論文の文献リストにおける研究データ引用の検出

生駒流季 [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 情報基盤センター
ikoma.tomoki.d0@s.mail.nagoya-u.ac.jp

概要

オープンサイエンスの機運の高まりとともに，研究データに関する情報を集約した研究データリポジトリの構築が進められている．しかし，その構築には多大な時間と労力を要する．本論文では，研究データリポジトリの自動構築に向けて，学術論文の参考文献リストから，研究データとして引用されている文献を検出する手法を提案する．引用されている文献の内容に言及する文字列を特定し，その周辺のテキストも用いて，研究データを検出するモデルを学習する．実験により，本手法の有効性を確認した．

1　はじめに

オープンサイエンスの機運の高まりとともに，研究において作成されたデータやツールなど，研究データの共有や再利用への需要が高まっている．自然言語処理の分野では，LDC[1]，CLARIN[2]，ISLRN[3, 4] など複数の機関が研究データの情報を収集し，研究データリポジトリを整備している．しかし，リポジトリの多くは人手で構築されており，多大な時間と労力を要するという問題がある．

研究データリポジトリの構築を自動化するために，学術論文からその研究において用いられた研究データの情報を取得し活用することが一つの方法である．論文の文献リストには通常の文献だけでなく，研究データを参照する文献が含まれることがあり，その情報を利用できる可能性がある．

本論文では，学術論文の参考文献リストから，研究データとして引用されている文献を検出する手法を提案する（図 1）．研究データの引用を検出するには，各文献が引用されている箇所の周辺テキストである引用文脈の情報が利用できる．引用文脈には，引用された文献の内容を表す記述と，引用した論文との関係を表す記述が含まれる．両者はそれぞれ異なる観点を示すため，研究データ引用の検出に，それらを区別して利用することが重要となる．

図 1　研究データ引用の検出（言語資源に関する国際会議 LREC では，研究データと通常の文献への引用が分別して文献リストに記載される）

本手法では，引用されている文献の内容に言及する本文上のテキストを特定し，その周辺のテキストも用いて，研究データを検出するモデルを学習する．自然言語処理分野の国際会議の発表論文を使用した実験により，本手法の有効性を確認した．また，本手法を広範な分野の論文に適用し，様々な分野の研究データ引用の検出可能性を確認した．

2　関連研究

学術論文から利用された研究データの情報を取得して集約する研究として，Tohyama ら [5] は，既存の複数のリポジトリからメタデータを自動収集し，多様な研究データの情報を検索できる環境を構築した．また，Kozawa ら [6] は，研究データの名称を含む文を学術論文から抽出し，その研究データの利用目的に関する記述を抜き出し，リポジトリの構築に利用している．Ikoma ら [7] は，人手で設計した複数の素性項目に基づき，引用されている文献を研究

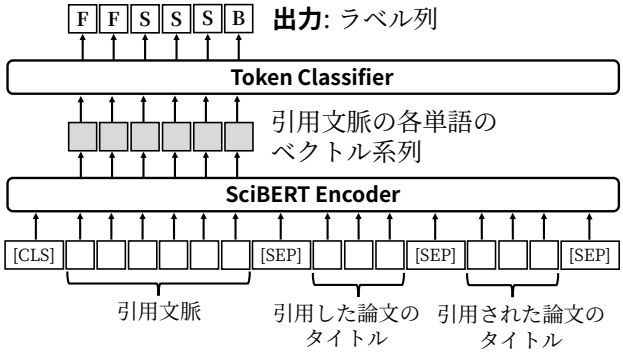

図 2　引用タグ範囲の抽出におけるモデルの入出力

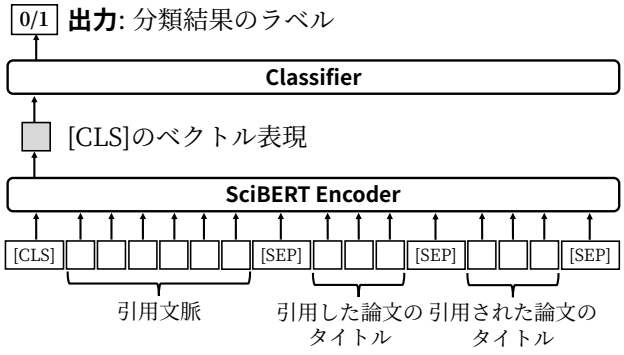

図 3　引用対象の分類におけるモデルの入出力

データと通常の文献に分類する手法を提案した．

なお，同様の目的の異なるアプローチとしては，Tsunokake ら [8] の研究が挙げられる．Tsunokake らは，論文中の URL に着目し，URL 参照の対象を研究データとその他に分類する手法を提案している．本研究とは対象とする引用の種類が異なる．

