研究データ利活用の促進に向けた論文における URL による引用の分類

角掛正弥 [1]　松原茂樹 [1,2]
[1] 名古屋大学大学院情報学研究科　[2] 名古屋大学情報連携推進本部
tsunokake.masaya.z3@s.mail.nagoya-u.ac.jp　matubara@nagoya-u.jp

概要

論文における研究データの引用を識別・解析することは，研究データリポジトリの拡充や研究データの検索，推薦，評価などに繋がる．論文において URL による引用の多くが研究データを参照している．本論文ではそのような引用に対し，参照先リソースが研究で果たす役割とその種類，および，引用した理由を求める分類問題に取り組む．提案する手法では従来手法のフレームワークに，節タイトルや脚注の文を入力素性として追加する．データセットを作成し，提案手法の有効性を検証した．

1　はじめに

オープンサイエンスは，論文や研究データ[1)]など研究資源の共有や利活用を促進する活動である．この方策にリポジトリの整備が挙げられる．論文についてはリポジトリが整備され，論文検索サービスも普及しアクセス性向上に寄与してきた．近年，研究データについてもそれら[2)]の整備が進んでおり，アクセス性向上や利活用が注目されている．

研究データリポジトリ整備には，研究データの登録やそのメタデータの生成が必要となる．これらの作業を可能な限り自動化することで，リポジトリ整備の効率化や登録される研究データの増加が期待できる．この実現のために，論文における研究データの引用の活用が考えられる．引用には，研究データの名称や用途など，メタデータとして記載すべき情報が記されることが多い．加えて，既存のメタデータには存在しない情報 [4] も得られる可能性がある．

1）　研究データの定義は文脈ごとに様々であるが [1]，本研究では研究の過程で生成/使用されるデジタルオブジェクト（e.g., 計測/試験データ，プログラム，ソフトウェア）を想定する．これは ACM が定める “artifact”[2] に近い．

2）　リポジトリの例：Zenodo (https://zenodo.org/), Mendeley Data (https://data.mendeley.com/)　検索サービスの例：DataCite Search (https://search.datacite.org/), Google Dataset Search[3] (https://datasetsearch.research.google.com/)

しかし，論文等の文献を対象とした引用（**文献引用**）とは異なり，研究データの引用方法は多様でありその識別が求められる．

本研究では，論文における URL による引用（**URL 引用**）に着目し，研究データの引用の識別・解析を目指す．URL 引用の例を図 1 に示す．URL 引用では論文以外にも多様なリソース（e.g., データセット，ソフトウェア，ホームページ，記事）が参照される．URL 引用の解析は研究データの識別に繋がるうえ，文献引用に比べ自由度の高い URL 引用の実態を解明する効果も期待できる．

本論文では，論文における URL 引用について下記を求める分類手法を提案する．

1. URL で参照するリソースが研究で果たす役割
2. URL で参照するリソースの種類
3. 著者がその引用を行った理由/目的

Zhao ら [5] は同様の問題に対しマルチタスク学習による分類手法を提案している．具体的には，引用箇所の周囲の単語列（**引用文脈**）を BERT[6] に入力し，得られた埋め込み表現を各タスク用の分類層に投入する．本論文で提案する手法では，新たな分類の素性として節タイトルと URL が記載された脚注を加える．また，参考文献を介した URL 引用も新たに分類対象としている．

2　関連研究

2.1　引用分類

論文における引用は一様ではなく，長年分析の対象であった [7, 8, 9]．Garfield[7] は著者が引用を行う理由を考察し，15 種類の動機を挙げている．Moravcsik ら [8] は，確認的・否定的，概念的・操作的など 2 項対立的な分類を 4 種類定義し，引用を調査した．また，引用を自動分類する手法が提案されている [10, 11, 12, 13]．Teufel ら [10] は引用機能（著

