テキスト情報の表現を利用した文献グラフの表現学習

片桐 脩那　井田 龍希　三輪 誠　佐々木 裕
豊田工業大学
{sd19027,sd22401,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

本研究では，論文の文献情報を表す文献グラフを対象に，高品質な文献グラフの表現の構築に向けて，大規模事前学習モデルから得られるテキストの表現とグラフ埋め込み手法・グラフニューラルネットワークを用いたリンク予測手法の利用可能性を調査する．文献グラフでは題目や要旨などのテキスト情報や著者・掲載雑誌などの情報など様々な情報が引用関係や共通する著者などで繋がっている．実験では，ACL Anthology に含まれる文献情報を対象に，リンク予測タスクを用いて評価を行い，大規模事前学習モデルから得られる表現のグラフ埋め込み手法への有効性を示した．

1　はじめに

膨大な論文の中から，探したい・関連のある論文を見つけ出すには，検索クエリの工夫や著者や引用関係などの様々な視点からの文献調査が必要である．従来のキーワード型の検索システムでは，クエリに一致する論文しか検索できないため，検索のみで見つけられる論文は限定的である．特に，近年は出版される論文が日々爆発的に増大しており，大量の論文から関連する情報に辿り着くための手段への需要が増大している．

従来の機械学習では，テキストやメディア情報，人物や属性などの様々な種類の情報を表現し，統一的に扱うことは難しかったが，深層表現学習により統一的に扱うことが可能となってきている．さらに，BERT (Bidirectional Encoder Representations from Transformers) [1] に代表される自己教師あり学習により，様々な種類の情報における事前学習モデルが提案されており，高品質な表現を学習の初期から利用することができる．このような深層表現学習・自己教師あり学習を背景に，ヘテロな情報をノード，それらの関係をエッジとしたグラフを構築し，グラフ上での自己教師あり学習により，グラフ上でのノード間の関係を考慮しながらグラフ情報の埋め込み表現を獲得するヘテロなグラフ埋め込み手法の研究が盛んに行われている [2].

論文の文献情報の中の題目や要旨をノード，引用関係などをエッジとした文献グラフにグラフ埋め込み手法を適用し，グラフ上のあるノードからそれに関連するノードを予測するリンク予測を用いることで，様々なヘテロな情報を対象とした柔軟な検索システムを構築する研究が行われている [3, 4]．この研究では，大規模事前学習を用いたモデルは利用できておらず，このヘテロな情報に対する有用性は評価できていない．また，グラフのノード間の表現においては，グラフ埋め込みだけではなく，グラフニューラルネットワーク (Graph Neural Networks; GNNs) [5] を用いたリンク予測手法 [6, 7] など，より深い構造を利用した表現に関する研究も行われているが，このような手法も評価されていない．その他にも引用関係や共著者などを対象としたグラフ表現学習の研究も盛んに行われている [2] が，文献情報の限られた要素を対象としているものがほとんどである．

このようなことから，本研究では，文献情報に含まれる様々な要素とその関係をグラフ構造として表現し，グラフの埋め込みを獲得する際の，大規模事前学習モデルによるテキストの表現とグラフ埋め込み手法・GNN を用いたリンク予測手法の利用可能性について調査を行う．具体的には，文献情報である題目や要旨などのテキスト情報に事前学習モデルによる表現を利用し，リンク予測における性能向上を図る．また，グラフ埋め込み手法，GNN を用いたリンク予測手法それぞれを用いて，リンク予測での性能の評価を行い，それぞれの手法の利点・問題点について調査を行う．本研究の貢献は次の通りである．

- グラフ埋め込み手法，GNN を用いたリンク予測手法の文献グラフでのリンク予測性能の評価.

- 文献グラフにおけるテキストノードの大規模事前学習モデルによる初期化によるリンク予測性能の改善.

