データセット間の関連性推定におけるメタデータの利用

伊藤滉一朗 [1]　松原茂樹 [1,2]
[1] 名古屋大学大学院情報学研究科　[2] 名古屋大学情報基盤センター
{ito.koichiro.z5, matsubara.shigeki.z8}@f.mail.nagoya-u.ac.jp

概要

近年，オープンサイエンスが世界規模で推進され，データセットやコードなどの研究成果の公開促進に加えて，そのアクセス性が重要視されている．アクセス性を高めるための要素の 1 つとして，研究成果間で関連付けることが挙げられる．そこで本論文では，研究成果の一種であるデータセットを対象に，それらの間の関連性を推定することの実現性を検証する．データセットに付与されたメタデータを用いて BERT ベースのモデルによって関連性を推定する手法を実装し，実験により推定性能を評価した．

1　はじめに

近年，オープンサイエンスが世界規模で推進され [1, 2]，研究成果の公開が推奨されている．研究成果としては，論文のほか，研究の過程で作成されたデータセットやコードなどが挙げられる．ただし，研究活動の加速化のためには，研究成果の公開だけでなく，そのアクセス性も重要となる．

研究成果へのアクセス性を高めるための要素として，研究成果間で適切に関連付けられていることが挙げられる．また，研究成果間の関連性には，いくつかのタイプを想定することができる．例えば，研究成果が産み出された背景が共通している場合や，研究成果の利用用途が共通している場合などがある．タイプごとに関連性を獲得できれば，図 1 に示すような，研究成果への的確なアクセスの促進に貢献できる．

そこで本論文では，研究成果の一種であるデータセットを対象に，それらの間の関連性を推定することの実現可能性を検証する．データセット間の関連性の推定をタイプごとに実施し，それぞれの推定可能性を検証する．関連性の推定手法を実装し，その性能を実験によって評価した．本手法では，データセットのメタデータを入力として，BERT [3] ベースのモデルによって，関連性の有無を推定する．実験の結果，本手法によって，一定の水準でその推定が可能であることを確認した．

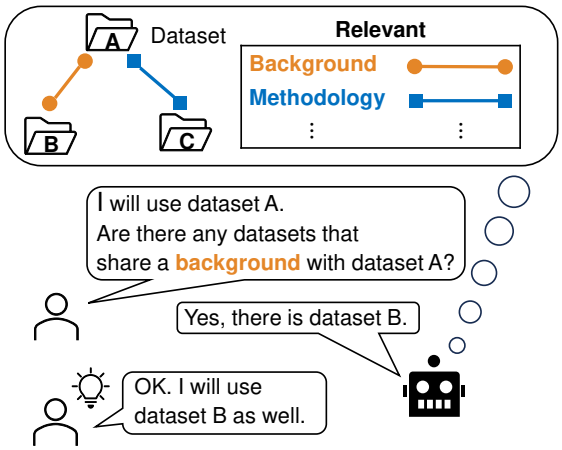

図 1　研究成果間の関連性とその利用

本論文の構成は以下の通りである．2 章では，研究成果間の関連性について述べる．3 章では，データセット間の関連性の推定手法について説明し，4 章では，その性能を実験によって評価する．最後に，5 章で本論文をまとめる．

2　研究成果間の関連性

本章では，研究成果間の関連性とその有用性について説明したのち，従来研究について述べる．

2.1　関連性の有用性

研究成果間の関連性は，ある研究成果の発見者に，関連する別の研究成果の情報を提供するために利用できる．これにより，研究成果の探索の効率化や，研究成果の見落としの防止が期待できる．また，所望の研究成果の特徴を適切に言語化することが難しい場合においても，関連する研究成果を順に辿ることで，所望の研究成果にアクセスできる可能性がある．

研究成果間の関連性には，いくつかのタイプを考えることができる．例えば，研究成果が産み出され

た背景が共通している場合や，研究成果の利用用途が共通している場合などがある．タイプごとに関連性を獲得できれば，研究成果への的確なアクセス促進に貢献できる．本論文では，研究成果の一種であるデータセットを対象に，それらの間の関連性を推定することの実現可能性をタイプごとに検証する．

