BERT-based Bi-Ranker による文脈を考慮した引用論文推薦

杉本 海人 †
† 東京大学理学部情報科学科
kaito_sugimoto@is.s.u-tokyo.ac.jp

相澤 彰子 ‡†
‡ 国立情報学研究所
aizawa@nii.ac.jp

1 はじめに

学術論文の総数は昨今急速に増大しており [1]，引用すべき論文を見つけることが困難になりつつある．この問題に対処するべく，引用論文推薦システム（citation recommendation system; 以下，論文推薦システム）に関する研究が進められてきた．これは入力として何らかのテキストを与えたときに，その内容を裏付けるような論文を出力するというもので，Google Scholar [1)] のようなキーワードベースの論文検索システムよりも精度の高い推薦が期待される．

論文推薦に対するアプローチは，大域的な論文推薦と局所的な論文推薦の 2 種類に大別される [2]．前者はある論文の大域的な情報，すなわち本文や要旨全体などを入力とするのに対し，後者は引用文脈と呼ばれる一文から数文程度の短い単語列を入力とする．後者は特に，文脈を考慮した論文推薦とも呼ばれる．表 1 に例を示す（引用元論文の出典は [3]，引用先論文の出典は [4]）．

近年，他の自然言語処理のタスクと同様，ニューラルネットワークが論文推薦の研究においても広く用いられるようになっている．しかしながら 2 節で述べるように， BERT [5] のような Transformer ベースの言語モデルは十分には活用されていない．そこで本研究では BERT-based Bi-Ranker と呼ばれるモデルを，文脈を考慮した論文推薦において用いる．このモデルでは，クエリの引用文脈と推薦する候補の論文の文章の双方を独立に BERT で埋め込み，候補の論文をコサイン類似度でランク付けする．同種のモデルは既にエンティティリンキングや対話行為予測などのタスクで効果が確認されている [3, 6]．

2 種類のデータセットを用いた評価実験により，提案手法が近年のニューラルベースの手法よりも推薦精度が高いことを確認した．また，大域的な情報を追加情報として用いるモデルの有効性も合わせて検証した．

表 1 文脈を考慮した論文推薦における入力・出力例

入力	Instead of using the poly-encoder as a trade-off between cross-encoder and bi-encoder, we propose to train a bi-encoder model with knowledge distillation (Buciluundefined et al., 2006; [*]) from a cross-encoder model to further improve the bi-encoder's performances.
出力	Distilling the knowledge in a neural network (Hinton et al., 2015)

1） https://scholar.google.com/

2 関連研究

文脈を考慮した論文推薦手法は He ら [7] によって提案された [2]．Huang ら [8] は Word2Vec [9] のような 2 層のネットワークを構築し，引用文脈内の単語の埋め込みと引用先論文の埋め込みを同時に取得することで，それらを用いた論文推薦が既存のニューラルネットワークを用いない手法よりも効果的であることを示した．以降，より質の高い埋め込みを得る研究 [10, 11, 12] や，入力の引用文脈から出力の論文までをネットワークで end-to-end に結びつける研究 [13] などが進められてきた．

論文推薦においても BERT [5] を用いた先行研究は存在するが，改良の余地がある．Jeong ら [14] は BERT と GCN [15] を用いて引用文脈の埋め込みを得ることで，LSTM ベースの既存手法 [11] に対してスコアの向上を確認したが，この研究では候補の論文のテキストを活用していない．一方，Cohan ら [16] は，BERT ベースの論文の埋め込みが様々な論文関連のタスクにおいて有効であることを示したが，この手法では引用文脈を入力対象としていないため，文脈を考慮した論文推薦には対応できない．2 つの研究は相補的であり，本研究の提案手法は引用文脈の埋め込みと候補の論文の埋め込みの双方を活用するものと位置づけられる．

3 提案手法

本研究では，文脈を考慮した論文推薦のための BERT-based Bi-Ranker モデルを提案する．このモデ

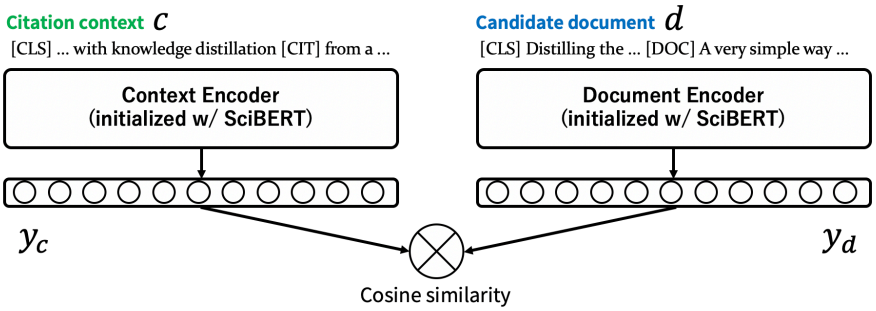

図 1　モデルの概要図

ルは Context Encoder と Document Encoder の 2 種類の BERT エンコーダからなる．前者は引用文脈をエンコードし，後者は論文の内容をエンコードする．2 つのエンコーダはともに学術論文テキストで事前学習が行われた SciBERT [17] で初期化し，訓練データによって fine-tuning を行う．

