論文における URL による引用を考慮した引用要否判定

和田和浩 [1]　角掛正弥 [2]　松原茂樹 [1,3]
[1] 名古屋大学大学院情報学研究科　[2] 日立製作所　[3] 名古屋大学情報基盤センター
wada.kazuhiro.s8@s.mail.nagoya-u.ac.jp　masaya.tsunokake@gmail.com
matubara@nagoya-u.jp

概要

論文執筆や査読支援に向けて，ある文に引用が必要か否かを判定する引用要否判定が行われている．既存研究は文献タグによる引用のみを対象としているが，学術論文では URL による引用（以下 **URL 引用**）も行われる．データセットなどの引用は URL 引用で実施されることが多く，それらの引用も引用要否判定の対象に含める必要がある．本論文では，文献タグによる引用に加え，URL 引用を対象とした引用要否判定を提案する．データセットを作成し[1)]，既存手法の性能を確認した結果，URL 引用に限った再現率は低く，改善の余地があることが判明した．また，URL 引用に特有の言い回しへの対応が課題であることが分かった．

1　はじめに

学術論文において，引用は先行研究を尊重し，著者の研究の立ち位置を明確にするために重要な行為である．加えて，論文の読者の理解や関連するリソースへのアクセスを支援するためにも，引用は適切に行われる必要がある．しかし，適切な引用が行われているか否かを確認する作業には熟練した技術が必要なうえ，近年の論文数の増加に伴い確認作業に要する時間も肥大化している．

論文の執筆・査読支援を促進するため，ある文に引用が必要か否かを判定する引用要否判定の研究が行われている [1, 2, 3]．引用要否判定は，引用の抜け漏れの防止・検出を可能とし，執筆・査読の双方を支援できる重要なタスクである．既存研究の引用要否判定で対象としている引用は主に，文献リストに書誌情報として記される文献の引用である．[1, 2, 3].

しかし，学術論文では文献以外にもデータセットやプログラムなど，様々な研究資源が引用される．そして，それらは本文中に文献タグによって引用されるとは限らない．特に，データセットやプログラム，ツールなどは URL によって引用（以下 **URL 引用**）されることが多い [4, 5]．したがって，文献タグによる引用のみが対象である従来の引用要否判定では，文献以外の引用を十分に考慮できない．一方で，論文の読者による研究の理解や再現，リソースへのアクセスを支援するためには，データセットやプログラム等も適切に引用されることが望ましい．

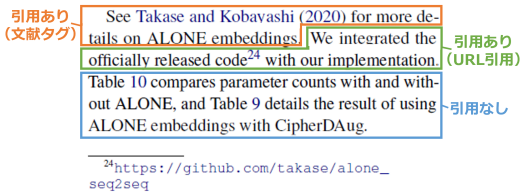

図 1　引用要否判定（本文は [6] より引用）

そこで本論文では，文献タグによる引用に加えて，URL 引用も含めた引用要否判定タスク（図 1）を提案する．具体的には，既存研究で「引用あり」とされていた文集合に，URL 引用が存在する文を加える．これにより，従来の引用要否判定では十分に考慮できていなかった引用も網羅できる．

このタスク設定に従い，ACL Anthology[2)] より収集した ACL，NAACL，EMNLP の本会議における論文からデータセットを作成した．従来の引用要否判定の手法から，BERT[7] などの文脈埋め込みモデルを使用した Gosangi らの手法 [3] をベースラインとして，URL 引用を含めた引用要否判定における性能を確認した．その結果，URL 引用の引用要否判定では，前後の文脈が重要であることが分かった．また，引用の種類を区別せず評価した場合は「引用あり」の文の F 値は高い一方で，URL 引用に限った再現率は低い傾向にあり，URL 引用の要否判定には改

1)　以下のリポジトリでデータセットを公開する．https://github.com/matsubara-labo/URL_Citation_Worthiness

2)　https://aclanthology.org/

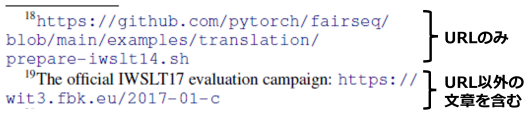

図 2　脚注での URL 引用（脚注は [6] より引用）

善の余地があることが判明した．加えて，URL 引用を誤って判定した文を分析した結果，データセットやツールを表す固有名詞や URL 引用に特有の言い回しへの対応が課題であることが分かった．

2　関連研究

引用要否判定は，ある文が引用を必要とするか否かを判定するタスクであり，杉山らによって提案された [8]．このタスクに対する取り組みは多くあり，Bonab らは ACL-ARC[9] を基に引用要否判定用のデータセットである SEPID-cite を作成し，Word2Vec[10] と CNN を用いた手法を提案した [1]．Gosangi らは引用要否判定に段落レベルの広い文脈が有用であると主張し，広い文脈情報を加えた ACL-cite を作成した [3]．