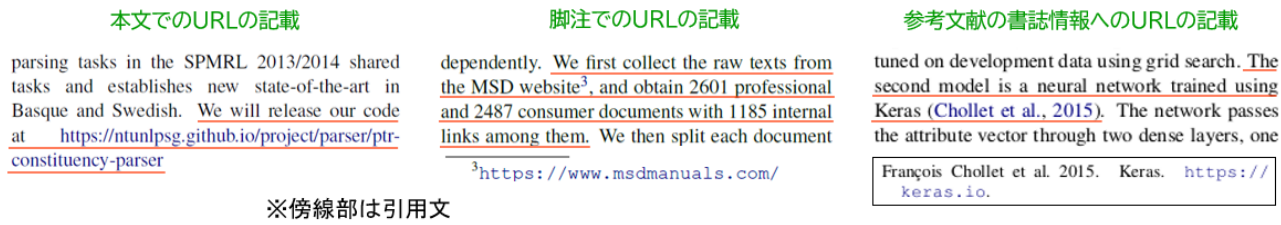

図 1　URL 引用の例

者が被引用論文を引用した理由）の観点から引用を分類する手法を提案した．Cohan ら [13] は引用を背景情報の提示，手法等の利用などに分類する手法を提案すると共に，大規模なデータセットを作成した．その他に，引用の意義や引用対象の種類など様々な観点の分類が存在する [14]．

Ding ら [15] は，引用をその内容に基づき分析/分類するこれらのアプローチを Content-based Citation Analysis(CCA) として整理した．CCA は論文の要約や推薦，検索等の応用が存在する [15]．また，引用機能は学術動向の分析 [11, 12] や引用文の自動生成 [16]，論文の被引用数予測 [12] へも寄与している．

2.2　研究データの引用

近年，論文における研究データの引用方法について統一的な慣習を定める動きがある．例えば，FORCE11 は “Data Citation Principles”[17] や “Software Citation Principles”[18] を宣言している．一方で，慣行に従わない多くの引用も存在する [19]．そのため，論文から研究データの参照を自動的に識別する試みがある [20]．例えば，論文の文章からデータセットの名称やそれに言及する文字列を特定する研究 [21, 22, 23]，ソフトウェアに対して同様のアプローチを採る研究 [24, 25, 26] が存在する．明示的な引用から研究データを識別する試みもあり，Ikoma ら [27] は Reference 節から言語資源を指す書誌情報の特定を目指した．URL 引用は研究データ参照の手がかりにもなる [24] ことから，研究データを参照する URL の識別も行われている [28]．

2.3　URL 引用の分類

論文における URL 引用の増加を受け，URL で参照されたリソースの活用を目指す研究も存在する．例えば，生医学分野のオンラインリソース検索システム [29] や自然言語処理分野の情報ポータル [30] の構築において，URL が論文から抽出されている．Nanba[31] は論文から URL とそれを表すタグを分散表現を基に抽出する手法を提案している．

各 URL 引用を分類する研究も存在する．Zhao ら [5] は広範な学術リソースの検索/推薦システムや知識グラフの構築に向け，URL 引用に対し 2.1 節で述べた CCA を適用した．具体的には，引用文脈から，参照先リソースの引用元論文における役割や引用機能を求める分類手法を提案した．

本研究では研究データのメタデータ生成を見据え，Zhao らの方法を発展させる．Zhao らが定義したリソースの役割のうち，Material と Method が研究データに相当する．この分類タスクを解くことで，研究データの引用の識別に加え，役割や引用の理由など研究活動におけるその研究データの利用に関した情報を取得できる．URL 引用は書誌情報が列挙される文献引用に比べて提供される情報が少なく曖昧性が高い．これを解析可能にすることは学術コミュニティにとっても有意義だと考えられる．

3　問題設定

本研究の分類タスクでは，各 URL 引用に対し，参照したリソースが研究で果たす役割（**Resource Role**）とその種類（**Resource Type**），および引用機能（**Citation Function**）を求める．Zhao ら [5] の設定では，Resource Role を一般/詳細の 2 階層で定義していたが，本研究では詳細な Resource Role を Resource Type と捉え再定義している[3)]．同じ URL であっても著者によって意図している参照先リソースは変わり得るため，いずれのタスクも引用文脈等から判断する必要がある．

分類対象の URL 引用はその実施方法から (1) 本文での URL の記載，(2) 脚注での URL の記載，(3) 参考文献の書誌情報への URL の記載 に大別できる．図 1 に例[4)]を示す．Zhao ら [5] が分類対象としていたのは (1)，(2) のみであった．しかし，Web 上のリ

