推薦理由提示のための アブストラクトの観点に基づく学術論文推薦

小林恵大 [1]　Qi Yang[1]　成松宏美 [2]　南泰浩 [1]
[1] 電気通信大学　[2]NTT コミュニケーション科学基礎研究所
{k2231042,s2331068}@edu.cc.uec.ac.jp　hiromi.narimatsu.eg@hco.ntt.co.jp
minami.yasuhiro@is.uec.ac.jp

概要

学術論文数の急増によって研究者の論文調査の負担が増大し，論文推薦システムの重要性が高まっている．本稿では，類似する論文リストだけでなく，要旨の観点に着目した推薦理由も提示する推薦手法を提案する．従来手法は，クエリの論文と推薦候補論文群との類似性を文章全体の内容で判断しており，どの点で類似するかを説明することは困難であった．本稿では，要旨中の各文を背景，手法，結果の観点に分類し，Transformer Encoder を用いて観点の埋め込み表現を生成し，観点ごとに類似度を算出することで観点の類似性を根拠として論文推薦を行う．観点に基づく論文推薦タスクのベンチマークデータ CSFCube を用いて提案手法を評価した結果，従来手法を超える精度を示した．

1　はじめに

学術論文の出版数は急速に増加している [1]．この情報源の増加は研究の深化を促進し，新たな発見の可能性を広げる一方で，論文を調査，理解するために必要な時間と労力も増加している．このような背景から，研究者の論文調査や執筆を支援する論文推薦システムに関する様々な研究が行われている [2, 3, 4]．論文推薦は，研究の初期段階での論文調査や試行錯誤の段階での新しいアイデアの発見を支援するものであり，研究者の負担を大幅に軽減したり，見落としていた論文を発見したりする効果が期待されている [5]．

論文推薦では，論文のタイトルやアブストラクト（以下「要旨」とする）を入力として，類似性の高い論文を提示することが求められる．例えば，研究背景や手法，得られた知見や評価方法などが類似する論文を提示する．しかし，これらの要素が部分的に類似する論文の場合，ユーザはタイトルを見ただけではその類似性を判断できない．したがって，論文推薦システムには，単に類似論文のリストを提示するだけでなく，背景やアプローチが類似するといった具体的な観点に基づく推薦理由を提示することが期待される．

従来の論文推薦研究は，類似論文や引用論文のリストを推薦することに主眼が置かれており，推薦理由を提示することを目的としていないものが多い [2]．推薦理由の提示につながる観点ごとの類似に着目した研究では，「背景」「手法」「結果」のいずれかの観点を条件として類似する論文を推薦するタスクを提案し，その評価のために観点に沿った類似度がアノテーションされたベンチマークデータ CSFCube が提案されている [6]．

本研究では，論文推薦において，論文間の観点の類似性を根拠として論文を推薦するモデルを構築し，評価する．具体的には要旨分類モデル [7] を用いて論文の要旨を「背景」「手法」「結果」に分類し，Transformer Encoder [8] を用いて各観点の埋め込み表現を生成する．この埋め込み表現に基づき Triplet Loss [9] を用いた距離学習によってモデルを学習し，観点に基づく論文推薦タスクのベンチマークデータである CSFCube [6] を用いて評価する．

2　関連研究

論文推薦に関する研究は主に 2 点のアプローチに分けられる．1 点目はユーザの過去の論文閲覧履歴や執筆した論文，共著者，引用などの情報を利用する手法である．ユーザの閲覧履歴を利用した協調フィルタリング手法や，引用ネットワークを分析し著者や論文の類似性を特定するグラフベースの手法が提案されている [2]．これらの手法は，既に広く引用・閲覧されている論文を効果的に推薦できる

が，被引用数が少ない論文や新しく発表された論文を見逃す可能性がある．

2 点目は論文のタイトルや要旨などのテキストの類似度に基づくものである．このアプローチでは，論文のタイトルや要旨を Doc2Vec [10] や BERT [11] を用いて埋め込み表現に変換し，論文間の類似性を判定する [12, 13]．Cohan らは引用関係を元にして正例負例をサンプリングし，距離学習を行ったモデル SPECTER を提案している [14]．また，SPECTER を改良し，引用グラフの埋め込みを利用した最近傍サンプリングに基づいて論文の類似性を学習したモデル SciNCL も提案されている [15]．これらの手法は論文間の全体的な類似性に基づく推薦タスクでは高い精度を達成している．しかし，推薦の理由を具体的に示すことは困難である．

