研究データ検索における論文上の引用文脈の利用

角掛正弥
名古屋大学大学院情報学研究科
tsunokake.masaya@a.mbox.nagoya-u.ac.jp

松原茂樹
名古屋大学情報連携推進本部
matubara@nagoya-u.jp

1 まえがき

オープンサイエンスは，論文や研究データなど研究成果の共有や利活用を促進する活動である．近年，データ中心科学の広まりと共に，研究データの公表や論文での引用が増えており，研究データを共有する取り組みが進んでいる．例えば，特定の分野を対象とした研究データリポジトリの構築 [1]，あるいは，The Australian National Data Service (ANDS)[1)]，European Open Science Cloud (EOSC)[2)]，The National Data Service (NDA)[3)] など国家的単位での研究データ管理基盤の構築も進んでいる．

研究データへのアクセス性向上において重要な役割を担うのが検索システムである．研究データの検索は，一般に，研究データの作成者が登録する[4)]メタデータに基づき行われる．しかし，作成者によるメタデータだけでは，利用者が発行するクエリに対応するには十分でなく，提示すべき研究データを検索結果として提示できない可能性がある．その理由は，メタデータの記述が不十分であること [2, 3] や，作成者が想定していない研究データの特徴や用途が存在する可能性があるためである．

本稿では，「研究データの検索において論文上の引用文脈を利用することは有用である」という仮説を設定し，これを検証する．論文における研究データの引用文脈を収集し，それを研究データのメタデータに追加することで，既存のメタデータとは異なる言い回しや表現，作成者が想定していない特徴や用途などを獲得できる可能性がある．これにより，研究データに関する記述が拡充され，研究データへのアクセス性向上が期待できる．

本研究では，まず，引用文脈をメタデータに加えることで，研究データに関する記述を拡充できるかを調査した．具体的には，メタデータとして与えられている研究データに関する説明文書と，論文における引用文脈を比較し，その重複を調査した．次に，メタデータへの引用文脈の追加前には検索結果として提示されなかった研究データが，追加により提示できるようになるかの検証実験を行った．

2 研究データの検索と引用文脈

データセットの検索手法について研究がされている [3]．実際の検索サービスとして，Google Dataset Search[5)] [2] が存在する．同様に，研究データを対象とした検索サービス DataCite Search[6)]が存在する．研究データとは，研究の過程で収集・生成されたデジタル情報であり，プログラムやソフトウェア等のツールや，計測・試験データが一定の形式で整理されたデータセット等を含む[7)]．Zenodo[8)]や Mendeley Data[9)]においても研究データを検索できる．

研究データの検索は，各研究データに付与されたメタデータに対するキーワード検索で実現される．メタデータとは，名称，作成者，出版年，種類など，その研究データに関する各種情報を記述するデータである．例として，言語資源メタデータデータベース SHACHI[10)] [5] における，WordNet[6] のメタデータの一部を表 1 に示す．このメタデータは，DublinCore[11)]に準拠したメタデータ語彙で記述されている．メタデータ語彙には，DCAT[7]，schema.org[8]，DataCite Metadata Schema[9] などが存在し，これらに従いメタデータを記述することで一貫性を保持できる．メタデータには，データセットに関する説明を記す項目（以下，**内容記述**（Description））がある．

1） https://www.ands.org.au/
2） https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
3） http://www.nationaldataservice.org/
4） リポジトリの管理者が作成することもある．
5） https://datasetsearch.research.google.com/
6） https://search.datacite.org/
7） 一方で，包括的な定義を与えるのは難しく [4]，その定義は文脈によって異なる側面もある．
8） https://zenodo.org/
9） https://data.mendeley.com/
10） http://shachi.org/
11） http://dublincore.org/

表 1　SHACHI における WordNet のメタデータ（一部）

property	value
title	WordNet
creator	George A. Miller
subject	a large lexical database of English
type	[Text]
type .purpose	[lexicography] [developing_technologies]
identifier	http://wordnet.princeton.edu/
description （内容記述）	WordNet® is a large lexical database of English, developed under the direction ... tool for computational linguistics and natural language processing.