2 関連研究

2.1 文献情報のベクトル表現による検索システム

論文の文献情報中の様々な要素を考慮して，埋め込みを行い，検索に利用する研究が行われている [3, 4]．これらの研究では，著者・題目・要旨・出版年・引用情報を対象に，LINE (Large-scale Information Network Embedding) [8] 及び TransE [9] に基づいて文献情報のベクトル表現を獲得し，関連する論文要素の検索を行う手法を提案している．この手法では，テキスト情報である題目・要旨は独立した句の集合として表現している．

2.2 グラフ埋め込み

グラフ埋め込みでは，グラフ構造を始点ノード h, 終点ノード t とそれらのノード間のエッジの関係 r を用いたトリプル (h, r, t) の集合で表すのが一般的であり，それらの関係を表現できるように h, r, t の埋め込み表現とその関係を表すパラメタを学習する．グラフ埋め込みの代表的な手法に TransE [9] がある．TransE は，トリプルの要素 h, r, t の埋め込み表現のみで，トリプルの関係を表現する手法である．式 (1) のように，始点ノード h のベクトルを関係 r のベクトルで平行移動させた表現ベクトル $\boldsymbol{h}+\boldsymbol{r}$ と終点ノードのベクトル $\boldsymbol{t}$ が一致するように h, r, t の埋め込み表現を更新する．

$$\boldsymbol{h}+\boldsymbol{r} \approx \boldsymbol{t} \tag{1}$$

この手法により，ノードや関係の埋め込み表現を獲得できるとともに，あるノードにある関係で関連した別のノードを予測するリンク予測が可能となる．

2.3 GNN によるリンク予測

グラフ構造を表現する GNN（[5, 10] など）を利用したリンク予測の研究が盛んに行われている [6, 7]．ここでは，本研究で採用するリンク予測タスクにおいて高い性能を挙げている Graph Inductive Learning (GraIL) [7] について説明する．GraIL はトリプルの 2 つのノードとその周辺を含むサブグラフからトリプルのスコアを計算する．GraIL はサブグラフの抽出，ノードの対象ペアに対する相対位置のラベル付け，GNN によるスコア計算の 3 ステップでリンク予測を行う．まず，トリプルの始点・終点の両方のノードと一定以下のホップ数で繋がっているノードのみを含むサブグラフを抽出する．次に，ノードのラベル付けではノードに（始点ノードからの距離，終点ノードからの距離）というラベルをつける．始点・終点ノードのラベルは (0,1) と (1,0) というラベルになる．最後に，ラベル付けをしたサブグラフについて，ノードをラベル表現で初期化し，GNN によるノード表現を学習し，スコア計算をする．サブグラフの全てのノードに対応する GNN のノード表現を平均したサブグラフ全体の表現 $\boldsymbol{h}_{G_{(h,t,r)}}$，始点・終点ノードの GNN の表現 $\boldsymbol{h}_h, \boldsymbol{h}_t$，関係の埋め込み表現 $\boldsymbol{e}_r$ を結合した表現に重みをかけて，式 (2) のように，トリプルのスコアを計算し，正解のトリプルが他のトリプルのスコアよりも高くなるように学習を行う．

$$score(h, r, t) = \boldsymbol{W}^{\boldsymbol{T}} [\boldsymbol{h}_{G_{(h,t,r)}} \oplus \boldsymbol{h}_h \oplus \boldsymbol{h}_t \oplus \boldsymbol{e}_r] \tag{2}$$

GraIL の GNN では，R-GCN [11] に着想を得て，エッジの種類ごとに異なるパラメタを利用し，エッジに対する注意機構を利用してノード情報を集約している．

3 文献グラフの表現学習

本研究では，高品質な文献グラフの表現の獲得に向けて，文献情報を文献グラフとしてトリプルを用いて表現し，SciBERT [12] を用いた大規模事前学習モデルによるテキスト情報の埋め込みの影響と，TransE と GraIL それぞれの表現学習手法の比較を行う．本研究の流れを図 1 に示す．

事前学習の影響については，TransE を対象に，初期表現にランダムな表現を用いた場合と題目や要旨などのテキストに大規模事前学習モデル SciBERT を用いた埋め込みを行った場合を比較する．TransE の初期表現において，題目や要旨などのテキストノードについて埋め込みを用いることで，ノードがテキスト情報と論文間の関係の両方を考慮した表現を獲得できると考えられる．