3.1　通常モデル

モデルの概要を図 1 に示す．以下の手順にしたがって計算を行う．

(a) クエリの引用文脈 c を Context Encoder を用いてベクトル $\boldsymbol{y_c}$ に変換する．
(b) 候補の論文 d を Document Encoder を用いてベクトル $\boldsymbol{y_d}$ に変換する．
(c) スコアをコサイン類似度によって計算する．

$$\text{score}(c, d) = \boldsymbol{y_c} \cdot \boldsymbol{y_d} \tag{1}$$

(a) の詳細について，Context Encoder に入力する引用文脈は，以下のようなトークン列とする．

`[CLS]` lctxt `[CIT]` rctxt

ここで，lctxt は引用位置の左側の文字列，rctxt は引用位置の右側の文字列を表す．また，`[CIT]` は引用位置を意味する特殊トークンである．引用文脈の埋め込み $\boldsymbol{y_c}$ は Context Encoder の `[CLS]` トークンの位置にあたる最後の隠れ層の出力から得る．

(b) の詳細について，Document Encoder に入力する論文に関しては，本研究では論文の題目と要旨を入力ソースとして用いる．入力トークン列は

`[CLS]` title `[DOC]` abstract

とする．ここで，title は題目，abstract は要旨を表す．`[DOC]` は題目と要旨を分ける特殊トークンである．論文の埋め込み $\boldsymbol{y_d}$ は前と同様，Document Encoder の `[CLS]` トークンの位置の出力から得る．

モデルの訓練時においては，ある引用文脈と対応する正解ラベルの論文のスコアが，訓練データの同じバッチ内の他の論文とのスコアよりも高くなるよう学習する．すなわち，引用文脈と対応する論文の組 n 個 (c_i, d_i) $(i = 1, ..., n)$ からなるバッチ内における損失関数を以下のように定める [3, 18]．

$$L(c_i, d_i) = -\text{score}(c_i, d_i) + \log \sum_{j=1}^{n} \exp(\text{score}(c_i, d_j)) \tag{2}$$

推論時においては，入力の引用文脈に対して，候補のそれぞれの論文とのスコアを計算し，スコアが上位の論文から順に推薦を行う．

3.2　大域モデル

Medić ら [12] は，文脈を考慮した論文推薦において，引用元の論文の題目や要旨のような大域的な情報を用いた大域モデルを提案し，特に引用文脈の長さが短い場合には推薦の精度が向上することを確認した．これに倣い，本研究においても引用元の論文の内容を用いた大域モデルの有効性を検証する．

このモデルにおいては，通常モデルと同様に引用文脈の埋め込み $\boldsymbol{y_c}$ を得るだけでなく，引用元の論文の埋め込み $\boldsymbol{y_g}$ を Document Encoder を用いて求める．入力形式は候補論文の埋め込み $\boldsymbol{y_d}$ を得る場合と同様である．

引用文脈の埋め込み $\boldsymbol{y_c}$ と引用元の論文の埋め込み $\boldsymbol{y_g}$ を足し合わせたものを最終的なクエリの埋め込み $\boldsymbol{y'_c}$ とし，スコアを計算する．その他の訓練方法は通常モデルに揃える．

$$\boldsymbol{y'_c} = \boldsymbol{y_c} + \boldsymbol{y_g} \tag{3}$$

$$\text{score}(c, d) = \boldsymbol{y'_c} \cdot \boldsymbol{y_d} \tag{4}$$

4　実験

4.1　データセット

表 2　実験結果（BM25 で得られた候補に対する結果）

	ACL-600		ACL-200		RefSeer	
	Recall	MRR	Recall	MRR	Recall	MRR
NCN [13]	-	-	-	-	0.291	0.267
SemCon [12]	0.568	0.306	0.537	0.291	0.340	0.166
SemEnh [12]	0.553	0.290	0.546	0.285	0.445	0.216
提案手法（通常モデル）	**0.657**	**0.408**	0.611	0.381	0.627	**0.365**
提案手法（大域モデル）	0.651	0.394	**0.638**	**0.382**	**0.645**	0.350