一方，特定の観点に沿った論文を推薦する手法やベンチマークデータも提案されている．Takahashi らは，ユーザの検索クエリに着目し，「研究背景や目的が類似しているが，手法が異なる」論文を推薦する手法を提案している [16]. この手法では，BERT [11] ベースの要旨分類モデル [7] を用いて，要旨中の「背景」「目的」「手法」の観点に対応する文を抽出し，SciBERT [17] を用いてそれらの観点の埋め込み表現を生成し，コサイン類似度で推薦論文を決定する．Mysore らは，指定された「背景」「手法」「結果」のいずれかの観点を条件として類似する論文を推薦するタスクに着目し，専門家によってアノテーションされたベンチマークデータ CSFCube [6] を提案している．

本研究では，推薦理由の提示を可能にする各観点の類似度に着目し，要旨分類モデル [7] と，SPECTER [14] や SciNCL [15] と同様の論文間の類似性を反映した埋め込みの学習手法を組み合わせて，観点ごとの距離学習が可能なモデルを構築し，CSFCube データセット [6] で評価する．

3 提案手法

論文間で「タイトル」「背景」「手法」「結果」の各観点の類似度を算出するために，観点の埋め込み生成に特化したモデルを構築する．以下では，モデルの構造と学習方法について述べる．

3.1 モデル構造

提案モデルは，論文のタイトルと要旨を入力すると，「タイトル」「背景」「手法」「結果」の 4 つの観点の埋め込み表現（以下「観点埋め込み」とする）を出力する（図 1）．このモデルは，(1) 論文要旨の各文を観点に分類する要旨分類モデル（以下「SSC」[1] とする），(2) 論文のタイトルと要旨を入力として，入力の各トークンに対応するベクトルを出力する Transformer Encoder，(3) 観点埋め込みを生成する Pooling Layer，の 3 点から構成される．

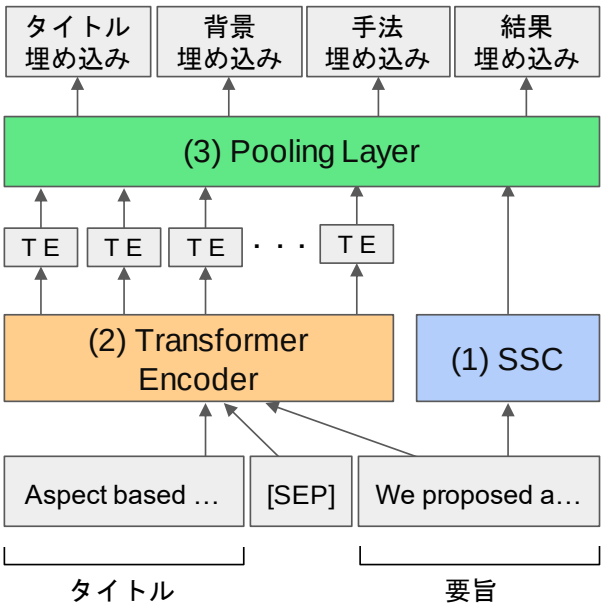

図 1 提案モデルの全体像．ここで，「TE」は各入力トークンに対応する埋め込み（Transformer Encoder の最終隠れ層の出力）を表す．タイトルと要旨を入力として，各観点の内容を判別し，観点の埋め込み表現を生成する．

(1) SSC. Cohan らの学習済みモデル [7] を追加学習することなくそのまま利用し，入力要旨の各文がどの観点に属するかを判定する．このモデルは要旨を入力すると，各文に「背景」「目的」「手法」「結果」「その他」のいずれかのラベルを付与するものであり，各文に上記観点のアノテーションが付与された 2,189 件のコンピュータサイエンス分野の英語論文要旨で SciBERT [17] が学習されている．ただし，「その他」は観点が曖昧であるため本研究では使用しない．また，「目的」と「背景」は同じ 1 文に記述されることが多く，自動認識で「背景」に分類され「目的」の観点が要旨に含まれなかったり，「目的」そのものが「背景」と類似する内容が記述されることが多い．そのため本研究では「目的」の観点は「背景」の観点と同じものと考える．

(2) Transformer Encoder. 論文のタイトルと要旨を入力し，トークナイズされた各トークンに対応

1) Sequential Sentence Classification（https://github.com/allenai/sequential_sentence_classification）の略である．