一般的に，研究データの作成者がそのメタデータを登録する．しかし，作成者によるメタデータだけでは，提示すべき研究データを検索結果として提示できない可能性がある．なぜなら，メタデータの記述が不十分な場合や，作成者が想定していない研究データの特徴や用途が存在し得るからである．

メタデータにおける内容記述を充実させることで，マッチするクエリが増加する．内容記述を拡充する方策として，論文における引用文脈を用いることが考えられる．論文で研究データを引用する際，著者はその説明や用途等を引用文脈に記述する．引用文脈を利用することで，研究データに関する記述を機械的に収集でき，異なる言い回しや表現の獲得による量的な充実が期待できる．

加えて，引用文脈を記した著者は作成者とは異なるため，作成者が想定していない研究データの特徴や性質，関連トピック，用途などを拡充できる可能性がある．特に用途に関しては，利用者が新たに作り出す可能性もある．実際，論文から用途情報を自動獲得する研究 [10] も行われている．引用文脈を収集することで，質的な充実も期待できる．

Singhal ら [11] は，利用者自身の研究の文脈に沿ったキーワードをクエリとする研究データ検索システムを提案している．検索対象となるメタデータとして，引用された論文のタイトルやタグ情報を用いている．しかし，これらは研究データについて直接記した情報ではなく，研究データの特徴や用途などとは関係しない可能性が高い．本研究では，引用文脈をメタデータとして直接利用することを想定する．

本稿では，「研究データの検索において論文上の引用文脈を利用することは有用である」という仮説を設定し，実験的にこれを検証する．

3　内容記述と引用文脈の関係

本節では，メタデータへの引用文脈の追加により，研究データに関する記述を拡充できるか否かについて，調査した結果を述べる．

3.1　調査方針

メタデータに引用文脈を加えたとしても，その内容が既存のメタデータの内容記述に対し重なりが大きいようなら，量的に拡充できているとは言えない．本研究では，キーワード検索を想定し，同一の研究データに対する内容記述と引用文脈の語彙を比較することで，その重複度を調査する．

3.2　調査用データの作成

調査にあたり，同一の研究データに対する内容記述と引用文脈を対応付けたデータセットを作成した．

引用文脈の抽出対象論文として，2000 年〜2019 年の ACL，NAACL，EMNLP における本会議論文を ACL Anthology[12] から収集し，PDFNLT[13] [12] によりテキスト化した．収集した論文数は 11,855 件であった．明示的な研究データの引用として URL に着目し [13, 14]，URL が記載された段落を抽出した．抽出された段落は 31,637 件であった．なお，脚注や参考文献に記載された URL については，その本文中での引用箇所を特定したうえで段落を抽出している．

研究データの内容記述の収集には，LDC[14] と SHACHI[5] を利用した．LDC の利用では，抽出した段落に記載されている URL のうち，LDC のカタログページを示す URL について，そのリンク先から内容記述を取得した[15]．また，SHACHI では言語資源のメタデータとして，内容記述と所在を示す URL 等が与えられている．これらを収集することにより，論文から抽出した段落文と SHACHI の内容記述を URL で対応付けた．

対応付けられた研究データの内容記述と引用文脈の文書対のうち，同一の研究データを指していない

12）https://www.aclweb.org/anthology/
13）https://github.com/KMCS-NII/PDFNLT-1.0
14）https://www.ldc.upenn.edu/
15）LDC の各言語資源ページには，<div itemprop="description", data-hook="description"> タグ下に内容記述が記載されているので，これを取得した．

表 2　内容記述と引用文脈の重複度

	語彙サイズ		語彙の重複度		
	引用文脈	内容記述	Jaccard	Dice	Simpson
平均	40.0	78.1	0.07	0.13	0.24
最小値	3	6	0	0	0
中央値	38	47	0.06	0.12	0.21
最大値	153	295	0.25	0.40	0.76
標準偏差	19.1	62.0	0.04	0.07	0.13