これらの研究は主に言語処理分野やコンピュータサイエンス分野を対象にしているが，他の分野でも研究されている [2, 11]．Zeng らは医学分野の論文データセットである PubMed[3] を使用して PMOA-cite を作成した [2]．また，Khatri らは Caselaw Access Project[4] で提供されているデータを使用してアメリカの法律文書に対して引用要否判定を行った [11]．

3　URL 引用を含めた引用要否判定

本節では，本研究で提案する URL 引用も含めた引用要否判定タスクについて述べる．このタスクでは文ごとに引用が必要か否かの 2 値分類を行う．具体的には，文献タグによる引用，または URL 引用が行われた文を「引用あり」として検出することを目指す．図 1 に例を示す．1 文目は文献タグによる引用であり，従来の引用要否判定においても検出対象の引用である．一方，2 文目は URL 引用が行われた例であるが，従来の引用要否判定では検出対象としていない．本論文の引用要否判定では，このような文を「引用あり」として検出すべき文に追加する．[5]

URL 引用は主に本文，脚注，参考文献の 3 か所で行われる [4]．参考文献における URL 引用は文献タグによる引用と重複するため，本文と脚注での URL 引用のみを扱う．本文での URL 引用では，本文中の URL を含む文を「引用あり」とする．脚注での URL 引用では，URL が記載された脚注の文章によって，「引用あり」として検出対象にする文が異なる．図 2 の上側のように脚注に URL のみ記載されている場合，本文中の脚注番号を含む文を「引用あり」とする．なぜなら，本文の脚注番号が脚注の URL のみを参照しているため，本文での URL 引用と本質的には同じであり，脚注番号のある文を引用箇所と考えることができるためである．一方，図 2 の下側のように脚注に URL 以外も記載されている場合，脚注番号は URL 以外の文章も参照しており，本文での URL 引用と同じであるとはみなし難い．そのため，本文中の脚注番号を含む文ではなく，脚注において URL が記載された文を「引用あり」として検出すべき対象にする．

4　データセット

4.1　作成手順

以下の手順に従って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. 文献タグ，URL の検出

論文 PDF のテキスト化に PDFNLT-1.0[12] を使用した．文分割に Spacy[6] の en_core_web_lg[7] を使用した．本研究で提案する引用要否判定タスクでは脚注に URL 以外の文章も含まれるか否かによって，脚注番号が対応している本文を「引用あり」とする場合がある．一方，PDFNLT では脚注と本文の脚注番号の対応付けは行われない．そのため，本文の脚注番号と脚注の対応付けを行う．本文中の脚注番号の候補から前後の語，単語，ページ番号などを使用してスコアを計算し，最も高いスコアであったものを対応する脚注番号とした．手順 4 は，「引用あり」として検出対象とする文を判定するために実施する．文献タグ[8] と URL[9] の検出には正規表現を用いた．

3）https://pubmed.ncbi.nlm.nih.gov/
4）https://case.law/
5）文献タグによる引用，URL 引用がともに行われた場合も「引用あり」として検出対象とする．
6）https://spacy.io/
7）https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.7.1
8）Gosangi らの用いた正規表現 [3] を参考にした．
9）http，https，ftp，www から始まる文字列を URL とする．

表 1　データセットの比較（括弧内の数値は全体の文数に占める割合を表す．[3] より引用し，一部改変）

	SEPID-cite[1]	PMOA-cite[2]	ACL-cite[3]	作成したデータセット
論文数	10,921	6,754	17,440	12,942
文数	1,228,053	1,008,042	2,706,792	2,277,834
引用を含まない文	114,275	811,659	2,401,059	1,899,672
引用を含む文（合計）	85,778(0.075)	196,383(0.195)	305,733(0.127)	378,162(0.166)
引用を含む文（文献タグ）	85,778(0.075)	196,383(0.195)	305,733(0.127)	360,036(0.158)
引用を含む文（URL 引用）	0	0	0	22,693(0.010)
1 文当たりの文字数	131	132	141	141
1 文当たりの単語数	22	20	22	23