する埋め込みを生成する．モデルの初期値には，学術論文のコーパスで事前学習されたモデルである SciBERT[2] [17] を用いる．具体的には，タイトルと要旨を [SEP] トークンで連結したものを入力し，入力トークンに対応する最終隠れ層の出力（last_hidden_state）を出力する．例えば，入力トークン数が 100 だった場合，隠れ層の次元数を d_{model} として，$(100 \times d_{model})$ のサイズのテンソルを出力する．

(3) Pooling Layer. Transformer Encoder からの出力を集約し，各観点の文脈を反映した観点埋め込みを生成する．具体的には，SSC の判定結果を元に Transformer Encoder の出力の各トークン埋め込みを観点ごとにスタックし，各観点について (観点のトークン数$\times$ d_{model}) のサイズのテンソルを得る．それらを Attention モデルに入力し，(観点数$\times$ d_{model}) のサイズのテンソルを観点の埋め込みとして生成する．Attention モデルは以下の式で表される．

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{Q(W_K K)^T}{\sqrt{d_{model}}}\right)\{V + PE(V)\} \tag{1}$$

ここで，Q（Query）は学習可能なパラメータ，K（Key）・V（Value）はスタックされた各観点のトークンの埋め込み，W_K は Key に適用される線形変換の学習可能なパラメータ，PE は Positional Encoding を表す．この式では，Query と線形変換された Key の類似度スコアを計算し，そのスコアに基づいて，PE を加えた Value の重み付け和を取ることで，トークンの埋め込みを集約し，観点埋め込みを生成する．

3.2 学習

論文の類似性を反映した埋め込みを学習するため，SPECTER [14] や SciNCL [15] のような論文埋め込みモデルと同様に Triplet Loss [9] を用いた距離学習を行う．

Triplet Loss を用いた損失は次の式で表される．

$$L = \max\{0, m + d(q,p) - d(q,n)\} \tag{2}$$

ここで，q はクエリ論文の埋め込み，p は正例（類似する論文）の埋め込み，n は負例（類似しない論文）の埋め込み，$d(\cdot,\cdot)$ はユークリッド距離，m は調整可能なマージンを表す．ただし，q，p，n は，それらを一般に x として，次の式で表されるように観点の埋め込みを平均化したものである．

$$x = \frac{1}{|Labels|}\sum_{l \in Labels} x_l \tag{3}$$

ここで，$Labels$ は観点，x_l は観点 l の埋め込みを表す．

ただし，上記では学習データは SPECTER や SciNCL と同様に，論文全体の類似性に基づいて作成されたデータを用いる．これは，このデータセット規模の各観点の類似性を反映したデータセットが現状は存在せず，データセットを作成するにも観点のラベルをつけるコストが膨大であるためである．このデータを用いた学習では，どの観点が類似するかしないかという教師信号は与えられないため，実際に類似しない観点があった場合でも，全ての観点について埋め込みが互いに近くなるように学習される．

4 観点に基づく論文推薦評価

4.1 評価設定

評価データ. 評価には CSFCube [6] を用いる．CSFCube は，特定の観点を条件とした論文推薦を評価するためのデータセットである．例えば，クエリと類似する手法を持つ論文を探すようなケースが想定される．このデータセットは，50 件のクエリ論文と各クエリ論文に対する 100〜250 件の候補論文に指定の観点に沿った類似度が 4 段階でアノテーションされている．含まれる論文は全てコンピュータサイエンス分野の英語論文である．なお，このデータセットにおける観点は本研究の同様の要旨分類モデル [7] を用いて決定されており，観点の定義は本研究と一致する．

学習設定. 提案モデルは，S2ORC データセット [18] から構築された約 68 万トリプルの SciNCL 学習データ [15] を用いて学習される．学習設定は，epoch 数を 2，学習率を 2e-6，Triplet Loss の margin を 1 とし，バッチサイズは学習に使用した GPU（GTX 1080Ti，11GB）に収まる値で 2 に設定した．Optimizer は Adam [19] を用いた．

ベースライン. 比較対象として，論文推薦における最先端の手法である SPECTER [14]，SciNCL [15] を用いる．また，OpenAI の Embedding Model の text-embedding-ada-002 [20] も比較対象とする．このモ

