論文執筆支援を目的とした引用要否判定タスクのドメイン間比較

小山康平 [1] 小林恵大 [1] 成松宏美 [2] 南泰浩 [1]
[1] 電気通信大学情報理工学研究科 [2]NTT コミュニケーション科学基礎研究所
k2131071@edu.cc.uec.ac.jp k2231042@edu.cc.uec.ac.jp
hiromi.narimatsu.eg@hco.ntt.co.jp minami.yasuhiro@is.uec.ac.jp

概要

学術論文の執筆および査読支援を目的として，引用の不足や余分な引用を自動で検出する引用要否判定タスクを課題とした研究が行われている．この引用要否判定タスクは非常に高い精度の推定が可能であるが，実験に用いるデータセットには実用が想定される状況よりも簡単な問題が含まれる懸念がある．そのため，我々は引用要否判定タスクにおけるデータセットの妥当性について検討してきた．本稿では，論文のドメインの観点から妥当性を検討した．まず，異なるドメインから作成された複数のデータセットを用いて引用要否判定タスクを実施しその精度を比較した．ドメインごとに推定精度に大きな差があることが分かった．続いて，ドメインの差による引用要否判定タスク精度差の原因を調査するために，それぞれのデータセットに出現する単語パタンの比較を実施した．出現する単語に差があり，引用要否判定学習モデルの学習精度にも影響を与えている可能性があることが分かった．

1 はじめに

公開される学術論文の数は年々増加しており [1] 研究者はこれまで以上に早いスピードで論文化し公開することが求められている．この論文化の過程では，学術論文を執筆する際，多数の関連研究を読み，一文一文に気を配りながら適切に引用することが求められ，研究者の負担は増している．論文執筆に慣れた研究者であれば，それらの作業を効率よく行うことができるが，そうでない場合には，執筆に関わる負担は大きい．また執筆した論文をチェックする共著者や査読者への負担の影響も大きい．

こうした背景から，論文執筆支援に関わる様々な研究が行われている．具体的には，既に検索済みの論文の閲読時間削減を目的とした論文要約 [2, 3, 4] や，未検索の論文の効率的な検索を目的とした参考文献推薦 [5]，論文執筆の効率化を目的とした引用要否判定 [6, 7]，被引用文献割り当て [8]，引用文生成 [9, 10] などである．本研究では，論文チェックにおいて最初に必要となる引用要否判定タスクに着目する．

引用要否判定タスクとは，論文中の任意のある文に対して引用が必要か必要でないかを推論するタスクである．その精度は９割程度と非常に高く [11]．前後の文脈の入力や学習モデルの改善により，さらなる精度の向上が期待できる [12]．しかしながら，支援システムを目的とするとき，これらの精度が真に得られているかは注意深く分析する必要がある．具体的には，引用箇所をそのまま取り除いているために発生する不完全文，引用をする際の明らかなパタン，論文カテゴリの偏りなどが判定精度を高めている可能性が示唆される．そのため，我々は引用要否判定を正しく評価し，その判定精度を向上を目的として，引用要否判定タスクのデータの妥当性分析を行ってきた [12]．

従来の引用要否判定タスクに用いられるデータセットは単一のドメインから作成されている．ドメインに偏りのあるデータセットを用いて引用要否モデルの学習を行った場合，ドメイン特有の引用，たとえば固有名詞などを特徴として学び取ることで，高い精度が出ている可能性があり，文全体から引用が必要かどうかは判断できていない可能性がある．以上の理由から，本稿では，異なるドメインから作成された複数のデータセットを用いて引用要否判定タスクを実施しその精度を比較する．もし，ドメインの偏りが引用要否判定タスクに影響を与えるなら，どのような情報が引用要否判定タスクに影響を与えるかを調査する必要がある．影響を与える要