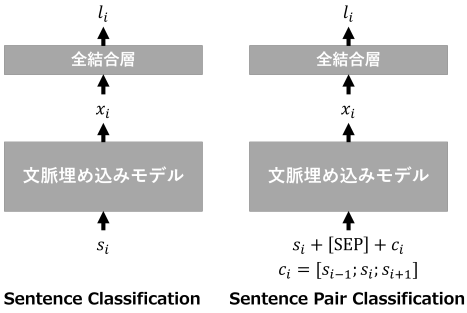

図 3　ベースラインのモデル図

表 2　ラベルごとの内訳

		文数	割合	文字数	単語数
引用なし		1,899,672	0.834	138	23
引用あり	文献タグのみ	355,469	0.156	157	25
	URL 引用のみ	18,126	0.008	130	20
	文献タグ +URL 引用	4,567	0.002	187	28

4.2　作成したデータセット

2000 年以降の ACL，NAACL，EMNLP の本会議における論文を ACL Anthology より収集し，データセットを作成した．

4.2.1　脚注番号の対応付けの正しさ

脚注番号の対応付けが正しく行われているかを確かめるために，各会議から 17 論文ずつ合計 51 論文を無作為に選択し，調査した．その結果，F 値が 0.91 であり，高い性能で脚注番号の対応付けを行えることを確認した．

4.2.2　先行研究のデータセットとの比較

作成したデータセットと先行研究のデータセットとの比較を表 1 に示す．SEPID-cite，ACL-cite は言語処理，PMOA-cite は医学分野のデータセットである．引用を含む文（文献タグ）の全体に占める割合は本研究で作成したデータセットが 0.158，先行研究のデータセットが 0.075~0.195 となっており，同分野ではやや多いが全体としては同水準である．また，1 文当たりの文字数，単語数はそれぞれ 141，23 であり，同分野の ACL-cite と同水準の数値であった．

表 3　データセットの内訳

	ドキュメントの数	文数
学習	10,353	1,823,058
検証	1,294	229,398
テスト	1,295	225,378

4.2.3　ラベルごとの内訳

表 2 に引用なし，および，引用の種類別に「引用あり」の文の内訳を示す．引用が行われなかった文は約 190 万文であり，全体の 83.4%を占めている．URL 引用が含まれている文は 1%と非常に少ない．また，URL 引用のみ行われた文では他に比べて，文字数，単語数ともに少ない．文献タグによる引用が行われた文は長くなりやすく，URL 引用も行われた場合には全ラベルの内，最長となっている．

5　ベースラインの性能

先行研究の手法をベースラインとし，本研究で提案する引用要否判定においてベースラインがどの程度の性能を達成可能かを確認するために実験を行った．ベースラインとして，文脈埋め込みモデルを使用したテキスト分類において標準的な方法を採用している Gosangi らの手法 [3] を選択した．

5.1　ベースラインの手法

Gosangi ら [3] の Sentence Classification と Sentence Pair Classification をベースラインとした．それぞれのモデル図を図 3 に示す．

Sentence Classification（SC）　SC のモデルは BERT[7] などの文脈埋め込みモデルと分類層である全結合層から構成される．文脈埋め込みモデルに判定対象の文を入力し，[CLS] トークンに対応する出力を全結合層に繋げて各クラスのロジットを得る．

Sentence Pair Classification（SPC）　SPC のモデルは SC と同様である．入力には判定対象の文とその前後の 1 文を加えたものを文脈として，判定対象の文と文脈を [SEP] により結合して使用する．

ただし，Gosangi らの手法では全結合層のみが学習対象であるが，本研究では，全てのパラメータを

表 4　実験結果（太字は最も高い性能を示したものを指す.）

	Weighted-F1	Precision（引用あり）	Recall（引用あり）	F1（引用あり）	Recall（URL 引用）
ACL-cite に対する SC[3]	0.921	0.782	0.496	0.607	N/A
SC(full) + RoBERTa-base	0.991±0.000	0.973±0.001	**0.975±0.000**	0.974±0.000	**0.779±0.006**
SC(full) + RoBERTa-large	**0.992±0.000**	**0.978±0.001**	0.974±0.001	**0.976±0.000**	0.777±0.005
ACL-cite に対する SPC[3]	0.932	0.820	0.562	0.667	N/A
SPC(full) + RoBERTa-base	**0.993±0.000**	**0.976±0.001**	0.982±0.001	0.979±0.001	0.932±0.007
SPC(full) + RoBERTa-large	**0.993±0.000**	**0.976±0.001**	**0.984±0.000**	**0.980±0.000**	**0.939±0.009**

学習する方法（SC(full) と SPC(full)）を検証する[10].