3　研究データの引用の検出

本手法では，学術論文における参考文献リストに記載されている各文献を，研究データと通常の文献に分類する．

研究で利用された研究データは，その研究データの作成について記述された文献を引用することで示されることがある．その場合，通常の文献と同様に参考文献リストに記載される．そのため，論文の文献リストには研究データを示すものと，その他の通常の文献を示すものが混在する．一方，先進的な取り組みとして，言語資源に関する国際会議 LREC では 2016 年より，研究データと通常の文献への引用を分別して記載することを著者に対し求めている[1)]．本論文では，このような研究データと通常の文献への引用の分別を自動化する手法を提案する．

4　引用文脈の利用

引用されている文献が研究データであるか否かの判別には，その文献が引用されている箇所のテキストである引用文脈が利用できる．本論文において，引用文脈はその文献への引用を含む文（引用文）と定義する．引用文脈には，引用されている文献がどのようなものであるかを表す記述と，引用した論文とその文献がどのように関係するかを表す記述が含まれる．例えば，以下に示す引用文

1) http://lrec2016.lrec-conf.org/en/submission/authors-kit/

1. *Each construct was operationalized using at least two items for measurement and analysed using* **confirmatory factor analysis** (1).

では，文中の太字部が前者に，斜体部が後者に相当する．それぞれの情報は，引用対象が研究データであるか否かを判別するのに利用できる．

引用文中において，引用された文献の内容について記述された文字列を**引用タグ範囲**と呼ぶ．上述の引用文 1 では，太字部が文献 (1) の引用タグ範囲に相当する．引用タグ範囲とその他の部分の分かれ方は引用文によって異なり，上述の例では引用タグの直前の名詞句が該当するが，以下に示す引用文

2. For example, previous studies have found that **cross institutional collaboration supports the diffusion of innovations and new ideas within a field** (2).

では，太字で示した節が該当する．また，以下に示す引用文

3. **The NACP was formally recognized by the United States in 2002 under the mantle of the nation's overall climate change management strategy** (3).

のように，引用文の全体が該当することもある．

引用されている文献が研究データであるか否かを判別するには，各引用文のうち引用タグ範囲に該当する範囲を特定するとともに，引用タグ範囲の内外を区別して利用することが重要である．

5　提案手法

本手法では，SciBERT[9] をベースとしたモデルに引用文脈，引用した論文のタイトル，引用された論文のタイトルを入力して，引用対象を研究データと通常の文献に分類する．

モデルは 2 段階で学習する．第 1 段階では，引用

文脈から引用タグ範囲を抽出するタスクでモデルを学習する．第 2 段階では，引用対象を研究データと通常の文献に分類するタスクでモデルを学習する．

5.1 引用タグ範囲の抽出

引用文脈，引用した論文のタイトル，引用された論文のタイトルを SciBERT に入力し，以下の手順で引用文脈中の各単語が引用タグ範囲に含まれるか否かを予測する（図 2）．

1. 引用文脈中において，対象の文献に対応する引用タグを [REFTAG] と置換する．
2. 引用文脈，引用した論文のタイトル，引用された論文のタイトルを [SEP] で連接し，単語列に変換して入力系列を生成する．
3. 入力系列を Encoder に入力し，引用文脈中の各単語のベクトル表現を生成する．
4. 生成したベクトルを Token classifier に入力し，引用文脈中の各単語が引用タグ範囲の手前 (F)，内部 (S)，後続部 (B) のいずれに該当するかを示すラベル列を生成する．

5.2 研究データの引用の検出

引用文脈，引用した論文のタイトル，引用された論文のタイトルを入力し，以下の手順で引用対象を研究データと通常の文献に分類する（図 3）．

1. 引用文脈中において，対象の文献に対応する引用タグを [REFTAG] と置換する．
2. 引用文脈，引用した論文のタイトル，引用された論文のタイトルを [SEP] で連接し，単語列に変換して入力系列を生成する．
3. 入力系列を Encoder に入力し，入力した引用文脈やタイトル情報の特徴を表すベクトルを生成する．Encoder は，入力系列の全体の特徴量を [CLS] に対するベクトルとして出力する．
4. 生成された [CLS] に対応するベクトルを Classifier に入力し，引用対象が研究データか否かを表すラベルを出力する．