素を特定するために，引用要否判定タスクに用いたデータセットをドメインと引用の要否の観点から分類し，N-gram や Tf-idf を用いて分析を行う．

2 関連研究

引用論文推薦の前段階として，引用の要否を判定する研究が行われている [11, 13]．引用要否判定タスクとは，論文中の各文に対して，引用が必要かどうかを判定するタスクである．初期の研究では，サポートベクターマシン (SVM) や決定木を使った判定器を学習する方法が提案されてきた [14, 15]．彼らは，ACL をはじめとする論文データベースから構築したデータセットを用いて評価を行なっていた．

近年では，公開される論文数の増加に伴い，より大規模な論文データベースからデータを作成できるようになった．ARC [16] などは特に幅広く使われている大規模データセットである．これによりデータ量が必要な深層学習をベースとする手法も提案されてきている [17]．さまざまな自然言語処理のタスクで高い性能を発揮している大規模汎用言語モデルの一つである BERT [18] を用いた研究もある．堂坂らは，BERT を引用要否判定タスクの少量のデータで転移学習することで，Bonab らが公開した Citation Worthiness データセット [19] にて，CNN をベースとする手法よりも高い精度が得られ，F 値で 0.7 に到達することを示している．成松らは，arXiv の論文を対象にデータセットを構築し，BERT で評価をした結果，F 値で 0.9 を達成しており，高い精度で要否の判定が可能であることを示している [20]．

しかしながら，いずれの研究においても，ドメインの違いが引用要否判定タスクに及ぼす影響の分析は行われていない．本研究では，異なる引用要否判定研究の結果を比較し，より適切な推定手法やデータセットを作成するためには異なるドメイン間の引用要否判定を実施する．

3 データセット

本稿では 2 種類のデータセットを使用する． 1 種類目のデータセットは AxCell [21] である．このデータセットは arXiv から集めた，コンピュータサイエンス分野の論文を基に作成されている．本来は引用要否判定のために作成されたデータセットではないが，より汎用的な論文執筆支援タスクに応用するために加工を行なった [22]．引用要否判定に用いた文は 707560 文である． 2 種類目のデータセットは PMOA-CITE である [23]．このデータセットは PubMed から集めた，医学分野の論文を基に作成されている．データ量は 1008060 文である．

4 異なるドメインの引用要否判定

異なるドメインから作成された引用要否判定モデルの推定精度に差が生じるかを確かめるために，上記の医療分野とコンピュータサイエンス分野のデータセットを利用し，それぞれに引用要否判定タスクを行う．また，単一ドメインのデータセットを学習基にした引用要否判定モデルを他分野に応用できるか確かめるために，学習データと評価データを異なるドメインにした引用要否判定も実施する．

4.1 実験条件

引用が必要か不要かの二値分類の性能を言語モデルを用いて評価する． 本稿では，汎用言語モデルとして最初に成功を納めた BERT [18] を基に，科学論文を事前学習に使用した SciBERT [24] を用いて引用要否判定を実施する．データセットには上記の AxCell [21],PMOA-CITE [23] を使用し，データ数同数になるよう，データ量を (train:200,000 文,dev:50,000 文,test:50,000 文) とした．学習モデルのパラメータは (学習率:1e-7,Epoch 数:10, batch_size:32) とした．

4.2 実験結果

実験結果を表 1 に示す．学習と評価共に同一ドメインの結果を見ると，AxCell の方が PMOA よりも推定精度が高い．このことから，コンピュータサイエンス分野のデータセットは医療分野のデータセットよりも推定が容易であることが分かる．ドメインによって引用要否判定の難易度が異なる場合があることが予想される．また，学習用データと評価データで異なるドメインのデータセットを使用した場合に精度が低下していた．このことから，引用要否判定タスクの精度はドメインごとの固有情報から影響を受ける可能性があり，ある単一分野から学習した引用要否判定モデルは他分野に転用できないことが分かる．学習データに AxCell, 評価データに PMOA-CITE を用いた結果が学習データに PMOA-CITE, 評価データに AxCell を用いた結果を下回ったことから，PMOA-CITE には AxCell よりも幅広い情報が含まれている可能性がある．どのような固有情報が引用要否判定タスクに影響を与えるかを明らかにするために，次章では PMOA-CITET と

