4N-06

# 研究データを参照する文献の引用文脈に基づく識別

生駒 流季 † 松原 茂樹 †‡

† 名古屋大学 大学院情報学研究科 ‡ 名古屋大学 情報連携統括本部

## 1 序論

学術論文では，本文で引用した論文や著書が参考文献として参照される．近年，データセントリックなアプローチが進み，論文において研究データが参照されることも多い．このような研究データの参照を，他の文献の参照と区別して示すことは，研究データの共有や再利用の促進に有意義である．

本稿では，論文における参考文献リストのエントリにおいて，研究データを他の文献と識別することを試みる．本方式では，参考文献リストの書誌要素（文献情報）に加え，本文における引用の文脈情報を利用する．国際会議論文を用いた識別実験を実施し，文脈情報利用の有効性を検証した．

## 2 文献リストにおける研究データの参照

学術論文の参考文献リストにおける論文や著書の書誌要素の書き方は，統一的に規定されている．一方，研究データを文献リストに記載する方法は，学会や会議で明確な規定が存在しないことも多く，著者の裁量に委ねられることも少なくない．実際，研究データの文献リストでの参照には，

- 研究データの作成に関する学術論文を記載する
- Web 上の研究データの URL を記載する
- 研究データの仕様書や利用ガイドを記載する

など，さまざまな様式が存在する．このため，研究データの利用に関心のある読者にとって，研究データの参照が他の文献と区別して示されることは有用である．そのような事例として，言語資源に関する最大の国際会議である LREC では，2016 年より，予稿集に掲載する論文において，通常の文献参照を記す Bibliographical References（以下，BRs）と言語資源（自然言語処理分野の研究データ）の参照を記す Language Resource References（以下，LRRs）の 2 つのリストに分割して記載することを著者に求めている [1]．本研究は，このようなリストの分割の自動化を目指すものであり，論文データからの研究データリポジトリの構築 [2, 3, 4] に応用できる．

表 1 書誌情報に関する手がかり表現

| 出現箇所 | 手がかり表現 | LRRs/出現数 (%) |
|---|---|---|
| 文献リスト | corpus/corpora | 32.5 (25/77) |
| 文献リスト | data/set/bank | 42.4 (25/59) |
| 文献リスト | 言語の名称 | 25.3 (25/99) |
| 文献リスト | university<br>institute<br>center/centre | 41.5 (17/41) |
| 文献リスト | proceedings<br>journal | 8.8 (22/249) |
| 文献リスト | http(s)://<br>www | 45.2 (14/31) |
| 文献リスト | LDC/CLARIN<br>ISLRN/LREC | 41.8 (28/67) |

表 2 引用文脈に関する手がかり表現

| 出現箇所 | 手がかり表現 | LRRs/出現数 (%) |
|---|---|---|
| 節タイトル | corpus/corpora | 22.9 (22/96) |
| 節タイトル | data/set/bank | 45.2 (19/42) |
| 節タイトル | introduction<br>related work<br>conclusion | 10.0 (24/229) |
| 節タイトル | experiment<br>evaluation | 9.1 (5/55) |
| 文中の引用 | corpus/corpora | 21.9 (23/105) |
| 文中の引用 | data/set/bank | 22.5 (16/71) |
| 文中の引用 | tool/parser | 3.2 (1/31) |
| 文中の引用 | word embedding<br>word2vec/WordNet | 12.8 (5/39) |
| 文中の引用 | we | 14.1 (37/253) |
| 文中の引用 | they | 5.1 (2/39) |
| 文中の引用 | use/adopt/apply/etc. と<br>言語資源を表す語が出現 | 25.6 (22/86) |
| 文中の引用 | 引用タグの文頭への出現 | 7.1 (8/112) |
| 文中の引用 | 引用タグの文末への出現 | 9.3 (25/270) |

## 3 研究データの参照に関する調査

論文の文献リストにおける研究データの参照の特徴を分析した．LREC 2016 及び 2018 の予稿集 [5] に掲載の論文のうち，LRRs にエントリのある 265 論文を均等かつ無作為に 10 分割し（ブロック 0〜9），ブロック 8 を調査データとした．

調査データの文献リストには 588 エントリあり，その内訳は BRs が 505 件 (85.9%)，LRRs が 83 件 (14.1%) であった．LRRs のエントリについて，書誌要素上の手がかり表現，及び，本文で引用された文中の手がかり表現（文脈情報）を取り出し，手がかり表現が出現した文献数における LRRs の占める割合 (LRRs/出現数) を算出した．

Discriminating References to Research Data Based on Citation Context
Tomoki Ikoma, Nagoya University
Shigeki Matsubara, Nagoya University