3）　研究データのメタデータ生成等への応用を考慮した場合には，Zhao らの定義した詳細な Resource Role は種類と捉えるのが自然と考えたためである．
4）　(2),(3) では本文上の対応する周辺文が引用文脈に相当する．

表 1　Resource Role と Resource Type の一覧

Role	Type	説明
Material	Dataset	corpus, image set 等
	Knowledge	lexicon, knowledge graph 等
	DataSource	Dataset/Knowledge の構築元データ
Method	Tool	toolkit, software, system 等
	Code	codebase, library, API 等
Supple-ment	Document	tutorial, specification, guideline 等の Web 上の文書や記事
	Paper	conference paper 等
	Media	game, music, video 等
	Website	その他のリソース（e.g., services, 人/組織/イベント等の homepage）
Mixed	Mixed	複数のリソースが参照される場合

表 2　Citation Function の一覧

Function	説明
Use	参照先のリソースが引用元論文の研究でそのまま使用された場合の引用
Produce	参照先のリソースが引用元論文の研究で初めて作成/公開された場合の引用
Compare	参照先のリソースが他のリソースと比較されていた場合の引用
Extend	参照先のリソースが引用元論文の研究で使用され，かつ，その過程で改変されていた場合の引用
Introduce	参照先リソースや関連する情報（背景，特徴等）が紹介されていた場合の引用
Other	他のカテゴリに属さない場合の引用

ソースを引用する際に，参考文献として引用しその書誌情報へ URL を併記するケースも存在する．参考文献としての提示が推奨される場合もあることから (3) も考慮する必要がある．

Resource Role/Type の分類ラベルの一覧を表 1 に示す．リソースの種類によって果たしうる役割が定まるため，Resource Role と Resource Type の間には対応関係が存在する．なお分類ラベルは，研究データのメタデータ生成を見据え，Zhao ら [5] の設定を基本とし一部を変更している．また，URL 引用の中には複数のリソースが同時に参照されている場合がある．この場合，特定のラベルに分類できないため，対応するラベルとして「Mixed」を設けた．「Mixed」は一部の文献引用分類にも存在するラベルである [32, 16]．Citation Function の分類先ラベルの一覧を表 2 に示す．これは Zhao らの設定と同一である．

4　提案手法

本節では提案手法について記述する．Zhao ら [5] は同様の分類タスクに対し，SciResCLF という枠組みを提案している．SciResCLF では引用文脈を事前学習済み BERT[6] へ入力し，[CLS] の埋め込みを各タスク用の分類層へ投入する．また，各タスクのクロスエントロピーの加重和を用いてモデル全体を学習する．この枠組みに対し，本論文では，グローバルな文脈情報として引用箇所の節タイトル，および，脚注のテキストを用いた分類を提案する．

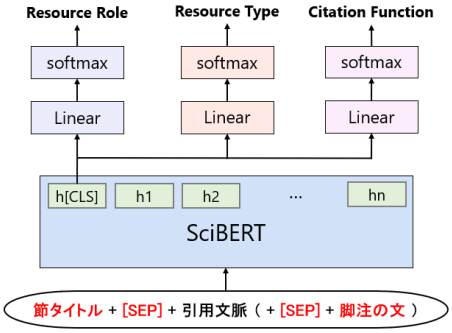

図 2　提案手法のアーキテクチャ

Jurgens ら [12] は文献引用において引用が出現する節と引用機能の間に関係性があることを報告している．本タスクにおいても，URL 引用が行われた節の情報が分類に寄与すると考え，節タイトルを活用する．[5] また，URL 引用の中には，本文中では参照先リソースを明示せず脚注にのみ記述する場合がある．そのため，URL 引用の分類においては脚注のテキストも重要な素性になると考えられる．

提案手法のアーキテクチャを図 2 に示す．本手法では，節タイトル，引用文脈，脚注の文[6] を [SEP] で連結し入力とする．そして Zhao らと同様にマルチタスク学習の枠組みでモデル全体を学習していく．