2.2 従来研究

研究成果は，研究によって産み出されたデータセットやコードなど多岐にわたり，論文もその 1 つである．論文間の関連を示す手がかりとして，論文の引用関係が挙げられる．これまでに，論文の引用に着目した研究が多数行われており，論文の引用情報を保持したグラフ [4, 5, 6, 7] やコーパス [8, 9, 10] の構築が進められている．

また，論文の引用の意図に着目した研究も存在する．例えば，論文の引用は，研究対象の問題や概念の説明のため，あるいは，研究で利用した既存の手法やデータの説明のために行われる．このように，論文の引用は，その意図の観点から，いくつかのタイプに分けられる．引用タイプを自動分類できれば論文の解析や読解に貢献できるという考えのもと，これまでに，論文の引用タイプがアノテーションされたコーパスの構築や引用タイプの分類手法の提案が行われている [11, 12, 13]．

上述した通り，論文については，その関連に着目した多くの研究が存在する．一方，論文以外の研究成果については検討が十分とはいえない状況にある．近年，研究成果の一種であるデータセットについて，その関連性の推定が試みられているものの [14]，関連性のタイプを考慮するには至っていない．

3 手法

本章では，本研究で実装するデータセット間の関連性を推定する手法を説明する．図 2 に，その概略を示す．一般に，リポジトリに格納されているデータセットには，その概要を表すメタ情報がメタデータとして付加されている．メタデータは，名称などの項目（Field）に対して値（Value）を持つ．本手法では，データセット間の関連性の推定に，そのメタデータを利用する．

メタデータは表形式の構造化データであるが，項目と値で構成される単純な文字列として言語モデルに入力することも試みられている [14, 15, 16, 17]．これらの研究に倣い，本手法では，下記の形式のテキストを入力として，BERT [3] ベースのモデルによって推定を行う．[ ] で表される文字列は，特殊トークンを表す．

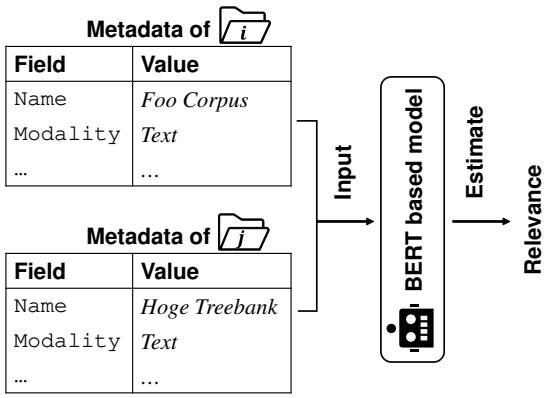

図 2 関連性の推定手法の概略

- [CLS] [FIEDL1] {*Value of* FIEDL1 *of dataset1*} [FIEDL2] {*Value of* FIEDL2 *of dataset1*} ··· [SEP] [FIEDL1] {*Value of* FIEDL1 *of dataset2*} [FIEDL2] {*Value of* FIEDL2 *of dataset2*} ··· [SEP]

最初に [CLS] トークン，次いで 1 つ目のデータセット（*dataset1*）の特徴を表すトークン系列と [SEP] トークン，最後に 2 つ目のデータセット（*dataset2*）の特徴を表すトークン系列と [SEP] トークンが続く．各データセットの特徴を表すトークン系列は，データセットのメタデータ項目名を表す特殊トークン（上記の [FIEDL1] や [FIEDL2] など）の後に，その項目に対する値が続く形式とする．推定モデルは，事前学習済みの BERT ベースのモデルに，2 クラス分類用の出力層を追加し，fine-tuning によって構築する．

本手法で利用する研究成果のメタデータについては，従来は人手で収集および整備されていたが，近年ではその自動生成が試みられている．論文には研究成果の作成や利用に関する記述が含まれることがある点に着目し，論文から研究成果のメタデータを抽出する研究が存在する [18, 19, 20, 21, 22]．また，既存のメタデータから，未入力のメタデータ項目を自動補完する研究も存在する [15, 17]．これらの研究の発展に伴い，メタデータを入力とする本手法を適用可能な場面の増加を期待できる．また，本手法はメタデータのみを利用するという点で汎用性が高く，メタデータを集約しているリポジトリやカタログとの親和性が高いことも特徴の 1 つである．