6 実験

6.1 データセット

実験データとして，LREC 2016, 2018, 2020, 2022 の予稿集[2)]を用いて構築したデータセットを使用した．2016 年以降の LREC の論文では，研究データ引用が Language Resource References (LRR) として，通常の文献引用 (BR) とは区別して記載されている．はじめに，2016 年以降の LREC 予稿集から，研究データ引用を 1 件以上含む論文 635 本の PDF を収集した．次に，論文解析ツール PDFNLT[3)] [10, 11] を用いて，各論文の PDF から文献リストと，引用を含む文を引用文脈として抽出した．引用タグに含まれている発行年と筆頭著者名をもとに，抽出した引用文脈と参考文献リストのエントリを対応付けた．

表 1 作成したデータセットのサイズ

	BR 件数	LRR 件数	合計件数
学習	8,802	1,489	10,291
開発	2,931	657	3,588
テスト	3,560	652	4,212

2016，2018，2020 年の論文を学習データとし，2022 年の論文を開発データとテストデータに等分した（表 1）．文献リストのうち，LRR への記載を正例，BR への記載を負例とした．

引用タグ範囲抽出の学習には Ikoma ら [12] が作成したデータを使用した．このデータは 3C Shared Task[13, 14] のデータをもとに作成されており，3,000 件の引用文に対して，引用タグ範囲に該当する部分が人手でアノテートされている．本論文では，そのうち 2,100 件の引用文を使用してモデルを学習した．

6.2 実験設定

研究データ引用を検出するモデルを，PyTorch[15] および HuggingFace Transformers[16] を利用して実装した．まず，Ikoma ら [12] が作成したデータを用いて，引用タグ範囲抽出のタスクで SciBERT を学習した．次に，学習データを用いて，引用された文献を研究データと通常の文献に分類するタスクでモデルを 10 エポック学習した．F 値の算出に用いる適合率及び再現率は，以下の通りとした．

適合率 モデルが LRR として分類した文献のうち，正しいものの割合．

再現率 データセット中で LRR として引用されている文献のうち，正しく分類されたものの割合．

6.3 引用タグ範囲抽出のエポック数の設定

開発データにおける研究データ引用の検出性能に基づき，引用タグ範囲抽出のエポック数を設定し

2） https://aclanthology.org/venues/lrec/

3） https://github.com/KMCS-NII/PDFNLT-1.0

表 2 引用タグ範囲抽出のエポック数の設定

エポック数	適合率 (%)	再現率 (%)	F 値 (%)
1	60.08	44.44	51.09
2	62.73	50.99	**56.26**
3	68.28	47.18	55.80
4	62.56	43.23	51.13
5	62.28	46.27	53.24

表 3 性能評価の結果

	適合率 (%)	再現率 (%)	F 値 (%)
ベースライン	68.77	40.18	50.73
提案手法	56.22	51.99	**54.22**

た．引用タグ範囲抽出のタスクで SciBERT を最大 5 エポック学習し，各エポック末のモデルの状態をチェックポイントとして保存した．各チェックポイントを起点として，引用されている文献を研究データと通常の文献に分類するモデルを学習し，開発データにおける F 値が 10 エポック以内で最良となった時点でのモデルを保存して性能を比較した．

表 2 に結果を示す．最良の性能は，エポック数を 2 としたときに記録された．

6.4 実験結果

6.3 節の結果に基づき，引用タグ範囲抽出を 2 エポック学習した状態の SciBERT を起点として，引用されている文献を研究データと通常の文献に分類するモデルを学習した．引用タグ範囲抽出の学習をせずに学習したモデルをベースラインとして，分類の性能を比較した．

結果を表 3 に示す．提案手法のモデルはベースラインより F 値で約 3.5 ポイントの性能上昇を示し，本手法の有効性が示された．

6.5 検出例

ベースラインでは，引用された文献が研究データであることを表す語句が引用文脈中に含まれている場合でも，その文献が研究データとして検出されないことが散見された．例えば，以下に示す引用文脈

4. The How2Sign Dataset (4) contains 83 hours of instructional videos from How2 19 trans-lated from English into American SL (ASL).