2） https://huggingface.co/allenai/scibert_scivocab_cased

表 1 CSFCube による評価結果（全観点を総合した結果）

Model	MRR	MAP	Recall@20	NDCG
SPECTER [14]	0.612	0.314	0.437	0.730
SciNCL [15]	0.550	0.348	0.511	0.745
OpenAI Emb [20]	0.629	0.386	0.574	0.775
提案手法	**0.647**	**0.412**	**0.579**	**0.781**

デルは OpenAI の API を介して埋め込みを生成することができ，情報検索（IR）タスクによる評価で最先端に近い性能を示したことが報告されている [20, 21]．これらの手法では，指定された観点に属する文章を入力とし，その出力（[CLS] トークンの出力もしくは API の返り値）を観点埋め込みとする．

評価方法. クエリ論文と候補論文集合の間で指定の観点の類似度を計算し，その類似度に従って候補論文をランク付けをする．その結果から，類似度が高い論文が上位にどの程度位置しているかを評価する．類似度の計算は，各手法を用いてクエリ論文と候補論文集合の観点埋め込みを生成し，埋め込み間のユークリッド距離を計算することによって行う．評価指標には，推薦タスクで一般的な指標である MRR，MAP，Recall，NDGC を使用する[3]．この評価プロセスは Mysore らによって公開されているプログラムを用いて実施した[4]．

4.2 評価結果

初めに，全ての観点の評価を総合した結果，すなわち観点によらず評価スコアを算出した結果を表 1 に示す．この結果では，全ての指標で提案手法がベースラインを上回った．これにより，提案手法のタイトルと要旨全体を入力として，Attention 機構で周辺文脈を反映した観点埋め込みを生成する手法が有効に働いたと考えられる．また，ベースラインの中では，OpenAI Emb [20] が最も高い性能を示した．このモデルは学術論文を扱うタスクに特化している訳ではなく，他の IR タスクでも高い性能を発揮していることから，分野を問わない高い汎用性を持つと考えられる．論文埋め込みに特化した既存手法では，論文間の離散的な引用関係から類似性を学習した SPECTER [14] よりも，引用グラフ埋め込みによる連続的な引用関係から学習した SciNCL [15] が高い性能を示した．

続いて，表 2 に示す各観点の評価では，「背景」

表 2 CSFCube による評価結果（観点ごとの結果）

観点：背景

Model	MRR	MAP	Recall@20	NDCG
SPECTER [14]	0.689	0.431	0.443	0.807
SciNCL [15]	0.725	0.449	0.538	0.831
OpenAI Emb [20]	0.651	0.468	0.581	0.841
提案手法	**0.838**	**0.532**	**0.601**	**0.866**

観点：手法

Model	MRR	MAP	Recall@20	NDCG
SPECTER [14]	0.385	0.193	0.332	0.645
SciNCL [15]	0.330	0.209	0.407	0.642
OpenAI Emb [20]	**0.497**	**0.271**	**0.510**	**0.681**
提案手法	0.409	0.266	0.500	0.669

観点：結果

Model	MRR	MAP	Recall@20	NDCG
SPECTER [14]	0.747	0.353	0.534	0.738
SciNCL [15]	0.609	0.391	0.591	0.767
OpenAI Emb [20]	**0.743**	0.424	0.634	0.808
提案手法	0.709	**0.445**	**0.638**	**0.816**

「結果」において提案手法がほとんどの指標でベースラインを上回った．「手法」では全ての手法が他の観点より大きくスコアを落としている．これは Mysore ら [6] が指摘するように，「手法」が直接的に類似するのではなく，構造的に類似するようなケースでモデルが判断を誤りやすいことが原因として考えられる．その中でも OpenAI Emb が「手法」において最も高いスコアを示している．OpenAI Emb が多くの IR タスクへの汎化性能を示していることもあり，構造的な類似性がある状況下での難しい類似性を捉える能力が高い可能性がある．

5 おわりに

本稿では，論文推薦において，研究者の論文リストと共に推薦理由も提示することを目的として，要旨の観点に着目し，観点の埋め込み表現を生成することに特化した論文推薦モデルを提案した．提案手法を観点に基づく論文推薦タスクのベンチマークデータである CSFCube を用いて評価し，従来手法を超える精度を示した．今後の展望としては，提案手法の学習において類似しない観点の埋め込みも近づける学習手法を改善し，より良い埋め込みを生成するモデルを構築することがあげられる．また，提案手法を組み込んだ論文推薦システムを実装し，ユーザによる実環境での評価も行いたい．

3） これらの評価指標の詳細は付録 A に記載
4） https://github.com/iesl/CSFCube

謝辞

研究の遂行にあたり，ご助言をいただきました，秋田県立大学 堂坂浩二教授，NTT コミュニケーション科学基礎研究所 杉山弘晃氏，東中竜一郎氏，大阪工業大学 平博順教授，工学院大学 大和淳司教授，国立研究開発法人科学技術振興機構 菊井玄一郎氏に感謝いたします．

参考文献

[1] Science and Technology Observatory (OST). **Dynamics of scientific production in the world, in Europe and in France**. 2019.