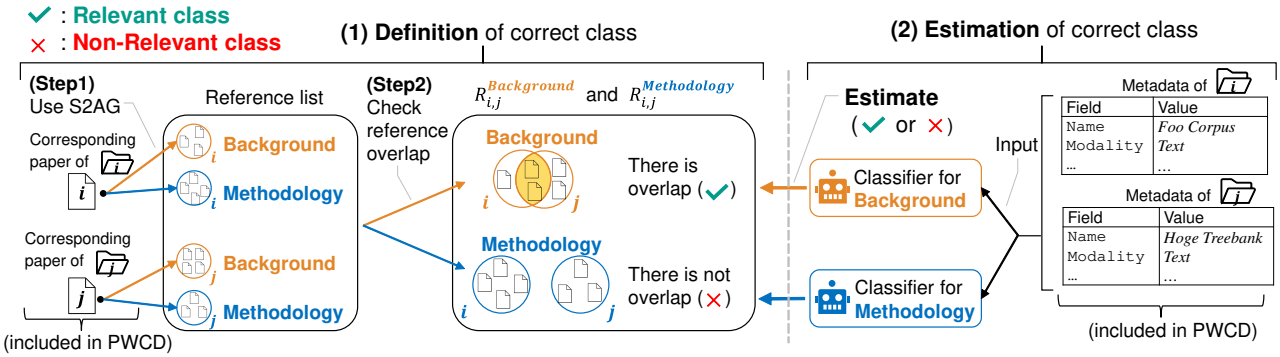

図 3　実験データの作成手順（左）と本手法による推定（右）

表 1　推定に利用するメタデータの例

Field	Value
Name	MultiNLI
Full Name	Multi-Genre Natural Language Inference
Modalities	Texts
Tasks	Natural Language Inference
Description	The **Multi-Genre Natural Language Inference** (**MultiNLI**) dataset has 433K sentence pairs. Its size and mode of collection are ...

表 2　推定対象の引用タイプ [12]

タイプ	定義
Background	The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field.
Methodology	Making use of a method, tool, approach or dataset

4　実験

データセット間の関連性を推定することの実現可能性を検証するために，その推定実験を行った．

4.1　実験データ

推定手法の性能を評価するためには，データセットのペアとそれらの間の関連性の正解ラベルが必要となる[1]．すなわち，1 つのエントリが，$(D_i, D_j, R^X_{i,j})$ のトリプルで構成されるデータが必要となる．ただし，D_i と D_j はデータセットを，$R^X_{i,j}$ は D_i と D_j の間のタイプ X における関連性を表す．

4.1.1　実験データの作成手順

実験データの作成手順の概略を図 3 の左部に示す．本実験では，Papers With Code が公開している Datasets のデータ [23]（以降，**PWCD** と表記）を用いた[2]．PWCD には，データセットのメタデータが記録されている．表 1 に，その例を示す．また，PWCD には，各データセットと対応する論文の情報も記録されている．例えば，表 1 で例に挙げた自然言語推論用のデータセットである MultiNLI [24] であれば，A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference という，MultiNLI の作成を報じている論文のタイトルが記録されている．

本実験では，論文間の関連度を測る手法である書誌結合（Bibliograpic coupling）[25] の考え方に基づき，データセット間に関連性があるか否かを定めた．すなわち，データセットに対応する論文間の参考文献リストの重複の有無によって，データセットのペアが関連性を持つか否かを定めた．以降では，関連性を持つペアが属するクラスを **Relevant**，そうでないペアが属するクラスを **Non-Relevant** と表記する．

各ペアが属するクラスを定めるために必要な参考文献リストは，Semantic Scholar Academic Graph (S2AG) [7] の API で取得した．引用タイプを判定する S2AG の機能により，引用タイプ別に参考文献リストを作成した（図 3: step1）．S2AG の引用タイプのセットは，文献 [12] の定義を踏襲している．そのうち，本実験では，表 2 に示す Background と Methodology の 2 つのタイプについて[3]，参考文献リストの重複に基づいて正解のクラスを定義し（図 3: step2），推定手法の性能の評価に用いた．

