を示す．従来データ[4]では正解率が 0.75 だが，新規データでは 0.83 であり，より正確に URL を抽出できている．

脚注番号の対応付け 無作為に 51 論文を抽出し，脚注の対応付けの性能を確認した．各論文の脚注番号の位置を正解としたときの適合率，再現率，F 値を表 3 に示す．全ての指標で性能が向上しており，Nougat を用いることでデータセットの質が改善していると言える．これは Nougat の出力の性質上，対応付く脚注番号が段落まで絞り込めることと，テキスト化が正確なため対応付ける際のスコアの計算がより正確になったためと考えられる．

5 実験

データセットの品質とドメインの違いが性能に与える影響を明らかにするために実験を行った．

5.1 実験設定

学習設定 エンコーダは RoBERTa[5] [13] を採用した．最適化手法は AdamW [16] を学習率 4e-6，その他のパラメータは Huggingface Trainer のデフォルト

4） PDFNLT の出力に合わせた後処理が追加されている．
5） https://huggingface.co/roberta-base

表 4 自然言語処理分野において学習データを先行研究（従来）と作成したデータ（新規）にしたときの性能．下付き文字は標準偏差を，太字は最も高い数値を表す．

学習データ	適合率	再現率	F 値
従来データ [9]	$0.894_{0.023}$	$0.577_{0.016}$	$0.701_{0.010}$
新規データ	$\mathbf{0.946_{0.019}}$	$\mathbf{0.758_{0.010}}$	$\mathbf{0.842_{0.002}}$

値[6]とした．また引用が必要なクラス（1）の F 値を指標として Early Stopping（patience は 2）を適用した．バッチサイズは 32，勾配累積のステップ数は 8 である．論文の出版年が新しいものから 100 万件程度を評価データ，残りを 9:1 の割合で分割したものを学習，検証データとした．各学習データについて，異なるシード値で 5 回実験を行い平均を取った．

評価指標 本タスクではデータセットの分布が偏っているため，引用が必要なクラスについての適合率，再現率，F 値を評価指標として使用する．引用要否判定タスクの評価には，通常これに加えて重み付きの F 値も使用するが，データセットの偏りが非常に大きいことで常に 1.0 になるため省略する．

5.2 学習データの品質がモデルの判定性能へ与える影響

表 4 に自然言語処理分野について学習データを先行研究（従来データ）と本研究で作成したもの（新規データ）としたときの結果を示す．評価データはより質の高い，新規データを使用した．全ての指標において本研究のデータセットを学習データとした手法が最も高い数値を示しており，特に再現率が 18.1%，F 値が 14.1%と大幅に向上している．この結果は学習データの品質は URL 引用の要否判定において大きな影響を与えることを示している．

5.3 他ドメインにおける要否判定の性能とドメイン間の転移の可能性

表 5 に学習と評価データのドメインを変えた結果を示す．まず，学習と評価データのドメインが同じ場合について，全てのドメインにおいて F 値が 0.8 以上と高い数値となっている．どのドメインにも URL 引用の形態には一定のパターンがあり，そのパターンを捉えて要否判定が十分に実行できたと考えられる．また，適合率が 0.9 以上と高く，再現率が低い傾向は全体で共通している．これはデータセットのほとんどが引用が必要ない文であり，分布が大きく偏っているためであると考えられる．ドメイン別では NLP，AST，ENG の順に F 値が高く，データセットの標本数が少ない順と一致しているため，学習データのサイズが影響した可能性がある．

次に，学習と評価データのドメインが異なる場合について，どの組み合わせにおいても F 値が 0.8 を超えておりドメインが同じ場合と匹敵する性能を維持している．この結果は URL 引用の形態はドメインによる差異が少ないことを示唆しており，単一ドメインで学習したモデルだけでドメインを横断して要否判定を行うことができる可能性を示している．

表 5 学習・評価データのドメインを変えたときの性能．下付き文字は標準偏差，太字は評価データが同じものの中で最も高い数値を表す．NLP は自然言語処理分野，AST は天体物理学，ENG は電気工学・システム科学を表す．

学習	評価	適合率	再現率	F 値
NLP	NLP	$0.946_{0.019}$	$\mathbf{0.758_{0.010}}$	$\mathbf{0.842_{0.002}}$
AST	NLP	$\mathbf{0.950_{0.008}}$	$0.728_{0.003}$	$0.824_{0.003}$
ENG	NLP	$\mathbf{0.950_{0.009}}$	$0.737_{0.004}$	$0.830_{0.002}$
NLP	AST	$0.916_{0.032}$	$0.761_{0.008}$	$0.831_{0.009}$
AST	AST	$0.966_{0.004}$	$\mathbf{0.775_{0.005}}$	$\mathbf{0.860_{0.002}}$
ENG	AST	$\mathbf{0.971_{0.006}}$	$0.745_{0.004}$	$0.843_{0.002}$
NLP	ENG	$0.926_{0.026}$	$0.826_{0.005}$	$0.873_{0.009}$
AST	ENG	$0.948_{0.006}$	$0.834_{0.006}$	$0.887_{0.003}$
ENG	ENG	$\mathbf{0.971_{0.005}}$	$\mathbf{0.847_{0.003}}$	$\mathbf{0.904_{0.001}}$