表 1　複数のドメインの引用要否判定

traindata	testdata	Acc	Pre	Rec	F1
AxCell	AxCell	0.899	0.930	0.863	0.896
PMOA	PMOA	0.812	0.798	0.835	0.816
AxCell	PMOA	0.684	0.808	0.484	0.605
PMOA	AxCell	0.762	0.712	0.877	0.786

AxCell のデータセットに現れる単語パタンを分析し比較する．

5　異なるドメインの単語パタン比較

PMOA-CITE 及び AxCell の引用要否判定データセットに含まれる固有情報を明らかにするために，それぞれのデータセットの N-gram 単語頻度の調査を行う．

まず初めに，ドメインごとに重要度の高い単語パタンを調査し，データセットの特徴を明らかにする．そのために，データセットの文中から単語 bigram 情報を抽出し，PMOA と AxCell それぞれにおける Tf-idf 値を計算する．

続いて，引用手段の特徴の差を明らかにするために，片方のドメインでは引用の要否によって単語 bigram パタンの出現頻度が大きく異なるが，もう片方のドメインでは出現頻度が変わらない単語を調査する．今回は，引用の要否による出現頻度の偏り (特有度) 独自に定義してドメインごとに計算する．

5.1　実験条件

特有度の計算は式 (1) に示す．(x, y は対象 bigram が引用の要・否それぞれに出現する割合を表す)
調査には (4.1) 節で定義したデータセットと同様のものを使用する．

$$\text{特有度} \quad = \quad log|x - y| \tag{1}$$

5.2　実験結果

ドメインごとの単語 bigram パタンの重要度を図 3 に示す．縦軸が PMOA-CITE における Tf-idf 値, 横軸が AxCell における Tf-idf 値を表している．PMOA では”associated with”,”this study”のようなパタンの tf-idf が高い一方で，AxCell では”which is”,”For example”のようなパタンの tf-idf が高かった．以上のことから，ドメインごとの出現単語に大きく差がある単語 bigram パタンが存在し，深層学習を用いて作成した引用要否判定モデルに影響を与える可能性があることが分かる．

ドメインごとの単語 bigram パタンの特有度を図 4 に示す．縦軸が PMOA-CITE における重要度, 横軸が AxCell における重要度を表している．この図では左上, 右下に位置する単語 bigram パタンほど引用の要否及びドメイン間での重要度の差が大きいことになる．こうした単語 bigram パタンには”deep learning”,”in patients”をはじめとするを含む表現や”focus on”,”effect of”のような言い回し表現も含まれる．以上のことから引用要否判定データセットにはドメイン特有の単語パタンが含まれていることが分かる．

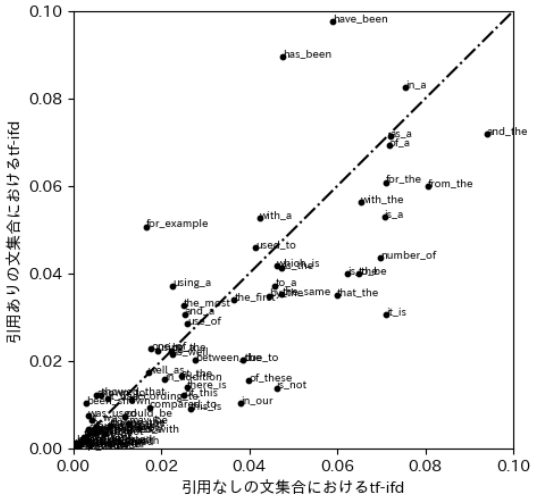

図 1　引用要否ごとの Tf-idf(AxCell)