[2] Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. Scientific paper recommendation: A survey. **IEEE Access**, Vol. 7, pp. 9324–9339, 2019.

[3] Christin Katharina Kreutz and Ralf Schenkel. Scientific paper recommendation systems: A literature review of recent publications. Vol. 23, No. 4, 2022.

[4] Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira. Task definition and integration for scientific-document writing support. In **Proceedings of the Second Workshop on Scholarly Document Processing**, pp. 18–26, 2021.

[5] 成松宏美, 小山康平, 堂坂浩二, 田盛大悟, 東中竜一郎, 南泰浩, 平博順. 学術論文における関連研究の執筆支援のためのタスク設計およびデータ構築. 言語処理学会第 27 回年次大会, 2021.

[6] Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. CSFCube – a test collection of computer science research articles for faceted query by example. **arXiv preprint arXiv:2103.12906**, 2021.

[7] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. Pretrained language models for sequential sentence classification. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 3693–3699, November 2019.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[9] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In **International Workshop on Similarity-Based Pattern Recognition**, 2014.

[10] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, **Proceedings of the 31st International Conference on Machine Learning**, pp. 1188–1196, 2014.

[11] 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, June 2019.

[12] Zafar Ali, Guilin Qi, Khan Muhammad, Bahadar Ali, and Waheed Ahmed Abro. Paper recommendation based on heterogeneous network embedding. **Knowledge-Based Systems**, Vol. 210, p. 106438, 2020.

[13] Andrew Collins and Joeran Beel. Document embeddings vs. keyphrases vs. terms for recommender systems: A large-scale online evaluation. In **2019 ACM/IEEE Joint Conference on Digital Libraries**, pp. 130–133, 2019.

[14] 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, July 2020.

[15] Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, December 2022.

[16] Tetsuya Takahashi and Marie Katsurai. Solutiontailor: Scientific paper recommendation based on fine-grained abstract analysis. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, **Advances in Information Retrieval**, pp. 316–320, 2022.

[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**, pp. 3615–3620, November 2019.

[18] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, July 2020.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **International Conference on Learning Representations Vol. 5**, 2015.

[20] Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. New and improved embedding model, 2022. https://openai.com/blog/new-and-improved-embedding-model, Accessed: 15 December, 2023.

[21] Jimmy Lin, Ronak Pradeep, Tommaso Teofili, and Jasper Xian. Vector search with openai embeddings: Lucene is all you need. **arXiv preprint arXiv:2308.14963**, 2023.

A　評価指標

ここでは，4 章で使用した，論文推薦タスクの評価指標を説明する．

Mean Reciprocal Rank (MRR): MRR は，類似する各論文が最初に現れるランクの逆数の平均を計算する．つまり，類似する各論文について，その論文がランク付けされた位置の逆数を計算し，それらの平均を取る．

Mean Average Precision (MAP): MAP は，ランク付けされた各論文における平均精度の平均値を示す．各論文での平均精度を計算し，それらの平均をとることで MAP が求められる．

Recall@20: Recall@20 は，類似するすべての論文の中で，上位 20 件にランクされたものの割合を計測する．これは，検索結果の上位 20 件が全体の類似論文をどれだけ網羅しているかを示す指標である．

Normalized Discounted Cumulative Gain (NDCG): NDCG の計算は，まず各ランクにおける Discounted Cumulative Gain（DCG）を計算することから始まる．DCG は次のように定義される：

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{4}$$

ここで，p は考慮されるランキングのポジション数，rel_i は位置 i における論文の類似度を示す．

次に，理想的なランキングにおける DCG，すなわち Ideal DCG（IDCG）を計算する．これは，類似度が最高から最低までの順序で論文をランク付けした場合の DCG である．NDCG は，得られた DCG を IDCG で正規化することによって計算される：

$$\mathrm{NDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p} \tag{5}$$

NDCG は 0 から 1 の間の値をとり，1 に近いほどランキングの品質が高いことを示す．