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

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 [2]*NTT* コミュニケーション科学基礎研究所
k2131071@edu.cc.uec.ac.jp k1810249@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．引用要否判定は主に，公開された学術論文の引用情報を正解として，その引用情報を自動的に削除することで構築した文がタスクの評価に用いられる．しかしながら，そのように構築されたデータセットは，執筆中の原稿において引用要否を判定する場合よりも，簡単なタスクになっていることが懸念される．例えば，引用があることが前提の記述パタンや引用情報を除いたことにより不完全な文になるなどである．そこで，本研究では，実用環境におけるモデルの性能を正しく評価することを目的として，タスク評価用データの妥当性分析を行なった．結果として，自動で構築した引用要否判定タスクの評価用データセットは明確な特徴を持つものも多く含まれていることがわかった．

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．一方，研究者が学術論文を執筆する際には，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も求められており，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に着目し，そのタスクの判定精度向上を目指す．

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．基本的には根拠が必要な文が引用の必要な文であるため，既存研究の多くは入力された一文に対して引用の要否を判定するタスクを扱ってきた [11]．その精度は９割程度と非常に高い．一方，Amjad らは引用の前後の文脈によって，criticizing, basis, comparison などの引用の目的を推定するタスクの精度が変わることを指摘している [12]．引用要否判定タスクも前後の文脈の入力により，さらなる精度の向上が期待できる．

しかしながら，支援システムを目的とするとき，真にこれらの精度が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，引用要否判定タスクが，引用箇所をそのまま取り除いているために，不完全な文かどうかや，引用をする際の明らかなパタンが，判定精度を高めている可能性が示唆される．そこで本稿では，引用要否判定を正しく評価し，その判定精度を向上させることを目的として，現在の引用要否判定タスクのデータの妥当性分析を行う．実際には，成松らが，論文執筆支援システムの構築を目的として公開したデータセット [13] を用いて，その性能を評価し，正解・不正解となるデータのパタンから，性能の妥当性を検証する．

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, 14]．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当てるタスクである．初期の研究では，サポートベクターマシン (SVM) や決定木を使ったモデルで判定器を学習する方法が提案されてきた [15, 16]．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 [17] 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 [18]．Bonab らは，Word2Vec による単語の分散表現を入力とし，CNN を用いた引用要否判定モデルを提案している [19]．また，Färber らは，Glove による単語分散表現を入力とし，RNN と CNN を組み合わせて引用要否を判定するモデルを提案している [11]．従来の SVM や決定木を用いる手法と比較し，高い精度が得られている．さらに，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 [20] 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 [13]．

しかしながら，いずれの研究においても，そのデータの妥当性については検証されていない．実際に執筆支援システムとして引用要否判定を用いるには，実用時に近いデータでの性能評価が求められるが，自動で構築したデータには，不完全な文や引用があることを示唆する文字列など，モデルが判定しやすい特徴を含むことが懸念される．そこで，本研究では，統合的な執筆支援システムの実現を目的として構築されたデータセット [21, 13] からタスク用データを構築し，大規模汎用言語モデルを用いてその性能を評価するとともに，より実用に近い環境で評価できるタスクデータ設計の指針について議論する．

表 1　入力テキストとラベルの例 (引用要の場合)

元の文	They proposed a method for 〜 [1].
入力テキスト	They proposed a method for 〜.
ラベル	1

表 2　引用要否判定データ量

	train	dev	test
引用あり (文)	213,630	70,175	90,701
引用なし (文)	275,698	90,454	70,636
論文数	18,506	6,169	6,171

3　データセットの構築

文脈の要否による性能の違いも評価するために，統合的な執筆支援を目的として構築されたデータセット [13] をもとに，引用要否判定タスク用のデータを構築した．基本的な構築方法は，過去に著者らが引用文と被引用文のペア適正を判定するために構築したデータセット [22] と同様である．

Tex のソースが公開されている論文を AxCell [23] のリストから選択し，Tex のファイルに対して以下の手順で加工して構築する．

1. section{}タグの情報を用いて関連研究の章の本文を抽出
2. Tex 記号をもとに図表や注釈を本文から除去
3. nltk ライブラリでテキストを文単位に分割
4. 文から引用記号\cite{}を削除
5. 引用を含んでいた文を引用要，そうでない文を不要に分類

なお，文脈の有無による精度比較も可能とするため，判定対象の文の前後もコンテキストとわかる様にデータセットに含めた．今回作成したデータの量は表 2 となった．

4　性能評価

4.1　実験設定

引用が必要か不要かの 2 値分類の性能を，3 つの大規模汎用言語モデルを用いて評価する．汎用言語モデルとして最初に成功を納めた BERT [20] と，事前学習でより長い文が学習されている RoBERTa [24] に加えて，科学論文の文章を対象に BERT を事前学習した SciBERT [25] を用いる．いずれの手法においても，公開されている事前学習済みのモデルに対して，SentenceClassification の形式で転移学習した．