5.2　実験設定

データセット　4 節で作成したデータセットを文書単位で学習，検証，テスト用に 8:1:1 の割合でランダムに分割した．作成されたデータセットの内訳は表 3 に示す通りである．

モデルと各種パラメータ　Gosangi らの手法 [3] に従い，RoBERTa-base[11]，large[12][13] を文脈埋め込みモデルとして使用した．最適化手法は学習率 1e-5（base），4e-6（large）の AdamW[14] とし，その他のパラメータは Huggingface Trainer のデフォルト値[13] とした．また CrossEntropyLoss を指標に Early Stopping（patience は 2）を使用した．バッチサイズは 32，勾配累積のステップ数は 2 である．

5.3　実験結果

表 4 に Weighted-F1 と「引用あり」の適合率，再現率，F 値，URL 引用についての再現率を示す．ACL-cite は作成したデータセットとオーバーラップがあるため、その結果 [3] も参考として記載する．結果の数値はそれぞれ 3 回異なるシード値で学習した平均と標準偏差である．large は base よりも若干の性能が向上したが全体の傾向に変化はなかった．

SC(full) + RoBERTa-base について，「引用あり」の F 値が 0.974 であり高い性能を発揮している．一方で，URL 引用についての再現率は 0.779 であり「引用あり」の 0.975 に対して低い値となっている．

SPC(full) + RoBERTa-base について，前後の文脈を加えることで SC(full) + RoBERTa-base より高い性能となった．特に URL 引用についての伸びが 0.153 と大きい．この結果から，文献タグによる引用は判定対象の文だけでもある程度判定できる一方で，URL 引用の判定には広い文脈が有効であるといえる．

10）事前に検証した結果，全てのパラメータを学習させた方が高い F 値であったため (full) で検証を行った．
11）https://huggingface.co/roberta-base
12）https://huggingface.co/roberta-large
13）https://huggingface.co/docs/transformers/v4.36.1/en/main_classes/optimizer_schedules#transformers.AdamW

表 5　SPC(full)+RoBERTa-base で誤って判定した文

判定対象の文	被引用物の種類
We further conducted stemming on the words with Iveonik English Stemmer.	ツール
5 Extracted from one of the latest Freebase dumps (downloaded in mid-August 2015)	データ

5.4　エラー分析

URL 引用を正しく検出できなかった文を人手で分析した．SPC(full) + RoBERTa-base で正しく判定できなかった文の例を表 5 に示す．1 つ目の例ではツールの引用を行っており，ツール名に対応して URL が記載されている．したがって，固有名詞に着目することで性能が改善する可能性がある．2 つ目の例では脚注でデータの引用を行っているが，1 つ目の例とは異なりデータ名は明示されていない．そのため，脚注の “Extracted from ~ firebase dump” からデータの取得についての文であることを識別する必要があり，こうした URL 引用に特有の言い回しへの対応が必要である．

6　おわりに

本論文では，文献タグによる引用に加えて URL 引用も含めた引用要否判定タスクを提案した．このタスクのために ACL Anthology より収集した論文データから新たなデータセットを作成した．このデータセットに対する文脈埋め込みモデルを使用したベースラインの性能を確認した．その結果，文献タグによる引用と比較して URL 引用の要否判定の再現率が低く，改善の余地があることが分かった．また，URL 引用の要否判定では特に前後の文脈の有無が重要であり，文脈の追加により大幅に性能が向上することを示された．加えて，URL 引用についてのエラー分析を行い，URL 引用に特有の言い回しへの対応が必要であることが分かった．

本研究で作成したデータセットの URL 引用の標本数が極端に少ない．このことが URL 引用の再現率の低さの原因である可能性がある．そのため，アップサンプリング等の手法を試す必要がある．

謝辞

本研究は一部, JSPS 科研費 基盤研究（B）21H03773 の支援により実施した.

参考文献

[1] 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 '18, pp. 1061—1064, New York, NY, USA, 2018. Association for Computing Machinery.

[2] 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.

[3] 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.

[4] Masaya Tsunokake and Shigeki Matsubara. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. In **Proceedings of the First Workshop on Information Extraction from Scientific Publications**, pp. 8–19, Online, November 2022. Association for Computational Linguistics.

[5] 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 (EMNLP-IJCNLP)**, pp. 5206–5215, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Nishant Kambhatla, Logan Born, and Anoop Sarkar. CipherDAug: Ciphertext based data augmentation for neural machine translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1 (Long Papers)**, pp. 201–218, Dublin, Ireland, May 2022. Association for Computational Linguistics.

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

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

[9] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)**, pp. 1755–1759, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In **Advances in Neural Information Processing Systems**, Vol. 26. Curran Associates, Inc., 2013.

[11] Mann Khatri, Pritish Wadhwa, Gitansh Satija, Reshma Sheik, Yaman Kumar, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Citecaselaw: Citation worthiness detection in caselaw for legal assistive writing, 2023.

[12] 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 (COLING): System Demonstrations**, pp. 136–140, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[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, 2019.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.