

利用手法の類似性に着目した学術論文推薦手法の提案

QI YANG¹ 成松宏美² 南泰浩¹

¹電気通信大学 ²NTT コミュニケーション科学基礎研究所
s2331068@edu.cc.uec.ac.jp hiromi.narimatsu@ntt.com
minami.yasuhiro@is.uec.ac.jp

概要

学術論文数の急増により、論文推薦システムの重要性が高まっている。特に、「背景」「手法」「結果」といった観点のうち、どの観点が類似するかを示すことは、利用者の負担削減に有効である。これに対して、従来は要旨内の各文を観点に分類し、各観点類似度をもとに推薦理由を提示する手法がとられていた。しかし、これまで「手法」が類似している論文に対する推薦精度が低かった。本研究では、要旨内の「手法」の類似性に着目した新たな推薦手法を提案する。具体的には、要旨から分野情報を抽出し、分野情報を表す単語の埋め込みを調整することで、「手法」が類似している論文の推薦精度を向上させる。ベンチマークデータで提案手法を評価した結果、従来手法を上回る精度を示した。

1 はじめに

学術論文の出版数は急速に増加しており、この膨大な情報源は、研究の深化や新たな発見の可能性を広げる一方で、論文を調査し理解するための時間や労力を大幅に増加させている [1]。このような背景の中、研究者の負担を軽減し、効率的な文献調査や執筆支援を目的とした論文推薦システムの開発が進められている [2, 3, 4]。

論文推薦システムは、研究の初期段階にお

ける関連文献の探索や、新しいアイデアを発見するプロセスを支援するものであり、研究者の負担軽減や重要な文献の見落とし防止につながることを期待されている [5]。特に、研究背景や手法、得られた知見、評価方法などが類似する論文の提示が求められており、「背景」「手法」「結果」のどの観点が類似しているかの根拠と合わせて類似論文を推薦できる手法が重要である [6]。

しかし、従来の研究では「手法」に着目した推薦において、他の観点到に着目した推薦と比較して推薦精度が低いという課題があった [7]。本研究では、観点到に基づく論文推薦において「手法」の観点到での推薦精度向上を目的とし、新たな論文推薦手法を提案する。自然言語処理ツール stanza [8] を活用して要旨から研究分野に関連する重要なテキストを抽出し、埋め込み表現を適切に調整することで、従来手法を超える精度を実現する。また、提案手法の有効性を観点到に基づく論文推薦タスクのベンチマークデータセット CSFCube を用いて評価する。

2 関連研究

従来研究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が主である。具体的なアプローチとして、論文のタイトルや要旨を Doc2Vec [9] や BERT [10] 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 [11, 12] 方法がある。Cohan らは引

用関係をに基づき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 [13]. 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 [14]. 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。また、Mysore らは「背景」「手法」「結果」の観点ごとに類似する論文を推薦するタスクに着目し、専門家がアノテーションを行ったベンチマークデータセット CSFCube を提案している [6]. これに対し、我々は、SPECTER と観点分類モデル SSC を組み合わせ、「背景」「手法」「結果」のいずれかの観点を条件に、類似する論文を推薦するモデルを提案してきた [7]. この手法では、要旨内の各文に「背景」「手法」「結果」のラベルを付与し [15], Transformer Encoder [16] を用いて各文を埋め込み表現に変換する。その後、生成された埋め込みを観点ごとに分類し、分類結果を連結して Pooling Layer に入力することで観点的意味を抽出し、Attention を用いて観点ごとの埋め込み表現を生成する。各観点的の評価では、「背景」「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。しかし、「手法」では全ての手法で他の観点より大きくスコアを落としている。

本研究では、「手法」が類似している論文の推薦精度向上を目的とし、従来手法の課題を克服する新たな手法を提案する。

3 従来手法のエラー分析

従来手法は [7] 埋め込みを観点ごとに Attention モデルに入力し、SPECTER [13] と同様に TripletLoss [19] を用いた距離学習を行う。しかし、この手法によって観点ごとの推薦精度は向上したものの、「手法」の観点については、低いままであった。これは

クエリ論文 [17]: paperid: 80628431; YEAR: 2019

TITLE: Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages

ABSTRACT(日本語訳): 我々はこのような問題を解決するためのパラダイムを提案する。すなわち、1つ以上の類型論的パラメータにおいて英語と異なる英語の合成版を作成し、構文解析された英語コーパスに基づき、それらの言語のコーパスを生成する。これらの合成言語それぞれにおける動詞の一致特徴を予測するために RNN を学習した一連の実験について報告する。