表 3　引用要否判定タスクの結果

	追加の文脈	Acc	Pre.	Rec.	F1
BERT	なし	0.906	0.940	0.835	0.885
RoBERTa	なし	0.964	0.978	0.939	0.958
SciBERT	なし	0.922	0.935	0.855	0.893
BERT	直前	0.914	0.928	0.870	0.898
RoBERTa	直前	0.969	0.976	0.954	0.964
SciBERT	直前	0.919	0.935	0.876	0.904
BERT	直後	0.916	0.945	0.857	0.899
RoBERTa	直後	0.967	0.968	0.956	0.962
SciBERT	直後	0.919	0.960	0.850	0.902

また，引用要否判定の学習に文脈情報が効果的であるかを確かめるために，推論対象の直前の一文・直後の一文をそれぞれ入力に追加し学習をすることで精度を比較する．これらの追加の文脈は，対象の文とセパレータ記号でつなげて入力とする．

4.2　実験結果

結果を表 3 に示す．手法や文脈の有無によらず，0.9 に近い精度が得られている．モデルとしては，前後の文の入力によらず RoBERTa が最も高い精度を示した．転移学習に用いたデータサイズは手法間で同じであるため，事前学習の効果が大きいと考えられる．

5　データの妥当性分析

引用要否判定タスクの精度は，0.9 に近く，システムの性能としては十分に見える．しかしながら，自動で構築したデータには，不完全な文や引用があることを示唆する文字列などモデルが判定しやすい特徴を含んでいる可能性が考えられる．そこで，(1) 引用記号を含む文に特有の単語 N-gram の出現パタンがないかの調査，(2) 正例・負例内の不完全な文の数および不完全な文を除いた時の性能の評価，の 2 点を行う．

5.1　正例・負例に特有の文字列調査

モデルが，正例（引用を含む文）・負例（引用を含まない文）の文中に表れる特有のパタンを特徴量として分類をしている可能性がないかを調べる．その目的は，実用において入力されるデータと，タスク評価用データで人工的に作成した入力データとの違いとそれによる影響の調査である．例えば，今回用いたデータには “According to a method to 〜” というような引用があることを明示している文も判定対象に含まれる．しかしながら，実用のシーンでは，“According to 〜” の部分は入力されることはなく，後半の文節のみが入力されると考えられる．そこで，このような引用の要が明示的な文に表れる文字列の特徴やそれによる影響を調べる．具体的には，正例，負例ごとに 1-gram，2-gram の出現頻度を調査し，片方のみに突出して表れる語については，明示的なパタンであるとする．

特に出現率が高い 1-gram と 2-gram をプロットした図を (図 1, 2) に示す．横軸は引用を含む文における出現率，縦軸は引用を含まない文における出現率，を表している．引用を含む文に現れる傾向が強い “For example” などは直後に他者の論文に対する言及が予想されるため，引用が含まれる文であることが明確な文となっている可能性が示唆される．一方，引用が含まれない文では “our work” や “this paper” などのパタンが出現する傾向が強い．実際に，自分の研究に対する意見を述べる場合に，“our work” を含むことが多く，このような文では引用が不要であることが多い．以上のことから，評価用データには，実用時よりも簡易に判定可能なデータになっている可能性が高いことがわかる．したがって，実用時の判定対象に含まれていないことが想定される文字列のパタンについては，評価用データセットから削除することが必要と考えられる．

5.2　不完全な文の調査

引用のスタイルはさまざまで，“According to Devlin et al. 〜 [1]” と書く場合もあれば，“According to [1], 〜” のように引用記号を名詞の代わりに使用する場合もある．そのため，引用箇所を取り除くことで，入力される文が不完全になる可能性もある．そこで，不完全な文が推論に与える影響を調査するために，名詞の欠落を判定する推論モデルである非文判定器を作成する．非文判定器は BERT の事前学習済みモデルを転移学習することで作成する．学習データには，arXiv の API を使用して取得した 30,846 件の論文のアブストラクトを使用する．これらのアブストラクトを文単位に分割し，そのままの文を正例，名詞となる単語をランダムに消去した文を負例とする．学習した非文判定器 (精度は表 4 に記載) で引用要否判定のテストデータを完全な文・不完全な文に分類し，それぞれの文を用いて引用要否判定タスクの精度を算出する．