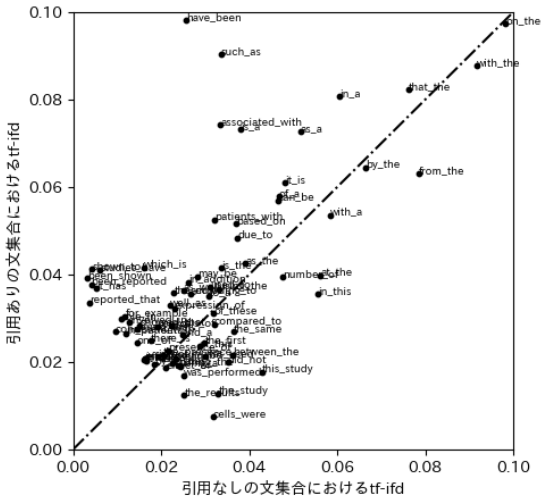

図 2　引用要否ごとの Tf-idf(PMOA)

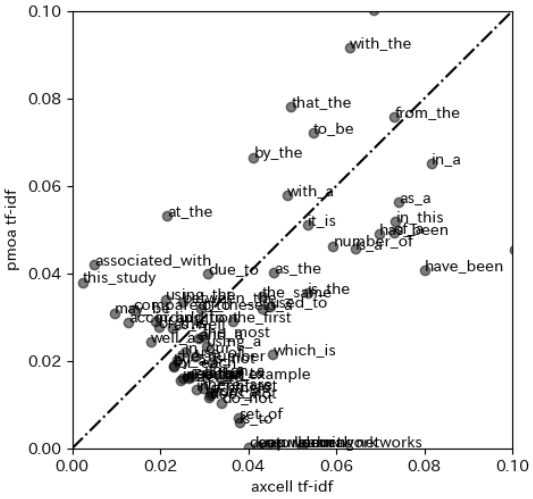

図 3　単語パタンの重要度

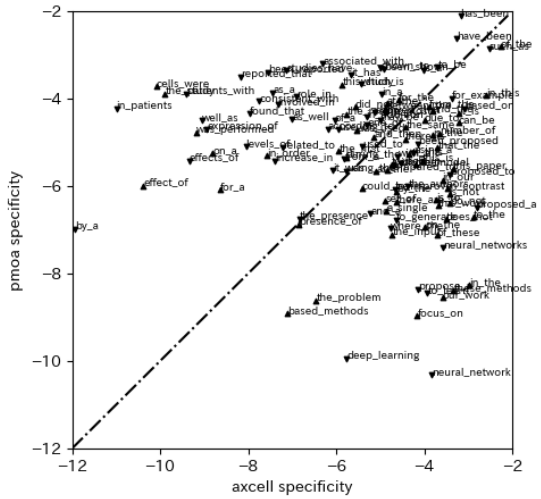

図 4　単語パタンの特有度

6　考察と今後の展望

本稿では，論文中のある文に対して引用が必要か必要でないかを推論するタスクである引用要否判定タスクにおいて，ドメインの種類が引用要否判定に影響を与えるかを調査するために，複数のデータセットで引用要否判定を実施した．コンピュータサイエンス分野と医学分野のデータセットでこの調査を行いその精度を比較した結果，コンピュータサイエンス分野のデータセットを用いた方が精度が高いことがわかった．このことから，引用要否判定はドメインごとに難易度に差があることが分かる．また，学習と評価で異なるデータセットを使用した場合には精度が大きく低下した．このことから，引用要否データセットにはドメイン固有の表現が含まれており，学習モデルの転用には，考慮が必要なことが分かった．

異なるドメインのデータセット内における単語の重要度を bigram の td-idf で比較した結果，ドメイン間で大きな差があることがわかった．また，単語 bigram パタンに対して引用要否ごとの偏りを表した特有度を計算した．その結果，様々なパタンでデータセットごとの特有度に差が見られた．以上のことから，カテゴリごとに単語パタンの出現頻度には偏りがあり，引用要否判定モデルの学習に影響を与えている可能性があることが分かる．