5　実験

5.1　データセットの作成

実験用データセットを作成した．2000〜2021 年の ACL, NAACL, EMNLP における本会議論文を ACL Anthology[7] から収集し，PDFNLT[8] [33] でテキスト化した．収集論文は 15,761 件であった．論文テキストに対し，URL[9]，脚注を参照する本文上の脚注番号，参考文献の引用アンカーを検出した．検出結果を用いて 3 節で述べた 3 種類の URL 引用に対

5）例えば，引用箇所が導入に当たる節である場合，リソースの役割が Supplement，引用機能が Introduce である可能性が高いと考えられる．
6）脚注を用いない URL 引用，および，URL のみを単体で脚注で記載していた場合は，[SEP] と脚注の文は連結しない．
7）https://aclanthology.org/
8）https://github.com/KMCS-NII/PDFNLT-1.0
9）“http://”，“https://”，“ftp://” のいずれかで始まる文字列

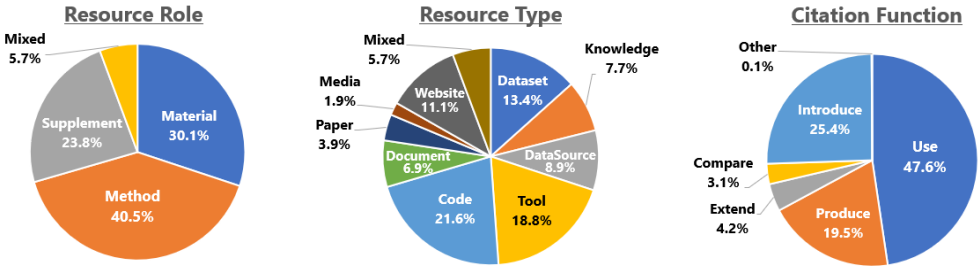

図 3　作成したデータセットにおけるラベルの分布

素性名	内容
節タイトル	Introduction
引用文脈	Online news platforms such as Google News [CITE] and MSN News have gained huge popularity for online digital news reading . Tens of thousands of news articles are streamed from ...
脚注	[CITE]

Resource Role : Supplement
Resource Type : Website
Citation Function : Introduce

素性名	内容
節タイトル	Experiments
引用文脈	.... was declared frozen before running with the formal evaluation data. All numbers reported here reflect this frozen system. [CITE]
脚注	The code and data are available from [CITE] , for replicability .

Resource Role : Mixed
Resource Type : Mixed
Citation Function : Produce

図 4　データセットの例

表 3　実験結果

Method	Resource Role	Resource Type	Citation Function
ベースライン	0.539	0.197	0.470
提案手法	0.613	0.255	0.446
- w/o 節タイトル	0.556 (↓)	0.258 (↑)	0.477 (↑)
- w/o 脚注の文	0.512 (↓)	0.279 (↑)	0.371 (↓)

し，本文上の対応する段落文を引用文脈として抽出した．ランダムに選択した 65 件の論文を用いて引用箇所の特定精度を評価した結果，適合率が 0.995 (199/200)，再現率が 0.948 (199/210) であった．

抽出した引用文脈を用いて自然言語処理の専門家がアノテーションを行った．データセットは 865 件の URL 引用（論文 287 編）を含む．図 3 にラベルの分布を示す[10]．図 4 にデータセットの例[11]を示す．

5.2　実験設定

作成したデータセットを用いて．提案手法の分類性能を検証した．データセットはランダムに分割し，学習セットを 519 件，検証セットを 173 件，テストセットを 173 件とした．ベースラインは Zhao ら [5] の SciResCLF である．ベースライン，提案手法ともに入力のエンコーダとして SciBERT[34] を採用した．損失関数には各タスクのクロスエントロピーロスの加重和を用いた．最適化関数は Adam[35] である．両手法ともに，各ハイパーパラメータの候補について最大で 50 エポック分学習し，各エポックで検証セットに対する分類性能を算出した[12]．各タスクについて，検証セットでの分類性能が最良であったモデルをテストセットへ適用し，手法を評価した．分類性能の評価にはマクロ平均 F1 値を用いた．

