学術論文で引用された Web 上の研究データの同定と分類

角掛正弥† 松原茂樹‡
† 名古屋大学工学部電気電子・情報工学科 ‡ 名古屋大学情報連携統括本部
tsunokake.masaya@a.mbox.nagoya-u.ac

1. まえがき

オープンサイエンスは、論文や研究データ[*1]の参照や利活用を促進するための活動である。この促進において重要な役割を担うのが論文における論文や研究データの引用である。このうち、論文の引用については、

- 引用された論文は文献リストに列挙される、及び、
- 論文の書誌要素（文献情報）の記法は定まっている[*2]

ことから、引用された論文の著者や題目の取得、種類の判別（雑誌論文か会議予稿か、など）も機械的に行える。一方、研究データの引用については、統一的な規定がなく、その書誌要素の記載箇所や記載方法は、著者の裁量に委ねられていることも多い。

これまで、著者ら [1] は論文から研究データのメタデータを抽出することを目指し、論文に記載された URL から研究データの同定とその種類の判別を行っている。URL の引用文脈からその分散表現を獲得し、それを入力素性とした分類器により実現している。しかし、アーカイブサイトなどの論文を指す URL は、研究データを指す URL と類似した文脈で出現することがあり、分類が困難になるという問題がある。

本稿では、URL の引用文脈だけでなく、URL を構成する文字列も考慮し、論文に記載された URL から研究データの同定とその種類の判別を行う手法について述べる。本手法では URL を構成するディレクトリ名などの文字列に意味があることに着目する。URL の文字列をドメイン名やディレクトリ名などの URL の構成要素に分解し、要素ごとに分散表現を獲得する。それらの分散表現を入力素性に活用した分類器で、研究データの同定と種類の判別を実行する。

2. 論文において研究データを引用する URL

2.1 研究データのリポジトリ

オープンデータはオープンサイエンスの動きの一つであり、研究データの共有による研究の加速化や研究データへのアクセス促進を図る運動である。また、近年はデータ中心科学が広まり、論文で研究データを引用するケースが増えている。論文で引用された研究データを機械的に収集し、リポジトリとして整備できれば、研究データの有効活用につながる。研究データをリポジトリとして整備するには、研究データの各種情報を表すメタデータが必要となる。図 1 にメタデータの例を示す。これまでに、小澤ら [2] は、学術論文から研究データの「用途」を自動抽出する手法を提案している。一方、本研究では、研究データの「種類」の抽出を目指す。

メタデータの属性一覧
研究データの名称
対応するURL群
作成者
帰属
作成時期
種類
用途
他の研究データとの関係
被引用論文

図 1 研究データリポジトリのメタデータ例

2.2 研究データの引用

研究データの引用には統一的な規定がなく、参考文献として列挙される場合もあれば URL の記載により引用される場合もある。生駒ら [3] は、参考文献から研究データを識別することを目的としている。本研究では、研究データの多くがインターネット上で利用可能であることから、URL で引用された研究データに焦点を当てる。しかし、論文に記載された URL の全てが研究データであるとは限らない。そこで本研究では、論文中の URL について、研究データであるか否か、研究データであればそれがツールであるかデータであるか、を判別するという 2 つの分類タスクに取り組む。

2.3 論文に記載された URL の予備調査

論文で記載された URL について予備調査を行った。ACL の本会議予稿集 2010〜2019 年分を ACL Anthology[*3]から取得した。さらに PDFNLT-1.0[*4][4] を用いて、PDF ファイルをその構造情報[*5]が保持されたテキストに変換し、3,837 件の論文データを作成した。

URL として “http://”、“https://”、“ftp://” で始ま

[*1] デジタル資料、計測データ、試験データ、プログラムなど、研究の実施や結果として収集・生成されたデジタル情報。
[*2] 独立行政法人科学技術振興機構：参考文献の役割と書き方，2011. https://jipsti.jst.go.jp/sist/pdf/SIST_booklet2011.pdf
[*3] https://www.aclweb.org/anthology/
[*4] https://github.com/KMCS-NII/PDFNLT-1.0
[*5] タイトル、著者、本文、図表、キャプション、脚注、参考文献リストといった論文を構成する要素