ドメインごとにデータセットには特徴があり，引用要否判定に与える影響にも大きく差が出る．そのため，より実用的な引用要否データセットを作成するためには，様々な領域全体をカバーするより大規模なデータを利用する，あるいは，厳密な実験条件を定義し，ドメインごとの明示的な単語パタンを取り除いたデータセットを作成することなども必要である．

謝辞

研究の遂行にあたり，ご助言をいただきました，秋田県立大学堂坂浩二教授，NTT コミュニケーション科学基礎研究所杉山弘晃氏，東中竜一郎氏，大阪工業大学平博順教授，工学院大学大和淳司教授，科学技術振興機構菊井玄一郎氏に感謝いたします．

参考文献

[1] 桑原真人. 各国の論文数の推移から見えるもの. 日本物理学会誌, Vol. 72, No. 4, pp. 246–251, 2017.

[2] Simone Teufel and Marc Moens. Articles summarizing scientific articles: Experiments with relevance and rhetorical status. **Computational Linguistics**, Vol. 28, No. 4, pp. 409–445, 2002.

[3] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In **Proceedings of AAAI 2019**, 2019.

[4] Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Enhancing scientific papers summarization with citation graph. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 14, pp. 12498–12506, May 2021.

[5] Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. Scientific paper recommendation: A survey. **IEEE Access**, Vol. 7, pp. 9324–9339, 2019.

[6] Michael Färber, Alexander Thiemann, and Adam Jatowt. To cite, or not to cite? detecting citation contexts in text. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, **Advances in Information Retrieval**, pp. 598–603, Cham, 2018. Springer International Publishing.

[7] 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**, pp. 4539–4545, Online, June 2021. Association for Computational Linguistics.

[8] Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. **International Journal on Digital Libraries**, Vol. 21, , 12 2020.

[9] Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. Automatic generation of citation texts in scholarly papers: A pilot study. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6181–6190, Online, July 2020. Association for Computational Linguistics.

[10] Qingqin Wang, Yun Xiong, Yao Zhang, Jiawei Zhang, and Yangyong Zhu. Autocite: Multi-modal representation fusion for contextual citation generation. In **Proceedings of the 14th ACM International Conference on Web Search and Data Mining**, WSDM '21, p. 788–796, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Michael Färber, Alexander Thiemann, and Adam Jatowt. To cite, or not to cite? detecting citation contexts in text. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, **Advances in Information Retrieval**, pp. 598–603, Cham, 2018. Springer International Publishing.

[12] 小山康平, 小林恵大, 南泰浩, 成松宏美. 引用要否判定タスクにおけるモデルの性能評価とデータの妥当性分析. 言語処理学会第 28 回年次大会, 2022.

[13] Hamed Bonab, Hamed Zamani, Erik Learned-Miller, and James Allan. Citation worthiness of sentences in scientific reports. 07 2018.

[14] 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. IEEE, 2010.

[15] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C Lee Giles. Citation recommendation without author supervision. In **Proceedings of the fourth ACM international conference on Web search and data mining**, pp. 755–764, 2011.

[16] 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'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[17] 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. **CoRR**, Vol. abs/2104.08962, , 2021.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[19] Hamed Bonab, Hamed Zamani, Erik Learned-Miller, and James Allan. Citation worthiness of sentences in scientific reports. In **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**, pp. 1061–1064, 2018.

[20] Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira. Task definition and integration for scientific-document writing support. In **Proceedings of the Second Workshop on Scholarly Document Processing**, pp. 18–26, 2021.

[21] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. Axcell: Automatic extraction of results from machine learning papers. **arXiv preprint arXiv:2004.14356**, 2020.

[22] 小山康平, 南泰浩, 成松宏美, 堂坂浩二, 田盛大悟, 東中竜一郎, 平博順. 学術論文における関連研究の執筆支援のための被引用論文の推定. 言語処理学会第 26 回年次大会, 2021.

[23] TONG ZENG. PMOA-CITE dataset. 6 2020.

[24] Iz Beltagy, Kyle Lo, 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 (EMNLP-IJCNLP)**, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.