4N-05

論文において研究データを引用するURLの同定

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 ‡

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

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。

本稿では、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う手法について述べる。本手法では、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する。

2. 論文における研究データの引用

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などを図る運動である。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。

2.2 研究データを引用するURL

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。

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

3. 論文に出現するURLの予備調査

3.1 調査データ

論文で記載されたURLについて予備調査を行った。自然言語処理分野の著名な国際会議であるACLの本会議予稿集2010〜2019年分をACL Anthology*2から取得した。さらにPDFNLT-1.0*3を用いて、PDFファイルをその構造情報*4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。

3.2 論文中のURL

URLとして “http://”、“https://”、“ftp://” で始ま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。記載箇所の内訳は、脚注が57.5%、参考文献リストが34.5%、本文が7.1%、とその多くが脚注や参考文献リストに記載されることがわかった。

4. 研究データを引用するURLの同定手法

4.1 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/`
- data: データ資源や知識のソースなど
  例 `http://qwone.com/~jason/20Newsgroups/`
- other: 研究データを指し示さないサイト
  例 `http://arxiv.org/abs/1301.3781`

Identification of URLs Citing Research Data in Scholarly Papers
† Masaya Tsunokake, Nagoya University
‡ Shigeki Matsubara, Nagoya University

*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 https://www.aclweb.org/anthology/
*3 https://github.com/KMCS-NII/PDFNLT-1.0
*4 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

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

4.2 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。この実現のためには、URL の引用文脈を利用することが考えられる。そこで本手法では、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を行う。分散表現の獲得には難波 [4] の方法を導入する。難波は学術リソースの所在である URL にキーワードを付与するため、word2vec[1] を用いて URL の分散表現を獲得している。また、URL の分散表現に対して類似度上位となる分散表現を持つ別の単語を、その URL のキーワードとしている。本研究では、この分散表現を分類器の入力素性として用いる。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. 論文データの各 URL に一意の id を付与する。
2. 各 URL を、対応する id を示すタグに変換する。
   例えば、論文中の全ての “`https://keras.io/`” をタグ “[URL8159]” に変換する。
3. 各タグの分散表現を獲得する。
4. 獲得した分散表現を入力素性とし、URL を分類する。

5. 実験

5.1 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 を挿入し、全 URL をタグに変換した。例えば以下の文、

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

は、“(McCallum, 2002)” が参照す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。その後、各論文の本文を連結し、分散表現獲得ツールへの入力を作成した。

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。

5.2 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[*5] で実装した。

分散表現の獲得には gensim[*6] の word2vec[1] モデルを用いた。文分割、単語分割も gensim を用いて行った。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決定した。その一部を表 1 に示す。

表 1　採用したハイパーパラメータ

エポック数	窓幅	分散表現サイズ	除外する低頻単語の閾値
20	10	1000	3

表 2　ラベルごとの各評価値

Type	Precision	Recall	F1-score
tool	0.771	0.845	0.804
data	0.776	0.824	0.786
other	0.768	0.611	0.668

5.3 実験結果

テストデータである 400 件の URL に対して 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。正解率は 0.768(307/400) であった。また、交差検定の各回でラベルごとの再現率、適合率、F1 値を計算した。その平均を表 2 に示す。

5.4 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、本分類タスクに分散表現を用いることの有効性を確認した。ラベルごとの結果を見ると、other の再現率が低い。これは異なるラベルに属する URL でも類似した引用文脈をとりうることが原因として考えられる。other ラベルが割り振られた URL には “arxiv.org” などアーカイブサイトが多い。論文を参照する URL は研究データの発表元として引用されることがあるため、研究データを指す URL と類似した文脈で出現し、分類が困難となった可能性がある。

6. 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アプローチを採用し、その有効性を確認した。今後の課題として、URL を構成する文字列を活用した手法の検討などがある。

参考文献

[1] 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.

[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] 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.

[*5] `https://scikit-learn.org/stable/`
[*6] https://radimrehurek.com/gensim/

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.