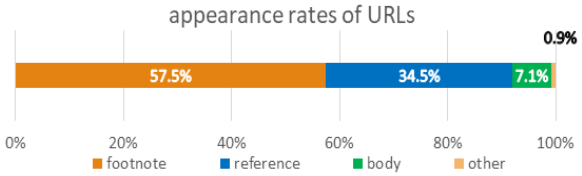

図 2　論文データにおける URL の記載箇所

る文字列を論文データから抽出した。URL の出現数は 12,568 件 (種類数 9,480 件) で、1 論文あたり平均 3.28 件であった。その記載箇所の内訳を図 2 に示す。URL の多くが脚注や参考文献リストに記載されている。

3.　研究データを引用する URL の同定手法

3.1　URL の分類問題

論文中の URL から研究データを引用する URL を同定し、その「種類」を判別する。本研究では URL を以下の 3 つに分類する多クラス分類タスクとして実現した。

- tool: コード、プログラム、ソフトウェア、ツールキット、API など
  例 `https://nlp.stanford.edu/projects/glove/`
  `https://github.com/google-research/bert`
  `http://www.nltk.org/`
- data: データ資源や知識のソースなど
  例 `http://qwone.com/~jason/20Newsgroups/`
  `http://babelnet.org`
  `http://answers.yahoo.com`
- other: 研究データを指し示さないサイト
  例 `http://arxiv.org/abs/1301.3781`
  `https://www.mturk.com`

3.2　URL の分散表現の獲得と利用

論文で URL がどのような目的で引用されたのかがわかれば、tool、data、other のいずれかに適切に分類できる。この実現のために、URL の引用文脈を利用することが考えられる。

3.2.1　URL 単位の分散表現

これまでに、著者ら [1] は 4.1 節の URL 分類問題に対して、論文に記載された URL の引用文脈を分散表現として獲得し、それを入力素性として分類する手法（以下、**従来手法**）を提案している。分散表現は、難波 [5] と同様に、word2vec[6] を用いて獲得している。従来手法 [1] では、この分散表現をそのまま分類器の入力素性として用いている。本稿ではこの入力素性を f_{url} と記す。すなわち、以下の手順で研究データを引用する URL を同定する。

1. 論文データの URL に一意の id を付与する。
2. URL を、対応する id を示すタグに変換する[*6]。
3. タグの分散表現を獲得する。
4. 分散表現を入力素性とし、URL を分類する。

3.2.2　URL 単位の分散表現による分類の問題点

前節で述べた手法では、URL を構成する文字列を考慮した分類ができない。例えば、“arxiv.org” などのアーカイブサイトといった論文を参照する URL は、研究データの引用元とされることがあり、研究データを指す URL と類似した文脈で出現する可能性がある。それを区別するために、URL を構成する文字列の情報も組み込んだ入力素性を検討する必要がある。

また、URL のリンク先の内容を判断する際、引用文脈だけでなく、URL を構成するドメイン名やディレクトリ名を考慮することが考えられる。例えば、

- `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

は、“tools” や “TreeTagger” という表現から、タグ付けツールを指す URL と推測できる。また、

- `http://trec.nist.gov/data/tweets/`

は “data” や “tweets” から、ツイートに関わるデータを指す URL と推測できる。

3.2.3　URL 要素単位の分散表現

URL を構成する文字列も考慮した分類を行う手法について述べる。本手法では、URL をドメイン名、ディレクトリ名、ファイル名、拡張子といった URL の構成要素 (以下、**URL 要素**) に分解する。論文に記載された URL を事前に URL 要素に分解し、URL 要素の分散表現を獲得する。これにより、“tools” や “data” といった URL 要素に固有の意味を学習することができる。分類対象となる URL が持つ、各 URL 要素の分散表現を基にした入力素性による分類を行う。すなわち、以下の手順で論文中の URL を分類する。

1. 論文データの各 URL を URL 要素へ分解する[*7]。
2. URL 要素に一意の id を付与する。
3. URL 要素を、対応する id を示すタグに変換する[*8]。
4. タグの分散表現を獲得する。
5. 獲得した URL 要素の分散表現に基づいた入力素性により、URL を分類する。

分類に用いる入力素性として、URL を構成する各 URL 要素の分散表現のベクトル和を提案する。この入力素性を f_{parts} とする。

[*6] 例えば、論文中の全ての “`http://nlp.stanford.edu/software/tagger.shtml`” をタグ “[URL2495]” に変換する。

[*7] 例えば、論文中の全ての “`http://nlp.stanford.edu/software/tagger.shtml`” を URL 要素 “nlp”、“stanford”、“edu”、“software”、“tagger”、“shtml” に分解する。