### 3.1 書誌情報に関する手がかり表現

表 1 に，書誌要素から取り出された手がかり表現と，その LRRs/出現数の値を示す．書誌情報では，以下の特徴が観察された．

**言語資源の種類や言語の名称を表す語** 言語資源の構築について記した論文は，“Corpus of reading comprehension exercises in German” のように，タイトルに corpus や data など言語資源の種類や，言語の名称を表す語を含むことが多い．

**URL の記載有無** 通常の文献の参照では，論文誌名や掲載ページ番号，学会名などが記載される．一方，言語資源の参照では，URL が記載されやすい．

### 3.2 引用文脈に関する手がかり表現

表 2 に，文脈情報から取り出された手がかり表現を示す．文脈情報では，以下の特徴が観察された．

**引用文を含む節のタイトル** 研究で利用した言語資源は，実験の設定を記述した節で言及されることが多い．一方，Introduction や Related Work の節での文献の引用では，手法や前提の言及が多く，言語資源への参照には該当しない場合が多い．

**言語資源の種類を表す語** 言語資源を参照する場合，引用タグは言語資源の種類を表す語の直後に現れることが多い．例えば，前述の文献 “Corpus of reading comprehension exercises in German” は，本文中では “Second, we provide POS and normalization annotation on top of the CREG Corpus (Merrers et al., 2011).” という文脈で参照されており，言語資源の種類 “corpus” の直後に引用タグが出現している．

**文頭に出現する引用タグ** 引用タグが文頭に出現している場合，引用タグはその文の主語を担っており，文全体がその文献について言及している場合が多く，通常の文献である可能性が高い．例えば，“Selinker (1972) coined the term interlanguage for these language variants of individual learners.” という文は Selinker(1972) の文献が提唱する概念を提示するものであり，通常の文献として参照されている．

以上の通り，書誌要素だけでなく，本文にも研究データの参照を識別する手がかりとなる表現が出現することを確認した．

## 4 研究データ参照文献の識別手法

本研究で提案する手法では，以下の手順によって文献リストのエントリを分類する．

1. PDFNLT ツール [6] を用いて，論文の PDF ファイルからテキストと文献リストを抽出する．
2. テキストから引用タグを抽出し，文献リストのエントリと対応付ける．
3. テキストおよび文献リストの記載内容から識別のための素性を抽出し，SVM により分類する．

識別に用いる素性は，表 1 および表 2 に示した手掛かり表現の出現の有無とした．

表 3 実験結果

| | 適合率 (%) | 再現率 (%) | F 値 |
|---|---|---|---|
| ベースライン | 73.7 (28/38) | 23.5 (28/119) | 35.7 |
| 提案手法 | 62.7 (32/51) | 26.9 (32/119) | 37.6 |

## 5 実験

### 5.1 実験の概要

研究データ参照の識別における文脈情報の有用性を検証するため，実験を実施した．手がかり表現を素性とし，通常の文献と研究データに分類する SVM を，scikit-learn[7] の SVM モジュールを用いて実現した．調査のために作成したデータ（3 節参照）のうち，ブロック 9 をテストデータとして，残りのブロックを学習データとして使用し，研究データの参照の識別性能を求めた．また，取り出した手がかり表現のうち，書誌要素に関するもの (表 1) のみを素性として用いる手法をベースラインとした．

### 5.2 実験結果

研究データ参照の識別実験の結果を表 3 に示す．精度は出力数に対する正解数の比を，再現率は参照数に対する正解数の比を，F 値は両者の調和平均を表す．ベースラインと比べ高い識別性能を示しており，本文中における引用文脈を利用することの有効性を確認した．

## 6 結論

本稿では，学術論文で引用された研究データを，通常の文献と区別して提示することを目的に，文献リストにおける研究データ参照の識別について述べた．国際会議論文を用いた識別実験を実施し，文脈情報利用の有効性を確認した．

## 参考文献


[1] LREC Author’s kit, 2016. http://lrec2016.lrec-conf.org/en/submission/authors-kit/

[2] 難波 英嗣. Web 上の学術リソースリポジトリの構築. 言語処理学会第 25 回年次大会発表論文集, 2019.

[3] 小澤 俊介，遠山 仁美，内元 清貴，松原 茂樹. 言語資源の用途情報の獲得と利用. 電子情報通信学会論文誌, Vol. 95, No. 7, pp. 611–622, 2012.

[4] H. Toyama et al. Construction of an infrastructure for providing users with suitable language resources. In *Proceedings of 22th International Conference on Computational Linguistics*, pp. 119–122, 2008.

[5] N. Calzolari et al. (eds.) *Proceedings of LREC 2016 and 2018*. http://www.lrec-conf.org/proceedings/

[6] PDFNLT. https://github.com/KMCS-NII/PDFNLT-1.0/tree/master/pdfanalyzer.

[7] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.