もの[16]や引用文脈が正しく抽出されなかったものを除外し，データセットを作成した．文書対の総数は 395 件，研究データの種類数は 98 件となった．

3.3　内容記述と引用文脈の重複度

3.2 節のデータセットを用い，内容記述と引用文脈の語彙の重複度を算出する．語彙は，以下の手順で前処理を行った後，重複を除去し作成する．

1. 文書のトークン化
2. 重要でない単語[17]を削除
3. 見出し語化，及び小文字化

トークン・見出し語化，ストップワードの判定には spaCy[18]を用いた．手順 2 で，検索においてクエリとして用いられる可能性が低い単語を取り除く．

語彙の重複度は，Jaccard 係数，Dice 係数，Simpson 係数により算出した．表 2 に算出結果を示す．語彙サイズが大きく異なるものの，いずれの係数も平均値や中央値の値が非常に低いことがわかる．

しかし，全く関係のない文書と内容記述の語彙を対象とした場合も，その重複度は低くなる．そこで，内容記述と，それとは内容的に関連が低い文書との対を作成し，重複度を算出する．その分布を，内容記述と引用文脈の文書対における分布と比較し，内容的な関連性は保ちながらも，引用文脈による拡充が可能であるかを検査する．一方で，関連が高すぎては拡充の効果が見られないため，内容的に強く関連する文書対も作成し，その重複度の分布とも比較する．異なる研究データ同士の内容記述は，関連が低いと考えられる．そこで，3.2 節で収集した全ての内容記述から文書対を作成し，これを関連が低い文書対とした．得られた文書対は 4,560 件であった．一方，LDC では研究データの内容記述以外にその README ファイルも配布されている場合がある．これらはいずれも作成者による同じ研究データに関する文書なので，関連が高いと考えられる．そこで，関連の高い文書対として，研究データの内容記述と README の文書対を作成した．得られた文書対は 40 件であった．

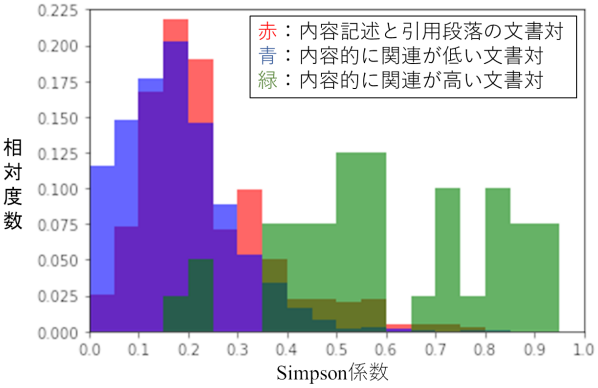

図 1　各文書対における Simpson 係数の分布（引用段落単位）

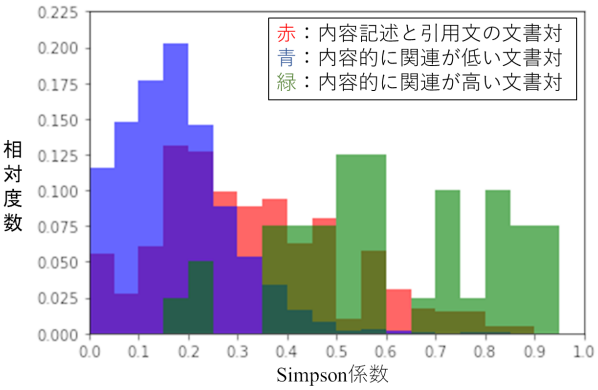

図 2　各文書対における Simpson 係数の分布（引用文単位）

図 1 に結果を示す．比較する語彙同士のサイズが大きく異なることが多いため，語彙サイズに頑健な Simpson 係数のみを表示している．関連が低い文書対は Simpson 係数の分布が低い位置に分布し，関連が高い文書対は Simpson 係数の分布が高い位置に分布している．内容記述と引用文脈の文書対では，その Simpson 係数の分布が上記 2 つの中間に位置している．内容記述と引用文脈には関連が高い文書対ほどの重複が存在せず，引用文脈をメタデータに加えることで記述を拡充できると考えられる．一方で，関連の低い文書対と比べると，高い位置に Simpson 係数が分布しており，内容記述と引用文脈との間には一定程度の関連が認められる．引用文脈をメタ