GraIL については GNN によるスコア計算の際に TransE で学習したノード表現をノードの初期表現としてラベル表現に結合して学習を行う．GraIL によるニューラルネットワークで表現を学習することで，周囲のノードの情報をより考慮した表現を学習できると期待される．さらに，ノード表現に TransE

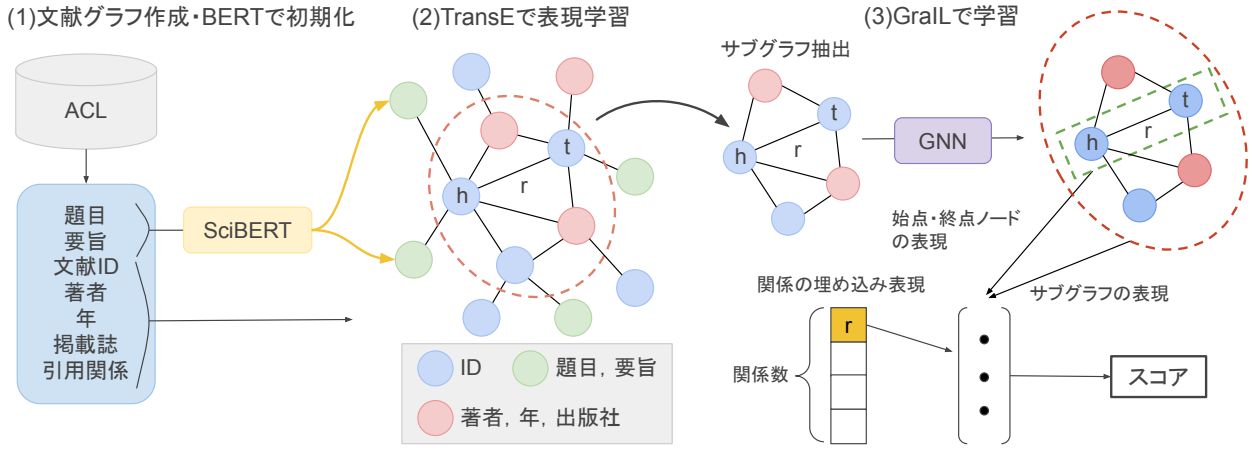

図 1　文献グラフの作成と表現学習の流れ

で学習したノード表現を用いることで，TransE のみによる表現学習に比べて，各ノードにおいてその周囲のノードの情報をより考慮した表現が獲得できると考えられる．

3.1　文献グラフの作成

まず，文献グラフを作成し，題目・要旨ノードを BERT で初期化する (図 1(1))．文献データの中で用いる文献情報には，題目，要旨，著者，掲載誌などがある．文献情報の中には，テキストカテゴリ（題目や要旨など）とそれ以外の非テキストカテゴリがある．テキストカテゴリに対して，初期表現を用いる際は，SciBERT [12] を用いたテキスト情報の埋め込みを行う．これにより，テキストカテゴリのノードがテキスト情報と論文間の関係の両方を考慮した表現を獲得できると期待される．

3.2　グラフ埋め込み

図 1 の (2) のように 3.1 節で作成したトリプルに対して TransE を用いてベクトル表現を学習する．TransE の学習により，各ノードが意味のある情報を持つようなノード表現を獲得できる．

3.3　GNN によるリンク予測

GraIL を用いた表現学習の手順は，図 1 の (3) の通りである．まず，サブグラフの抽出を行う．各トリプルの h と t を対象ノードとする．図 1 で，ノードに h と t と書かれているものが対象ノードである．サブグラフは対象ノードのいずれかから k ホップ以内にあり，かつ，対象ノード間のパス上にあるノードとその間のエッジで構成する．次に，サブグラフのノードのラベル付けを行う．最後に，TransE で学習したノード表現をラベル表現に追加して，GNN に入力し，2.3 節のスコア関数に基づいて GNN の学習を行う．

4　実験設定

データセットには ACL Anthology [13] に掲載されている 71,567 件の論文とその文献情報を使用した．引用情報は ACL Anthology に含まれないため，Semantic Scholar Academic Graph API [14] を用いて取得した．API には，時間による回数制限があるため，制限を超えないように時間を空けながらアクセスした．また，取得した文献情報は BibTeX 形式となっていたため，JSON 形式に変換を行った．