6 おわりに

本論文では学習データの品質とドメインの違いが URL 引用の要否判定の性能に与える影響を調査した．まず，高精度な PDF のテキスト化ツールである Nougat を用いて，自然言語処理，天体物理学，電子工学・システム科学の計 3 分野のデータセットを作成した．また，自然言語処理分野について先行研究と作成したデータセットを比較し，URL の抽出性能と脚注番号の対応付けがより正確であることを確認した．次に，先行研究と本研究のデータをそれぞれ学習データとした際の判定性能を比較した結果，本研究で作成したデータセットで学習したモデルの方が 14.1%高い F 値を示し，学習データの品質が URL 引用の要否判定に大きな影響を与えることがわかった．加えて，ドメイン間で性能を比較した結果，学習と評価データのドメインが異なる場合でもモデルの判定性能が維持され，ドメインを横断して URL 引用の要否判定が行うことができることを示した．

6） https://huggingface.co/docs/transformers/v4.36.1/en/main_classes/optimizer_schedules#transformers.AdamW

謝辞

本研究は JSPS 科研費 JP23K18506 の助成を受けたものです.

参考文献

[1] Davide Buscaldi, Danilo Dessí, Enrico Motta, Marco Murgia, Francesco Osborne, and Diego Reforgiato Recupero. Citation prediction by leveraging transformers and natural language processing heuristics. **Information Processing & Management**, Vol. 61, No. 1, p. 103583, 2024.

[2] Kazunari Sugiyama, Tarun Kumar, Min-Yen Kan, and Ramesh C. Tripathi. Identifying citing sentences in research papers using supervised learning. In **Proceedings of the 2010 International Conference on Information Retrieval & Knowledge Management (CAMP)**, pp. 67–72, 2010.

[3] Hamed Bonab, Hamed Zamani, Erik Learned-Miller, and James Allan. Citation worthiness of sentences in scientific reports. In **Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)**, pp. 1061–1064, New York, NY, USA, 2018. Association for Computing Machinery.

[4] Rakesh Gosangi, Ravneet Arora, Mohsen Gheisarieha, Debanjan Mahata, and Haimin Zhang. On the use of context for predicting citation worthiness of sentences in scholarly articles. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**, pp. 4539–4545, Online, June 2021. Association for Computational Linguistics.

[5] Tong Zeng and Daniel E Acuna. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. **Scientometrics**, Vol. 124, No. 1, pp. 399–428, July 2020.

[6] Peter Vajdecka, Elena Callegari, Desara Xhura, and Atli Ásmundsson. Predicting the presence of inline citations in academic text using binary classification. In **Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)**, pp. 717–722, Tórshavn, Faroe Islands, May 2023. University of Tartu Library.

[7] Mengnan Zhao, Erjia Yan, and Kai Li. Data set mentions and citations: A content analysis of full-text publications. **Journal of the Association for Information Science and Technology**, Vol. 69, No. 1, pp. 32–46, 2018.

[8] Min Sook Park and Hyoungjoo Park. An examination of metadata practices for research data reuse: Characteristics and predictive probability of metadata elements. **Malaysian Journal of Library & Information Science**, Vol. 24, pp. 61–75, 2019.

[9] 和田和浩, 角掛正弥, 松原茂樹. 論文における URL による引用を考慮した引用要否判定. 言語処理学会 第 30 回年次大会, pp. 2258–2262, 2024.

[10] Mann Khatri, Reshma Sheik, Pritish Wadhwa, Gitansh Satija, Yaman Kumar, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. CiteCaseLAW: Citation Worthiness Detection in Caselaw for Legal Assistive Writing. Vol. 379, pp. 257–262, 2023.

[11] Dustin Wright and Isabelle Augenstein. CiteWorth: Cite-worthiness detection for improved scientific document understanding. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 1796–1807, Online, August 2021. Association for Computational Linguistics.

[12] 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] 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:1907.11692**, 2019.

[14] Takeshi Abekawa and Akiko Aizawa. SideNoter: Scholarly paper browsing system based on PDF restructuring and text annotation. In **Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations**, pp. 136–140, Osaka, Japan, December 2016. COLING 2016 Organizing Committee.

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

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