4.1.2　実験データの規模

本実験では，PWCD 内のデータセットのうち，S2AG で参考文献のリストを取得できた 7,972 個のデータセットを用いた．データセットの集合を

1）　正確には，本実験の実施のために必要なものは，データセットそのものではなく，そのメタデータである．
2）　2024/09/02（JST）にダウンロードしたデータを用いた．
3）　Result タイプも存在するが，Background と Methodology に比べて，その出現が極少数であったため本実験の対象外とした．その関連性の推定可能性の検証は今後行う予定である．

表 3　実験データの規模

		Pair (Background)		Pair (Methodlogy)	
	Dataset	Relevant	Non-Relevant	Relevant	Non-Relevant
Train	6,377	70,563	1,961,655	170,228	1,861,990
Dev.	797	1,068	30,171	2,652	28,587
Test	798	12,140	305,863	30,271	287,732
Total	7,972	83,771	2,297,689	203,151	2,178,309
		(3.52%)	(96.48%)	(8.53%)	(91.47%)

8:1:1 の割合で，学習，開発，テスト用にランダムに分割した．分割ごとに，任意の 2 つのデータセットを取り出してペアとし，正解のクラスを定めた．ただし，実験における計算量の制約から，学習および開発用のデータセットのペアには，全ペアの約 10% をランダムサンプリングして利用した．表 3 に，実験データの規模を示す．実験データにおける関連性を持つペアの割合は低く，Background タイプでは 3.52%, Methodology タイプでは 8.53%であった．

4.2　実装

推定モデルには SciBERT [26] [4] を，入力には表 1 で示した 5 つのメタデータ項目とその値を用いた．メタデータ項目に対応する特殊トークン（[NAME], [FULL NAME] など）を用意して，その後に値を続けた．入力形式は，3 章で説明した通りである．

Fine-tuning における損失関数は Cross entropy loss として，バッチサイズを 32, Gradient accumulation step 数を 8，エポック数を 3 とした．モデルの最適化手法には，Weight decay を 0.01 とした AdamW [27] を用いた．学習率は，最大値を 1e-5，Warmup ratio を 10%とした線形スケジューリングで調整した．

4.3　評価方法と比較手法

推定性能は正解率と F 値で評価する．F 値については，Relevant クラスに対する F 値（F_{Rel}），Non-Relevant クラスに対する F 値（F_{Non}），これらの平均値であるマクロ F 値（F_M）の 3 つの結果を報告する．また，データセットのメタデータに基づかない手法として，ランダムに推定する下記の 2 つの手法を，Background タイプと Methodology タイプごとに実装した．

- $Rand_{even}$: 50%の確率でランダムに推定する手法
- $Rand_{dist}$: 学習データにおけるクラス分布に従って，ランダムに推定する手法

4）https://huggingface.co/allenai/scibert_scivocab_uncased

表 4　実験結果

	Background				Methodlogy			
	Acc.	F_M	F_{Rel}	F_{Non}	Acc.	F_M	F_{Rel}	F_{Non}
$Rand_{even}$	0.500	0.365	0.072	0.658	0.500	0.402	0.161	0.644
$Rand_{dist}$	0.929	0.500	0.036	0.963	0.838	0.501	0.091	0.911
本手法	**0.957**	**0.585**	**0.192**	**0.978**	**0.884**	**0.594**	**0.250**	**0.937**

4.4　実験結果

推定性能を表 4 に示す．まず，総合的な評価指標である正解率とマクロ F 値に着目する．メタデータを利用する本手法では，Background タイプの正解率は 0.957 でマクロ F 値は 0.585，Methodology タイプの正解率は 0.884 でマクロ F 値は 0.594 であった．いずれのタイプにおいても，ランダムな手法の結果を上回り，有意差が確認された（マクネマー検定：$p < 0.05$）．これらのことから，本手法によって，データセット間の関連性を一定の水準で推定できること，ならびに，データセットのメタデータを利用することの有効性が示された．

次に，Relevant クラスと Non-Relevant クラスの F 値に着目する．いずれのクラスにおいても，Background タイプと Methodology タイプともに，本手法はランダムな手法を上回った．しかし，Relevant クラスの F 値は，Non-Relevant クラスよりも低かった．表 3 で示した通り，Relevant クラスのペアが相対的に少ないことが原因の 1 つであると考えられるものの，Relevant クラスのペアの検出性能の向上は今後の課題といえる．