[*8] 例えば、論文中の全ての “nlp”、“stanford”、“edu”、“software”、“tagger”、“shtml” をそれぞれタグ “[PARTS7070]”、“[PARTS9479]”、“[PARTS3891]”、“[PARTS9344]”、“[PARTS9680]”、“[PARTS9182]” に変換する。

4. 実験

4.1 実験に用いたデータ

3 節の論文データを用いて分散表現を獲得する。予備調査の結果に基づき、脚注と参考文献リストのいずれかに出現した URL を対象とし、その引用文脈を捉えられるよう、以下の位置を機械的に特定した。

- 脚注に対応する本文での参照位置
- 参考文献に対応する本文での参照位置

本文での参照位置に URL を挿入後、URL をタグに変換したファイルと、URL 要素をタグに変換したファイル、の 2 種類を作った。例えば前者の場合、以下の文、

- We used the Maximum Entropy (MaxEnt) and Naive Bayes classifiers in the MALLET software package (McCallum, 2002) as initial baselines.

は、“(McCallum, 2002)” で引用される文献の書誌情報に “`http://mallet.cs.umass.edu`” が併記されているため、

- We used the Maximum Entropy (MaxEnt) and Naive Bayes classifiers in the MALLET software package [URL2229] as initial baselines.

と変換される。URL をタグに変換したファイルと、URL 要素をタグに変換したファイルのそれぞれで、各論文の本文を連結し、分散表現獲得ツールへの入力を 2 種類作成した。

分類器の学習や開発・テストに用いるデータセットを得るため、論文データ中の URL に頻度上位からラベル付けを行い、500 件の正解ラベル付き URL を作成した。500 件のうち 100 件を開発データとし、残りの 400 件をテストデータとした。

4.2 分類器の設定と分散表現の獲得

多クラス分類器のモデルとして one-versus-rest 法によるロジスティック回帰を用い、scikit-learn[*9] で実装した。

分散表現の獲得には word2vec[6] モデルを用いた。実装には gensim[*10]を用い、文分割、単語分割も同様に gensim により行った。word2vec のハイパーパラメータは、開発データに対する分類結果に基づき、各手法ごとに以下の計 40 組の組み合わせの中から最良のものを選んだ[*11]。

- epoch 数 10,20
- window 幅 5,10
- 分散表現のサイズ 100,200,⋯,900,1000

加えて各次元の特徴量を標準化するか否かも開発データに対する分類結果に基づき決定した。手法ごとに採用したハイパーパラメータを表 1 に示す。

表 1　手法ごとのハイパーパラメータ

入力素性	エポック数	窓幅	分散表現サイズ	特徴量の標準化
f_{url}	20	10	1000	なし
f_{parts}	10	5	800	あり

表 2　マクロ平均を基にした各評価値

入力素性	適合率	再現率	F1 値
f_{url}	0.772	0.760	0.766
f_{parts}	0.795	0.789	0.792

表 3　tool ラベルについての各評価値

入力素性	適合率	再現率	F1 値
f_{url}	0.771	0.845	0.804
f_{parts}	0.750	0.841	0.785

表 4　data ラベルについての各評価値

入力素性	適合率	再現率	F1 値
f_{url}	0.776	0.824	0.786
f_{parts}	0.739	0.710	0.714

表 5　other ラベルについての各評価値

入力素性	適合率	再現率	F1 値
f_{url}	0.768	0.611	0.668
f_{parts}	0.895	0.817	0.842

4.3 実験結果

テストデータである 400 件の URL に対して、3 つの手法を用いて 10 分割交差検定を行った。なお各回の学習に開発データも加えている。交差検定の各回でマクロ平均による適合率、再現率、それらの F1 値を算出した。その平均について手法ごとの結果を表 2 に示す。また、交差検定の各回でラベルごとの再現率、適合率、F1 値も計算した。tool、data、other ラベルについて、その平均をそれぞれ表 3、4、5 に示す。

4.4 考察

テストデータの URL における、各ラベルの割合は、tool が 39%、data が 33%、other が 28% であり、本分類タスクに URL の分散表現および URL 要素の分散表現を用いることの有効性を確認できた。

表 2 より、マクロ平均による適合率、再現率、それらの F1 値はいずれも $f_{url} < f_{parts}$ となっている。これは、表 5 の other ラベルでの評価値において、f_{parts} が f_{url} の評価値を大きく上回ったことが要因である。other に属する、“arxiv.org” などアーカイブサイトといった論文を指し示す URL は、ツールやデータの発表元や説明論文として引用されるケースがある。そのため、研究デー