4.1.1 ACL-ARC

ACL-ARC は自然言語処理分野の論文を収録したデータセットである [19]．本研究では [12] の実験設定に従い，ACL-200 と ACL-600 の 2 種類のサブセットを用いる．前者は引用文脈として引用位置の前後 600 文字ずつを，後者は 200 文字ずつを抽出したものである．訓練データ，検証データ，テストデータのサイズはそれぞれ約 30K，9K，10K となっている．また，予測候補の論文の総数は約 20K である．

4.1.2 RefSeer

RefSeer は Web 上に存在する様々な分野の学術記事からなるデータセットである [8]．訓練データ，検証データ，テストデータのサイズはそれぞれ約 3.5M，125K，127K となっている．また，予測候補の論文の総数は約 625K である．

4.2 ベースライン

4.2.1 NCN

Ebesu ら [13] は NCN (Neural Citation Network) というモデルを提案した．これは Attention 機構を備えた Encoder-Decoder モデルであり，引用文脈，引用元の論文の著者名，引用先の論文の著者名の埋め込みを連結したものを Encoder 側，引用先の論文の題目の埋め込みを Decoder 側に与えて学習が行われる．Ebesu らはこのモデルを RefSeer データセットにおいて評価した．具体的には，入力の文脈に対して, まず BM25 [20] で上位 2,048 本の論文を候補として取得した上で，それらの候補を NCN を用いて Re-ranking した．なお，BM25 の時点で正解ラベルの論文が得られなかった場合は, Re-ranking の候補に正解ラベルを加えている．

表 3　Ablation Study の結果（全候補に対する結果）

	ACL-600	
	Recall	MRR
BM25 [12]	0.095	0.049
提案手法（通常モデル）	0.535	0.306
題目のみ	0.461	0.271
BERT で初期化	0.465	0.258

4.2.2 SemCon，SemEnh

Medić ら [12] は, 文脈を考慮した論文推薦システムのための Semantic module と Bibliographic module を考案した．Semantic module は論文や引用文脈の内容を捕捉するのに対し，Bibliographic module は論文のメタ情報（引用数等）を考慮する．本研究では内容による推薦に焦点を置くため，Semantic module のみを比較の対象とする．

Semantic module を用いたモデルは特に SemCon と SemEnh の 2 種類に分かれる．SemCon は引用文脈と候補の引用先論文の内容それぞれの埋め込みを Bi-LSTM を用いて取得し，コサイン類似度を計算する．SemEnh は大域的な情報を用いて SemCon を拡張したモデルである．具体的には，引用元論文の埋め込みを別に計算し，引用文脈の埋め込みと足し合わせてから，コサイン類似度を計算する．

NCN の場合と同様に，Medić らはこれらのモデルを用いて， RefSeer データセットに対しては BM25 で得られた上位 2,048 本の論文を Re-ranking し，ACL-ARC データセットに対しては上位 2,000 本の論文を Re-ranking した．NCN の場合と同様，BM25 の時点で正解ラベルの論文が得られなかった場合は, Re-ranking の候補に正解ラベルを加えている．

4.3 評価手法

先行研究と同様に，提案した論文推薦システムが既存の論文の引用をどの程度予測できるかを測定する．本研究では，先行研究に揃えて，推薦された予測候補の上位 10 件に対して Recall と MRR (Mean Reciprocal Rank) [21] を計算し，ベースラインの手法と比較する．

4.4 結果

表 2 は，提案手法のスコアを既存研究で報告された値と比較したものである．比較の上では，既存研究が BM25 で得られた候補にもとづいて値を報告していることから，本研究でもこれに従った（候補は [12] と同一のものを用いた）．全てのデータセットにおいて，提案手法がベースラインの手法の値を上

表 4　BERT-based Bi-Ranker モデルによる ACL-200 データセットの予測例（太字が正解ラベルの論文を表す）

クエリの引用元論文	Language Identification using Classifier Ensembles
クエリの引用文脈	...be some redundancy in the large number of character ngram features and removing these might increase the diversity and thus accuracy of the ensemble using the feature analysis methodology outlined by [*] we analyzed the feature interactions using the training and development set this methodology uses yule's qcoefficient statistic which can be a useful measure of pairwise dependence between two cla...