16）　同じ URL で研究データ以外や別の研究データを指している場合が存在する．
17）　1 文字のトークン，ストップワード，句読点のみで構成されるトークン，数値表現
18）　https://spacy.io/

表 3　クエリにマッチした引用文脈の例

クエリ	bilingual lexicon induction	document classification	coreference resolution	parallel data
研究データ名	Chinese-English Translation Lexicon Version 3.0	20 Newsgroups	OntoNotes Release 4.0	European Parliament Proceedings Parallel Corpus
引用文脈	In the task of **bilingual lexicon induction**, we opt for Chinese-English Translation Lexicon Version 3.0 to be the gold standard.	To empirically verify our method, from 20 Newsgroups, a dataset for **document classification** or clustering, we chose 6 classes and randomly drew 100 documents for each class.	When testing on the UMIREC and N2 corpora with the state of the art Berkeley **coreference resolution** system trained on OntoNotes, our inference substantially outperforms the original inference on the CoNLL 2011 metric.	The setup for our experimental comparison is German-to-English translation on the Europarl **parallel data** set.

データに加えることで，効果のある拡充が行えることを示している．

ここまで述べてきた引用文脈は，抽出した引用箇所を含む段落（以下，**引用段落**）をそのまま利用した結果である．引用段落は数文に渡る引用文脈を網羅することが可能な一方で，引用文脈でない文を含む可能性もある．そこで，引用箇所に対応する 1 文のみ（以下，**引用文**）を用いて同様の調査を行った．図 2 に，Simpson 係数の算出に引用文を用いた場合の分布を示す．図 1 と比較して，内容記述と引用文脈の文書対分布は右へシフトしており，引用文脈でない文を除外できていることがわかる．一方で，関連が高い文書対に比べると分布は左に位置している．内容記述と引用文の文書対のほとんどが 0.5 以下であり，記述を拡充する効果が期待できる．

4　研究データ検索における有用性

引用文脈の追加による研究データに関する記述の拡充が，検索において有用であることを検証する．検索対象が，１）内容記述のみ，２）内容記述と引用文脈，の 2 条件で研究データ検索実験を行い，発見できる研究データ数を比較する．

4.1　検索実験の設定

利用者が研究データを検索する際のキーワードとして研究トピックや用途が考えられる．そこで，研究データを使用した過去の事例から用途を収集し，検索実験のクエリとする．収集には LREMap[19] を用いた．LREMap は，研究で使用・作成された言語資源のメタデータが登録されているリポジトリである．論文を投稿する際に，著者が使用・作成した言語資源をこのリポジトリに登録する．メタデータには，言語資源の名称，種類，用途などの様々な情報が記述される．用途情報は，LREMap 側で定義されたものから選択する，もしくは，著者自身が自由に記述することで作成される．LREMap に登録されている言語資源の使用事例を収集することにより，多様な用途情報を獲得できる．

19）https://lremap.elra.info/

LREMap から収集した全用途情報のうち，「,」を含む場合そこで分割し，クエリセットを作成した．作成したクエリの種類数は 728 件である．なお，本実験では，メタデータに加える引用文脈をできる限り研究データに関する記述に絞るため，引用文のみを引用文脈とした．

4.2　実験結果

3.2 節で作成した引用文と内容記述の文書対データを検索対象として実験を行った．

引用文脈を加えることにより，研究データ発見数が増えたクエリが 36 件存在した．引用文脈を用いることにより，内容記述のみでは発見できなかった研究データにマッチできた事例を表 3 に示す．クエリに対して，拡充した記述が有効に働いていることがわかる．