5.3　実験結果

実験結果を表 3 に示す．Resource Role/Type の分類において，提案手法はベースラインの性能を上回った．表 3 における提案手法より下の行では，各素性を除いた場合の結果を示している[13]．表 3 より，Resource Role の分類においては節タイトル，脚注ともに有効であることがわかる．Citation Function の分類では，脚注が素性として有効な一方で，節タイトルは分類性能を悪化させている．Resource Type の分類では各素性を抜くことで性能が向上しており，両者とも単体での使用は有効であることがわかる．これらの結果はタスクごとに有効な素性の組み合わせが異なることを示しており，引用文脈以外の素性を加えるアプローチとマルチタスク学習との組み合わせ方法について今後検証していく必要がある．

6　おわりに

本論文では，論文における URL 引用に対し，(1) 参照されたリソースが研究で果たす役割，(2) 参照されたリソースの種類，(3) 引用が行われた理由 を求める分類問題に取り組んだ．既存研究ではマルチタスク学習に基づく分類手法が提案されていたが，入力には引用文脈のみを用いていた．本論文では，節タイトルと脚注のテキストを入力素性として用いる分類手法を提案した．実験結果から Resource Role の分類において提案手法の有効性を確認した．

10)　記載方法別の内訳は，脚注が 0.79，本文が 0.11，参考文献の書誌情報が 0.10 であった．
11)　Zhao ら [5] と同様に，引用箇所や引用に用いた URL 自体を [CITE] に置き換えている．
12)　詳細は実装とともに付録へ記載
13)　数値が提案手法より下回る場合は “(↓)” を，上回る場合は “(↑)” を付与している．

謝辞

本研究は，一部，科学研究費補助金（基盤研究（B））（No. 21H03773）により実施したものである．

参考文献

[1] The Australian National Data Service. What is research data, 2017（2022-01 閲覧）. https://www.ands.org.au/__data/assets/pdf_file/0006/731823/Whatis-research-data.pdf.

[2] Association for Computing Machinery. Artifact review and badging – version 2.0, 2021（2022-01 閲覧）. https://www.acm.org/publications/policies/artifact-review-badging.

[3] Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In **WWW '19: The World Wide Web Conference**, pp. 1365–1375, 2019.

[4] Shunsuke Kozawa, Hitomi Tohyama, Kiyotaka Uchimoto, and Shigeki Matsubara. Collection of usage information for language resources from academic articles. In **Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)**, pp. 1227–1232, 2010.

[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, 2019.

[6] 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, 2019.

[7] Eugene Garfield. Can citation indexing be automated? In **Statistical association methods for mechanized documentation, Symposium Proceedings**, pp. 189–192, 1964.

[8] Michael J. Moravcsik and Poovanalingam Murugesan. Some results on the function and quality of citations. **Social Studies of Science**, Vol. 5, No. 1, pp. 86–92, 1975.

[9] Ina Spiegel-Rosing. Science studies: Bibliometric and content analysis. **Social Studies of Science**, Vol. 7, pp. 97–113, 1977.

[10] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In **Proceedings of the 2006 conference on empirical methods in natural language processing**, pp. 103–110, 2006.

[11] 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, 2013.

[12] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 391–406, 2018.

[13] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. 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. 3586—3596, 2019.

[14] 柴田大輔, 芳鐘冬樹. 学術文献における引用分類の観点. 情報知識学会誌, Vol. 26, No. 3, pp. 277–296, 2016.

[15] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. Content-based citation analysis: The next generation of citation analysis. **Journal of the association for information science and technology**, Vol. 65, No. 9, pp. 1820–1833, 2014.

[16] Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. BACO: A background knowledge- and content-based framework for citing sentence generation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1466–1478, 2021.

[17] Data Citation Synthesis Group. **Joint Declaration of Data Citation Principles**. FORCE11, 2014. https://doi.org/10.25490/a97f-egyk.

[18] Arfon M Smith, Daniel S Katz, and Kyle E Niemeyer. Software citation principles. **PeerJ Computer Science**, Vol. 2, p. e86, 2016.

[19] James Howison and Julia Bullard. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. **Journal of the Association for Information Science and Technology**, Vol. 67, No. 9, pp. 2137–2155, September 2016.

[20] Frank Krüger and David Schindler. A literature review on methods for the extraction of usage statements of software and data. **Computing in Science Engineering**, Vol. 22, No. 1, pp. 26–38, 2020.