非文判定器によって分類された完全な文と不完全な文ごとの引用要否判定タスクの結果を表 5 に示

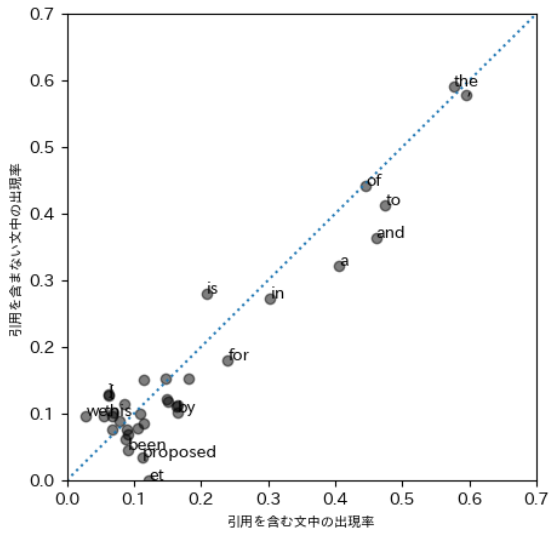

図 1　1-gram の出現率

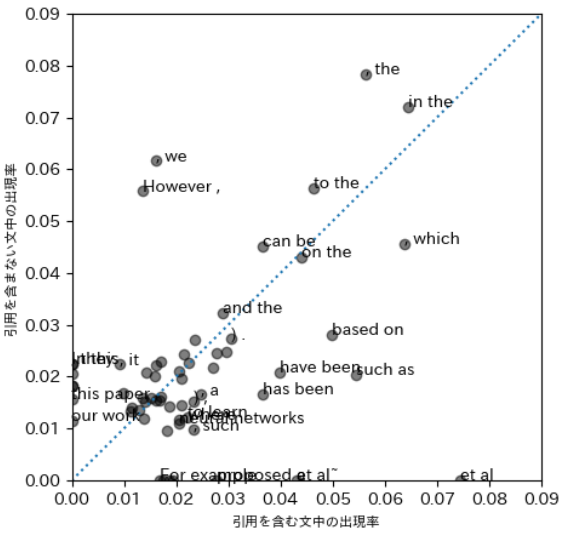

図 2　2-gram の出現率

表 4　非文判定器の性能

Acc	Pre.	Rec.	F1
0.816	0.746	0.959	0.839

表 5　非文判定後のデータを用いた引用要否判定の結果

	Acc	Pre.	Rec.	F1	数
完全な文と判定	0.898	0.930	0.768	0.841	121136 文
不完全な文と判定	0.899	0.967	0.886	0.925	39493 文

す．不完全な文と判定された文の Recall と F 値が完全な文を大きく上回る結果となった．これは，名詞の欠落をヒントにモデルが引用要否判定を行っていることを示唆している．

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を行った．モデルの種類や入力文脈などの条件を変えて精度を比較したところ，RoBERTa が最も良い精度を示し，入力文脈が長いほうが精度の高いモデルを作成できた．このことから，適切な事前学習と引用前後の文脈情報は引用要否判定タスクに対して効果的であると考えられる．入力する文脈長を調整し，適切な入力の形式を調査することが今後の課題である．

データセットを実用時の条件に近づけるために，妥当性分析を行った．引用が含まれていることが明示的な文はデータセットに適さないため，どのような明示的な引用のパタンが存在するかを N-gram を用いて調査した．結果，複数の明示的引用を伴いやすいパタンが，引用を含む文では出現率が高いことを確認した．これらのパタンで，明示的な引用文を簡単に発見できる可能性がある．引用の消去による名詞の欠落がモデルの推論に影響を与えるかを確かめるために，非文判定機を用いてテストデータを完全な文と不完全な文に分けて，引用要否判定タスクを行った．不完全な文であると推論された文の要否判定の精度が高かったことから，不完全な文は推論に影響を与えることが示唆される．引用要否判定タスクを実用に近い条件にするためにも，不完全な文や明示的な引用文をデータセットから見つけ出す必要がある．今後は，より実用を考慮したタスク設計を行い，論文執筆者支援を続けたい．

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，農研機構菊井玄一郎チーム長に感謝いたします．

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] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. Purpose and polarity of citation: Towards NLP-based bibliometrics. In **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 596–606, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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).

[18] 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.

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

[21] 成松宏美, 小山康平, 堂坂浩二, 田盛大, 東中竜一郎, 南泰浩悟, 平博順. 学術論文における関連研究の執筆支援のためのタスク設計およびデータ構築. 言語処理学会第 27 回年次大会発表論文集, pp. C6–4, 2021.

[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] 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.

[24] 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.

[25] 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.