適合性も考慮した実験結果を示すため，クエリとマッチした引用文の全ペアについて適当であるか否かを確認し，適当でないマッチは行わないように再度検索実験を行った．その結果，引用文脈を加えることにより研究データ発見数が増えたクエリは 29 件であった．その一覧を付録の表 4 に示す．引用文脈の追加により，研究データの発見数が増加することを確認した．なお，クエリとマッチした引用文のペアのうち，適当であったものの割合は 0.788(93/118) であった．

5　あとがき

本稿では，「研究データの検索において論文上の引用文脈を利用することは有用である」という仮説を設定し，実験による検証結果について述べた．まず，既存のメタデータにおける研究データに関する説明と論文の引用文脈について，その重複度を調査した．調査結果から，引用文脈を利用することで，効果のある拡充を行えることがわかった．次に，LREMap から収集した用途情報を用い検索実験を行った．引用文脈を用いた拡充により，発見できる研究データ数が増加し，研究データの検索における引用文脈の有用性を確認した．

参考文献

[1]大向一輝. オープンサイエンスと研究データ共有. 心理学評論, Vol. 61, No. 1, pp. 13–21, 2018.

[2]Dan Brickley, Matthew Burgess, and Natalya Fridman Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *Proceedings of The World Wide Web Conference*, pp. 1365—1375, 2019.

[3]Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. Dataset search: a survey. *The VLDB Journal*, Vol. 29, No. 1, pp. 251–272, 2020.

[4]The Australian National Data Service. What is research data, 2017（2021-01 閲覧）. https://www.ands.org.au/__data/assets/pdf_file/0006/731823/Whatis-research-data.pdf.

[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 in 22nd International Conference on Computational Linguistics: Companion volume: Posters*, pp. 119—122, 2008.

[6]George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, Vol. 38, No. 11, pp. 39—41, 1995.

[7]Riccardo Albertoni, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. Data catalog vocabulary (dcat) - version 2. In *W3C Recommendation*, 2020（2021-01 閲覧）. https://www.w3.org/TR/vocab-dcat-2/.

[8]Ramanathan V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of structured data on the web. *Communications of the ACM*, Vol. 59, No. 2, pp. 44—51, 2016.

[9]DataCite Metadata Working Group. Datacite metadata schema documentation for the publication and citation of research data (version 4.3), 2019（2021-01 閲覧）. https://doi.org/10.14454/7xq3-zf69.

[10]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 A, Vol. 95-A, No. 7, pp. 611–622, 2012.

[11]Ayush Singhal, Ravindra Kasturi, and Jaideep Srivastava. Datagopher: Context-based search for research datasets. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, pp. 749—756, 2014.

[12]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.

[13]角掛正弥, 松原茂樹. 論文で引用された研究データの同定と分類. 電気学会論文誌C（電子・情報・システム部門誌）, Vol. 140, No. 12, pp. 1357–1364, 2020.

[14]Hidetsugu Nanba. Construction of an academic resource repository. In *Proceedings of Toward Effective Support for Academic Information Search Workshop at ICADL-2018*, pp. 8–14, 2018.

A 付録

表 4　研究データ発見数が増えたクエリ一覧

クエリ	内容記述のみ	引用文脈と内容記述
semantic similarity	1	2
tagging	7	10
coreference	6	8
pos tagging	2	3
testing	3	6
bilingual lexicon induction	0	1
wsd	0	1
machine translation	8	9
word embeddings	0	3
wordnets	1	2
general	8	10
diverse	0	1
analysis	2	3
annotation	22	23
word embedding	0	1
term extraction	0	1
coreference resolution	1	4
sentiment	0	1
parallel data	0	1
feature extraction	0	1
document classification	0	1
semantics	5	6
standard	15	24
parsing	7	9
nlp	0	2
named entity recognition	0	1
translation	11	12
evaluation	6	11
terminology	0	1