本研究では簡単のため，ワークショップの論文を除いた過去 5 年の論文とその文献情報 (16,916 件) を使用して文献グラフを作成した．データの統計を付録 A に示す．文献 ID・題目・要旨・著者・年・掲載雑誌のノードを持ち，文献 ID のノードと文献情報のノード間に辺を張るスター型の文献グラフを作成した．また，引用関係にある文献 ID のノード間にも辺を張った．

文献 ID のノードと著者の関係にあるノードを予測するリンク予測で評価した．5 回以上掲載されている著者とその著者が執筆した文献 ID のトリプルから開発用データ 10,000 件と評価用データ 5,000 件を用意した．また，開発用データと評価用データに存在する全てのノードが訓練データに存在するトランスダクティブな設定とした．

評価指標には，[7] で使用されている Hit@N を用いた．評価データの各トリプルに対して終点ノード

をランダムにサンプリングした不正解のトリプルを49個用意する．そして，正解のトリプルの順位が上位N位以内に正解があれば1，なければ0として全評価用データで平均した指標である．ただし，終点ノードのサンプリングの際のノードタイプは正解のトリプルと同一のもの使用した．

本実験では，提案手法で述べたそれぞれの条件での結果を得るための，3種類のモデルを用意した．各モデルの説明は以下の通りである．

- **TransE** 初期表現にランダムな表現を用いて，TransEで学習するモデル．学習したノード表現は768次元である．
- **TransE+BERT** テキストカテゴリ（要旨と題目）について，SciBERTによる埋め込みを初期表現に用いたものを用いて，TransEで学習するモデル．学習したノード表現は768次元である．
- **GraIL** サブグラフを3ホップ抽出して，サブグラフのノードの表現にTransE+BERTで学習した表現を追加して学習するモデル．グラフ構造がスター型となっているため，サブグラフとして抽出されるノード数を考慮して，3ホップでサブグラフ抽出を行う．サブグラフ抽出時の，サブグラフの平均ノード数は4.33個となった．

5 結果

4節の各モデルの評価を表1に示す．参考として，最も良いリンク予測性能が得られたTransE+BERTのモデルについて，ノード表現の近い文献を表示した結果を付録Bに示す．

5.1 TransE と TransE+BERT の比較

まず，TransEとTransE+BERTのモデルを比較する．TransE+BERTのモデルでは，TransEのモデルに比べて，Hit@Nが高い値となった．TransE+BERTのモデルが高い性能を示せた理由として，題目と要旨の情報をもとにした高品質なノード表現を初期表現に使えたためであると考えられる．また，学習速度に関してもTransE+BERTのモデルでは，TransEのモデルに比べて半分以下のエポック数で学習を終えることが可能であった．

このことから，テキストカテゴリのノードの初期表現に大規模事前学習モデルによる埋め込みを用いることは性能と学習速度の両方において有効であるとわかった．

表1 各モデルの著者の正解率

モデル	TransE	TransE+BERT	GraIL
Hit@1	0.7196	0.7660	0.5002
Hit@3	0.8576	0.8870	0.5962
Hit@5	0.9004	0.9262	0.6610
Hit@10	0.9434	0.9624	0.7554

5.2 TransE+BERT と GraIL の比較

次に，TransE+BERTとGraILのモデルを比較する．GraILのモデルでは，TransEのモデルに比べて，Hit@Nが低い値となった．GraILのモデルが低い性能となった理由としてサブグラフの構造が考えられる．今回はグラフをスター型で構築したため，対象ノード間のパスを構成するのに最低必要なホップ数が大きくなってしまったと考えらえる．また．GraILは対象ノードの間に現れるノードしか対象にしないため，文献の周囲の情報を利用できなかったのではないかと考えられる．さらに，出版年などノード間に頻出して現れるノードも悪影響を及ぼしている可能性もある．これらのサブグラフの構築方法の調査は今後の課題である．