[*9] `https://scikit-learn.org/stable/`

[*10] https://radimrehurek.com/gensim/

[*11] 除外する低頻単語の閾値を 3 にしたことを除いて、その他のハイパーパラメータはデフォルトのままである。

表 6　入力素性が f_{parts} である場合の、誤り先ラベルの割合

正解ラベル	誤り先ラベル		
	tool	data	other
tool	-	84%	16%
data	83%	-	17%
other	45%	55%	-

タを指す URL と類似した文脈で出現しうる。例えば、

- `https://doi.org/10.3115/v1/P14-5010`

は Manning ら [7] が Stanford CoreNLP toolkit[*12]の設計と使用方法について説明している論文へのリンクである。上記 URL は研究データの所在を指している訳ではないが、Stanford CoreNLP toolkit の所在を指す URL

- `https://stanfordnlp.github.io/CoreNLP/`

と類似した文脈で出現しやすい。実際、引用文脈のみから作成される入力素性 f_{url} の場合、分類器は “`https://doi.org/10.3115/v1/P14-5010`” に tool ラベルを割り振ったが、入力素性が f_{parts} であると other ラベルを割り振った。URL を URL 要素に分解し、その分散表現を入力素性に用いる提案手法では、引用文脈だけでなく、構成する文字列も考慮して URL を分類できている。

一方で tool ラベルや data ラベルでは、f_{parts} が f_{url} に劣っており、URL を構成する文字列を考慮することで精度が下がった。入力素性が f_{parts} である場合の誤り先ラベルの割合を、表 6 に示す。

tool ラベルの誤り先の多くが data ラベルであり、逆もまた同様であることから、tool ラベルと data ラベル間での分類が困難になっていることがわかる。URL 要素に分解することで分類精度が下がる原因として、複数の URL 上で出現する出現頻度の高い URL 要素の存在が考えられる。URL 要素の多くは、特定の URL に対して局所的にしか出現しないが、ドメイン名などは多くの URL で出現する。そのため、それらに分類が著しく支配されてしまう。例えば、URL 要素 “com” を持つ URL について、その正解ラベルの割合と、各手法における分類器の予測ラベルの割合を表 7 に示す。入力素性が f_{parts} であると、正解ラベルの割合に対し、tool の予測ラベルの割合が高い。URL を URL 要素に分解することにより、“com” が最も属する tool ラベルに URL を割り振りやすくなってしまっている。一方、URL 要素に分解することで、分類精度の上がった other ラベルに属する URL 集合では、“arxiv.org”、“aclweb.org/anthology”、“doi.org” といった other ラベル特有の表現が多く出現する。これらは他ラベルに属する URL にはあまり出現しないため、出現数が多いことが良い方向に働いたと考えられる。

表 7　“com” を含む URL における、正解ラベルおよび、各入力素性による分類器の予測ラベルの割合

	tool	data	other
正解ラベル	51.9%	35.7%	12.4%
f_{url} での予測ラベル	48.1%	37.2%	14.7%
f_{parts} での予測ラベル	62.0%	30.2%	7.8%

5. まとめ

本稿では、論文で引用された研究データのメタデータ抽出を目指し、論文に記載された URL から研究データを同定し分類する手法について述べた。URL の引用文脈から分散表現を獲得し入力素性とする手法を採用した。加えて、URL を構成する文字列自体に意味があることに着目し、URL を構成する各 URL 要素の分散表現を獲得し、入力素性に利用する手法を述べた。後者の手法を採ることによって、URL を構成する文字列を考慮した分類が行えることを確認した。

参考文献

[1] 角掛正弥, 松原茂樹. 論文において研究データを引用する URL の同定. 情報処理学会第 82 回全国大会 講演論文集, 2020.

[2] 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. J95-A, No. 7, pp. 611–622, 2012.

[3] 生駒流季, 松原茂樹. 研究データを参照する文献の引用文脈に基づく識別. 情報処理学会第 82 回全国大会講演論文集, 2020.

[4] Takeshi Abekawa and Akiko Aizawa. Sidenoter: scholarly paper browsing system based on pdf restructuring and text annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 136–140, 2016.

[5] 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, pp. 1483–1486, 2020.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[7] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[*12] `https://stanfordnlp.github.io/CoreNLP/`