5　おわりに

本論文では，研究成果の一種であるデータセットを対象に，それらの間の関連性を推定することの実現可能性をタイプごとに検証した．Background と Methodology の 2 つのタイプに対して関連性の推定手法を実装し，その推定が一定の水準で可能であることを実験によって確認した．今後は，データセット以外の研究成果についても，それらの間の関連性の推定可能性を調査したい．

謝辞

本研究は，一部，科学研究費補助金 基盤研究 (B) 23K21844 により実施したものである．本研究における実験は，一部，名古屋大学のスーパーコンピュータ「不老」の一般利用制度により実施した．

参考文献

[1] UNESCO. Open science. https://www.unesco.org/en/open-science (Last accessed 2025/01/07).

[2] Annex 1: G7 open science working group (OSWG). https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/annex1_os.pdf (Last accessed 2025/01/07).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of NAACL-2019**, pp. 4171–4186, 2019.

[4] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and mining of academic social networks. In **Proceedings of KDD-2008**, pp. 990–998, 2008.

[5] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In **Proceedings of NAACL-2018**, pp. 84–91, 2018.

[6] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. **Quantitative Science Studies**, Vol. 1, No. 1, pp. 396–413, 2020.

[7] Alex D. Wade. The semantic scholar academic graph (S2AG). In **WWW'22: Companion Proceedings of the Web Conference 2022**, p. 739, 2022.

[8] Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. The ACL Anthology network corpus. In **Proceedings of NLPIR4DL-2009**, pp. 54–61, 2009.

[9] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In **Proceedings of ACL-2020**, pp. 4969–4983, 2020.

[10] Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. The ACL OCL corpus: Advancing open science in computational linguistics. In **Proceedings of EMNLP-2023**, pp. 10348–10361, 2023.

[11] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. **TACL**, Vol. 6, pp. 391–406, 2018.

[12] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In **Proceedings of NAACL-2019**, pp. 3586–3596, 2019.

[13] Suchetha N. Kunnath, David Pride, and Petr Knoth. Dynamic context extraction for citation classification. In **Proceedings of AACL-2022**, pp. 539–549, 2022.

[14] Koichiro Ito and Shigeki Matsubara. Estimation of relevance between datasets for enhancing accessibility of research artifacts. In **Proceedings of ICADL-2024, Part 2**, pp. 97–103, 2024.

[15] Hyunju Song, Steven Bethard, and Andrea Thomer. Metadata enhancement using large language models. In **Proceedings of SDP-2024**, pp. 145–154, 2024.

[16] Pawin Taechoyotin and Daniel Acuna. MISTI: Metadata-informed scientific text and image representation through contrastive learning. In **Proceedings of SDP-2024**, pp. 155–164, 2024.

[17] Koichiro Ito and Shigeki Matsubara. Estimating metadata of research artifacts to enhance their findability. In **Proceedings of e-Science-2024, Article 33**, pp. 1–2, 2024.

[18] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In **Proceedings of EMNLP-2018**, pp. 3219–3232, 2018.

[19] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In **Proceedings of ACL-2020**, pp. 7506–7516, 2020.

[20] Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In **Proceedings of EACL-2021**, pp. 707–714, 2021.

[21] David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. SoMeSci – A 5 star open data gold standard knowledge graph of software mentions in scientific articles. In **Proceedings of CIKM-2021**, pp. 4574–4583, 2021.

[22] Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. Capabilities and challenges of LLMs in metadata extraction from scholarly papers. In **Proceedings of ICADL-2024, Part 1**, pp. 280–287, 2024.

[23] Papers with code datasets. https://github.com/paperswithcode/paperswithcode-data (Last accessed 2024/09/02).

[24] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In **Proceedings of NAACL-2018**, pp. 1112–1122, 2018.

[25] Maxwell Mirton Kessler. Bibliographic coupling between scientific papers. **American Documentation**, Vol. 14, No. 1, pp. 10–25, 1963.

[26] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of EMNLP-IJCNLP-2019**, pp. 3615–3620, 2019.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **Proceedings of ICLR-2019**, pp. 1–8, 2019.