において，(4) として引用されている文献は，ベースラインでは研究データとして検出されなかった．対して提案手法では，引用対象について記述された部分を特定することで，当該の文献を研究データとして検出することができた．引用タグ範囲を抽出するタスクによる学習には，引用対象について記述された範囲と、その周辺の記載事項に基づいたうえで研究データ引用を検出できるようにする効果があると考えられる．

7 多分野の論文コーパスへの応用

本手法を多分野の論文コーパスに適用し，様々な分野の研究データの情報の取得を試みた．

広範な分野の論文を収蔵した論文コーパス S2ORC[17] から 10,000 本の論文を無作為に選出し，各論文のテキストからのべ 394,956 件の引用を抽出した．引用文脈，引用した論文のタイトル，引用された文献のタイトルを本手法のモデルに入力し，研究データ引用を検出した．

抽出した引用のうち，研究データとして検出された事例は 1,866 件あった．例えば，以下の引用文脈

5. Speech files for the evaluation are taken from the TIMIT corpus (5) where 20 files are chosen randomly.

からは，(5) として引用されている文献 “TIMIT acoustic-phonetic continuous speech corpus” が，研究データとして検出された．以下の引用文脈

6. The experimental data set (Australian Credit Approval Data Set) was taken from the UCI repository(6), which has 690 samples, each with 8 symbolic features and 6 numerical features.

からは，(6) として引用されている文献 “UCI repository of machine learning databases” が検出された．このように，本手法を用いることで，様々な分野の研究データを検出することができた．

8 まとめ

本論文では，論文の文献リストから，研究データとして引用されている文献を検出する手法を提案した．本手法では，引用されている文献の内容に言及する本文上のテキストを特定し，その周辺のテキストも用いて，研究データを検出するモデルを学習する．自然言語処理分野の国際会議の発表論文を使用した実験により，本手法の有効性を確認した．また，本手法を広範な分野の論文に適用し，様々な分野の研究データ引用の検出可能性を確認した．

謝辞

本研究は JSPS 科研費 JP21H03773 および JST 次世代研究者挑戦的研究プログラム JPMJSP2125 の支援を受けて行われた．また，本研究における実験は名古屋大学のスーパーコンピュータ「不老」を使用して実施した．

参考文献

[1] Eleftheria Ahtaridis, Christopher Cieri, and Denise DiPersio. LDC language resource database: Building a bibliographic database. In **Proceedings of the 8th International Conference on Language Resources and Evaluation**, pp. 1723–1728, 2012.

[2] Claus Zinn. Squib: The language resource switchboard. **Computational Linguistics**, Vol. 44, No. 4, pp. 631–639, 2018.

[3] Khalid Choukri, Victoria Arranz, Olivier Hamon, and Jungyeul Park. Using the international standard language resource number: Practical and technical aspects. In **Proceedings of the 8th International Conference on Language Resources and Evaluation**, pp. 50–54, 2012.

[4] Valérie Mapelli, Vladimir Popescu, Lin Liu, and Khalid Choukri. Language resource citation: the ISLRN dissemination and further developments. In **Proceedings of the 10th International Conference on Language Resources and Evaluation**, pp. 1610–1613, 2016.

[5] Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, and Hitoshi Isahara. Construction of an infrastructure for providing users with suitable language resources. In **Proceedings of the 22nd International Conference on Computational Linguistics 2008: Companion volume: Posters**, pp. 119–122, 2008.

[6] 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**, pp. 1227–1232, 2010.

[7] 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: Digital Libraries at Times of Massive Societal Transition**, pp. 149–156, 2020.

[8] Masaya Tsunokake and Shigeki Matsubara. Classification of URLs citing research artifacts in scholarly documents based on distributed representations. In **Proceedings of the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents collocated with ACM/IEEE Joint Conference on Digital Libraries**, pp. 20–25, 2021.

[9] 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**, pp. 3615–3620, 2019.

[10] 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, 2016.

[11] Kenichi Iwatsuki, Takeshi Sagara, Tadayoshi Hara, and Akiko Aizawa. Detecting in-line mathematical expressions in scientific documents. In **Proceedings of the 2017 ACM Symposium on Document Engineering**, p. 141–144, 2017.

[12] Tomoki Ikoma and Shigeki Matsubara. Identifying influential references in scholarly papers using citation contexts. In **Proceedings of the 25th International Conference on Asia-Pacific Digital Libraries: Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration**, pp. 152–161, 2023.

[13] Suchetha Nambanoor Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. Overview of the 2020 WOSP 3C citation context classification task. In **Proceedings of the 8th International Workshop on Mining Scientific Publications**, pp. 75–83, 2020.

[14] Suchetha Nambanoor Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. Overview of the 2021 SDP 3C citation context classification shared task. In **Proceedings of the 2nd Workshop on Scholarly Document Processing**, pp. 150–158, 2021.

[15] Adam Paszke et al. PyTorch: An imperative style, high-performance deep learning library. In **Advances in Neural Information Processing Systems 32**, pp. 8024–8035, 2019.

[16] Wolf Thomas et al. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 38–45, 2020.

[17] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4969–4983, 2020.