[21] Ayush Singhal and Jaideep Srivastava. Data extract: Mining context from the web for dataset extraction. **International Journal of Machine Learning and Computing**, Vol. 3, No. 2, pp. 219–223, 2013.

[22] Daisuke Ikeda and Yuta Taniguchi. Toward automatic identification of dataset names in scholarly articles. In **Proceedings of the 8th International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 379–382, 2019.

[23] Animesh Prasad, Chenglei Si, and Min-Yen Kan. Dataset mention extraction and classification. In **Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications (ESSP)**, pp. 31–36, 2019.

[24] Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. **Journal of the Association for Information Science and Technology**, Vol. 72, No. 7, pp. 870–884, 2021.

[25] David Schindler, Benjamin Zapilko, and Frank Krüger. Investigating software usage in the social sciences: A knowledge graph approach. In **Proceedings of the 17th European Semantic Web Conference Semantic Web (The Semantic Web)**, pp. 271–286, 2020.

[26] Kai Li and Erjia Yan. Co-mention network of R packages: Scientific impact and clustering structure. **Journal of Informetrics**, Vol. 12, No. 1, pp. 87–100, 2018.

[27] Tomoki Ikoma and Shigeki Matsubara. Identification of research data references based on citation contexts. In **Proceedings of the 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)**, pp. 149–156, 2020.

[28] 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 (EEKE2021) at JCDL2021**, pp. 20–25, 2021.

[29] Yasunori Yamamoto and Toshihisa Takagi. OReFiL: an online resource finder for life sciences. **BMC bioinformatics**, Vol. 8, No. 1, pp. 1–8, 2007.

[30] Monarch Parmar, Naman Jain, Pranjali Jain, P Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. NLPExplorer: Exploring the universe of NLP papers. **Advances in Information Retrieval**, Vol. 12036, pp. 476–480, 2020.

[31] Hidetsugu Nanba. Construction of an academic resource repository. In **Proceedings of Toward Effective Support for Academic Information Search Workshop at ICADL 2018**, pp. 8–14, 2018.

[32] John Cullars. Citation characteristics of italian and spanish literary monographs. **The Library Quarterly**, Vol. 60, No. 4, pp. 337–356, 1990.

[33] 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 (COLING 2016)**, pp. 136–140, 2016.

[34] 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, 2019.

[35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of the 3rd International Conference for Learning Representations (ICLR)**, 2015.

[36] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In **Proceedings of the 18th BioNLP Workshop and Shared Task**, pp. 319–327, 2019.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In **Advances in Neural Information Processing Systems 32 (NeurIPS 2019)**, pp. 8024–8035, 2019.

A　分類ラベルの設定に関する補足

研究データのメタデータ生成を想定する場合，一定の具体性を持ったリソースの種類を定義しその分類可能性を検証したい．しかし，Zhao ら [5] の設定では Material（Resource Role の 1 つ）に対応するリソースの種類が Data のみであった．500 件の引用を対象とした調査と議論のうえで，Data を Dataset, Knowledge, DataSource へ分割した．

B　実験設定の詳細

実験では，両手法ともに下記のハイパーパラメータの組み合わせを試した．

- バッチサイズ： 16, 32, 64
- 学習率： 1.0e-4, 5.0e-5, 1.0e-5, 5.0e-6
- 引用文脈の範囲[14)]： 1 文，前後 1 文を含めた 3 文，前後 2 文を含めた 5 文
- ドロップアウト率： 0.0, 0.3, 0.6
- 入力の最大系列長： 256

学習は最大で 50 エポック行うが，10 回以内に検証セットに対する最小ロスを更新しなかった場合は学習を終了する．また，ロスにおける各タスクの重みは等しく 1.0 としている．引用文脈の文分割には ScispaCy[15)] [36] を用いた．なお，入力に加えている脚注は URL が記載されている 1 文としている．実装には，PyTorch[37][16)] を用いた．

14） 本論文では，前後 2 文を引用文脈に含む場合であっても，同じ段落に含まれる文までに制限した．
15） https://github.com/allenai/scispacy
16） https://pytorch.org/docs/1.8.1/