6 おわりに

本研究では，文献グラフの高品質な埋め込み表現の獲得のために，大規模事前学習モデルによるテキスト情報の埋め込み，グラフの埋め込み手法，グラフニューラルネットワークを用いたリンク予測手法の利用可能性について，文献グラフ上でのリンク予測を対象に評価を行った．実験の結果，グラフの埋め込み手法の初期表現に大規模事前学習モデルによるテキスト情報の埋め込みを利用したTransE+BERTのモデルが最も良い性能を示した．これより，グラフ埋め込みにおける大規模事前学習モデルの有効性を示した．

今後の課題として，入力とするサブグラフの構造やリンク予測手法を変更した場合の性能への影響を検証する．また，高品質な文献グラフの表現を構築できた際には，柔軟に検索が可能な検索システムの構築などに繋げていきたい．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです.

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. **Advances in neural information processing systems**, Vol. 33, pp. 22118–22133, 2020.

[3] Takuma Yoneda, Koki Mori, Makoto Miwa, and Yutaka Sasaki. Bib2vec: Embedding-based search system for bibliographic information. In **Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics**. Association for Computational Linguistics, 2017.

[4] 米田拓真, 三輪誠, 佐々木裕. 文献情報の多様な要素を考慮したベクトル表現獲得. 言語処理学会 第 24 回年次大会 発表論文集, pp. 996–998. 言語処理学会, 2018.

[5] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. **IEEE transactions on neural networks**, Vol. 20, No. 1, pp. 61–80, 2008.

[6] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 31, pp. 5171–5181. Curran Associates, Inc., 2018.

[7] Komal K. Teru, Etienne G. Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In **Proceedings of the 37th International Conference on Machine Learning**, ICML'20. JMLR.org, 2020.

[8] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: large-scale information network embedding. In **WWW**, pp. 1067–1077, 2015.

[9] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In **Advances in Neural Information Processing Systems**, pp. 2787–2795. Curran Associates, Inc., 2013.

[10] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. **ICLR**, 2017.

[11] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, **The Semantic Web**, pp. 593–607, Cham, 2018. Springer International Publishing.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Association for Computational Linguistics, 2019.

[13] ACL anthology. https://aclanthology.org/ (2023 年 1 月 11 日アクセス).

[14] Semantic scholar academic graph api. https://www.semanticscholar.org/product/api (2023 年 1 月 11 日アクセス).

A　データセットの統計

実験に用いたデータ数の統計を表 2，トリプル数の統計を表 3 に示す．表 3 の文献 ID と文献 ID 以外のトリプル数は双方向の場合の数である．

表 2　データセットに含まれるデータ数

データの種類	データ数
文献 ID	16,916
題目	16,916
要旨	14,771
著者	26,055
年	5
掲載誌	24

表 3　トリプルの統計

トリプルの種類	トリプル数
文献 ID と題目	33,832
文献 ID と要旨	29,542
文献 ID と著者	137,780
文献 ID と年	33,832
文献 ID と掲載誌	33,740
文献 ID と文献 ID	246,801

B　ノード表現の評価の例

TransE+BERT のノード表現について，対象文献 ID に対して近傍の文献 ID 上位 5 件を出した例を表 4 に示す．

表 4　対象文献 ID の近傍の文献 ID の例 (最初の文献 ID は対象を示す)

例 1	
文献 ID	題目
N19-1423	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
N18-1202	Deep Contextualized Word Representations
P18-1031	Universal Language Model Fine-tuning for Text Classification
2020.tacl-1.5	SpanBERT: Improving Pre-training by Representing and Predicting Spans
W18-5446	GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding
P19-1356	What Does BERT Learn about the Structure of Language?
例 2	
文献 ID	題目
2022.acl-long.220	Learned Incremental Representations for Parsing
2022.acl-long.146	Investigating Non-local Features for Neural Constituency Parsing
2022.acl-long.155	Headed-Span-Based Projective Dependency Parsing
2022.acl-long.171	Bottom-Up Constituency Parsing and Nested Named Entity Recognition with Pointer Networks
2020.findings-emnlp.65	Rethinking Self-Attention: Towards Interpretability in Neural Parsing
2021.emnlp-main.826	A New Representation for Span-based CCG Parsing