令和元年度　情報工学コース卒業研究報告要旨

松原 研究室	氏　　名	角　掛　正　弥
卒業研究題目	学術論文で引用された Web 上の研究データの同定と分類	

オープンサイエンスは、論文や研究データ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、統一的な規定がなされており、著者や題目の取得、種類の判別を機械的に行える。一方、研究データの引用については、統一的な規定がない。

近年、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。図 1 に研究データリポジトリのメタデータの例を示す。現在、研究データの多くが Web 上で利用可能であり、それらは URL で引用されている。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。

本論文では、メタデータにおける研究データの「種類」の自動獲得を目指し、学術論文で引用された Web 上の研究データの同定と分類について述べる。本研究では、URL を以下の 3 つに分類する分類タスクとして実現した。

- tool: コード、ソフトウェア、ツールキットなど　例 `https://pytorch.org/`
- 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`

論文における URL の引用目的がわかれば、上記のいずれかに適切に分類できる。このため本手法では、URL の引用文脈を分散表現として獲得し、分類の入力素性として用いる。引用文脈の例を図 2 に示す。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 を構成する文字列を考慮することが考えられる。例えば、“`http://trec.nist.gov/data/tweets/`” という URL は、“data” や “tweets” などの文字列から、ツイートに関するデータを参照していると推測できる。そこで、URL をドメイン名、ディレクトリ名などの構成要素に分解し、それらの分散表現を獲得し、利用する。

実験では、自然言語処理分野の国際会議である ACL の予稿集 2010〜2019 年の論文 (3,837 件) から URL(異なり数 9,480 件) を抽出し、使用した。頻度上位の URL500 件に対してラベル付けを行い、分類器の学習・テストに用いた。分散表現は word2vec により獲得した。ロジスティック回帰による 10 分割交差検定の分類結果を、適合率、再現率、F1 値により評価し、本分類タスクにおける本手法の有効性を確認した。

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

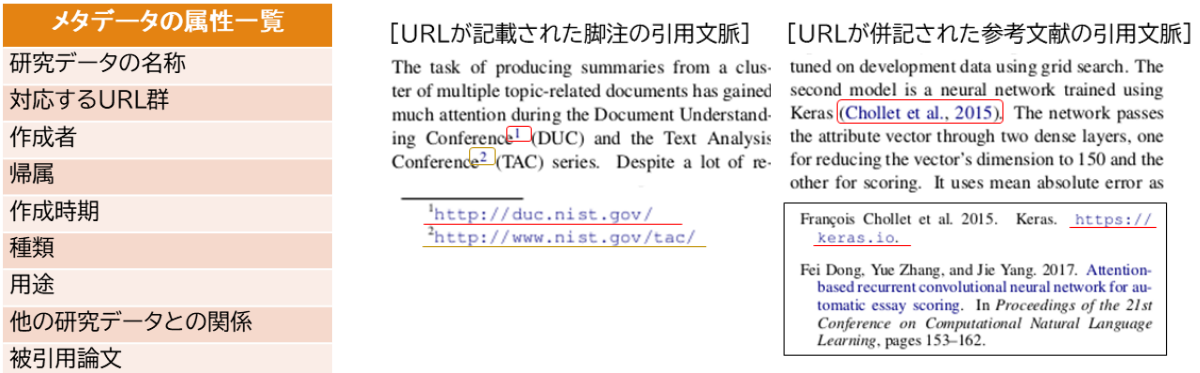

図 2　URL の引用文脈