BERT-based Bi-Ranker（通常モデル）	BERT-based Bi-Ranker（大域モデル）
1. A Two-level Classifier for Discriminating Similar Languages	1. Language Identification using Classifier Ensembles
2. Measuring Feature Diversity in Native Language Identification	2. A Two-level Classifier for Discriminating Similar Languages
3. Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing	3. Using Maximum Entropy Models to Discriminate between Similar Languages and Varieties
4. A Simple Baseline for Discriminating Similar Languages	4. A Simple Baseline for Discriminating Similar Languages
5. Feature Space Selection and Combination for Native Language Identification	5. The NRC System for Discriminating Similar Languages

クエリの引用元論文	A Computational Approach for Generating Toulmin Model Argumentation
クエリの引用文脈	...tween the claim and data one of the aspects of the toulmin model in terms of determining stance previous work has utilized attack or support claims in user comments as a method for determining stance [*] inspired by hashimoto et al 6's excitatory and inhibitory templates in this work we similarly compose a manual list of promotexy and suppressxy relations and rely on these relations coupled with pos...

BERT-based Bi-Ranker（通常モデル）	BERT-based Bi-Ranker（大域モデル）
1. Cats Rule and Dogs Drool!: Classifying Stance in Online Debate	1. A Computational Approach for Generating Toulmin Model Argumentation
2. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts	2. **Back up your Stance: Recognizing Arguments in Online Discussions**
3. Recognizing Stances in Ideological On-Line Debates	3. Contrasting Opposing Views of News Articles on Contentious Issues
4. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates	4. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions
5. Stance Classification in Online Debates by Recognizing Users' Intentions	5. Stance Classification in Online Debates by Recognizing Users' Intentions

回っている．また，ACL-600 と ACL-200 の結果を比較すると，引用文脈の長さが短い ACL-200 において大域モデルが効果的であることがわかる．

提案手法の予測例を表 4 に示す．上段の例においては，通常モデルの予測候補の論文は "Feature" という単語を含むものが多い．これはクエリの引用文脈の "feature analysis" という部分をより考慮に入れた結果であると考えられる．これに対し，下段の例においては，大域モデルの予測候補においてのみ "Arguments" や "Argumentation" という単語が見られる．これは，クエリの論文が Argumentation に関する論文であるという大域的な情報が予測に反映されていると考えられる．

論文推薦システムを実用化する上では，推薦候補の論文の要旨をテキストとして抽出するのが困難なケースも考えられる．そこで，提案手法において，論文の要旨を用いず，題目のみを埋め込んだ場合の性能を調べた．加えて，両方のエンコーダを SciBERT ではなく通常の事前学習済み BERT [5] で初期化した場合における性能を調べた．結果を表 3 に示す．題目のみを情報として用いた場合でも，[12] で報告された BM25 のベースラインのスコアを大きく上回っていることが分かる．また，BERT で初期化した場合，題目のみの場合と同程度スコアが下がってしまうことから，学術論文テキストによる事前学習の重要性が示唆される．

5　おわりに

本研究では，引用文脈と論文の文章の双方を BERT で埋め込んだモデルを提案し，文脈を考慮した論文推薦タスクにおいて効果を検証した．今後の課題として，論文の要旨以外の内容（手法や実験結果など）の埋め込みが推薦に役立つかどうかの検証や，既知の引用先の論文がある場合にそれを入力として有効活用する手法の検討などが挙げられる．

謝辞

本研究は，JST CREST-20218985 の支援を受けたものである．

参考文献

[1]Mark Ware and Michael Mabe. The stm report: An overview of scientific and scholarly journal publishing. 2015.

[2]Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, Vol. 21, No. 4, p. 375–405, Aug 2020.

[3]Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6397–6407, Online, November 2020. Association for Computational Linguistics.

[4]Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[5]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.

[6]Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. 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)*, pp. 673–683, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7]Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 421–430, 2010.

[8]Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C Giles. A neural probabilistic model for context based citation recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29, 2015.

[9]Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[10]Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. hyperdoc2vec: Distributed representations of hypertext documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2384–2394, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11]L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, L. Guo, and T. Dai. A lstm based model for personalized context-aware citation recommendation. *IEEE Access*, Vol. 6, pp. 59618–59627, 2018.

[12]Zoran Medić and Jan Snajder. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the First Workshop on Scholarly Document Processing*, pp. 97–103, Online, November 2020. Association for Computational Linguistics.

[13]Travis Ebesu and Yi Fang. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, p. 1093–1096, New York, NY, USA, 2017. Association for Computing Machinery.

[14]Chanwoo Jeong, Sion Jang, Eunjeong L. Park, and Sungchul Choi. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics*, Vol. 124, No. 3, pp. 1907–1922, 2020.

[15]Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, Vol. abs/1609.02907, , 2016.

[16]Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, July 2020. Association for Computational Linguistics.

[17]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)*, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[18]Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[19]Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[20]Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[21]Ellen M. Voorhees. The TREC-8 question answering track report. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, Vol. 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 1999.