図 1: クエリ論文

候補論文 [18]: paperid: 52113185; YEAR: 2018 類似度:(+3)

TITLE: Targeted Syntactic Evaluation of Language Models

ABSTRACT(日本語訳): 言語モデルは非文法的な文よりも文法的な文に高い確率を割り当てることが期待される。このデータセットを使った実験では、LSTM 言語モデルは多くの構文で悪い結果を出した。

図 2: 候補論文

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。以下では、実例の一つについて述べる。

図 1 と図 2 は、二つの論文は利用した「手法」が類似しているものの、従来手法が類似しないと判定したものである。

クエリ論文と候補論文は、どちらも自然言語の文法や文構造を学習することに焦点を当てている。また、手法の観点で見ると、どちらも RNN や LSTM といった再帰型のニューラルネットワークを用いており、手法は類似と判断されるのが望ましい。従来手法は要旨全体の文脈を反映した埋め込みを元に類似度を判定するため、類似しない他の観点的分野情報がノイズとなった可能性がある。

4 提案手法

従来手法は Attention 機構を利用し、周辺文脈に基づく埋め込みを生成することに

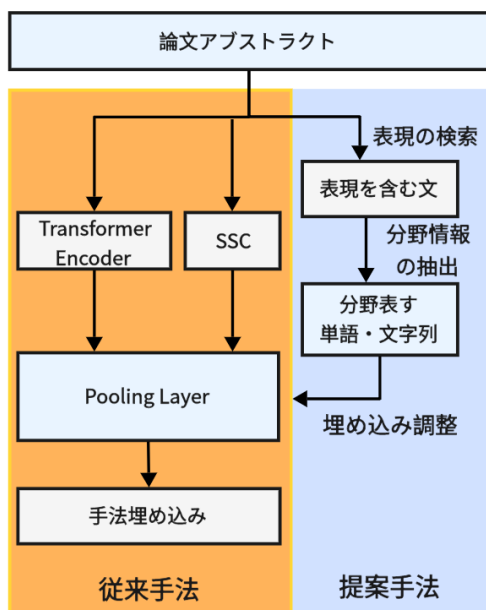


図 3: 従来手法および提案手法の全体像

よって、精度が向上したと推察できる。しかし、周辺文脈の「背景」「結果」に述べられている分野の情報に影響を受けることによって、逆に推定誤りを引き起こす可能性もある。そのため、例に示すような2つの論文では、使用した手法は類似しているが、論文の分野は異なることから、誤推定をしやすい対象となっている。本研究は周辺文脈に記載されている分野情報の影響を避けるための手法を提案する。以下では、提案手法のアルゴリズムの流れについて述べる。

4.1 分野情報を含む文の特定

大規模論文データベース [5] を用いて、732本の論文、5091文から、以下のような周辺文脈に分野情報を含む表現を特定する。検索する表現は大規模論文データベースでよく使われ、かつ文中に研究分野に関する情報を含む65の表現を特定しました。

- applied to
- used in

- employed in
- in the field of
- predicated on...

4.2 分野を特定する単語の抽出例

テキストデータから特定の分野に関連する単語やフレーズを抽出するために、前節で「分野情報を含む文」とされた文に対して、Stanza [8] にて構文解析を行い、固有名詞と判断された単語およびフレーズが、65個の表現の後に出現していれば、該当の単語およびフレーズを分野を表す単語として抽出する。

4.3 抽出した単語の重みの調整

抽出された単語に対し、従来手法のモデルの埋め込み層で重みを調整し、モデルが分野に依存せず正確な判断を行えるようにする。具体的には、従来手法モデルの埋め込み層から、抽出された分野特定単語の埋め込みベクトル v を取得する。各分野特定単語に対して、分野の単語の影響が小さくなるように重み調整係数 α を設定する。調整後の埋め込みベクトル v' は $v' = \alpha \cdot v$ 式で計算される。

ここで、 $0.200 < \alpha < 1.000$ とし、 α の値が小さくするほど単語の影響を低減できる。係数 α は実験的に最適な値を探索する。調整後の埋め込みをモデルの埋め込み層に反映させる。これにより、分野特定単語の影響力が制御でき、モデルが分野依存せずに他の文脈情報に基づいて判断を行えるようになる。

5 評価

5.1 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。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基づく類似度が 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。

5.2 評価方法

評価は Mysore らによって公開されているプログラムを用いた¹⁾。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、候補論文をランク付けする。その結果から、類似度が高い論文が上位にどの程度に位置しているかを評価する。類似度の計算では、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。評価指標には、推薦タスクで一般的な指標である Recall@20 を使用する。

5.3 評価結果

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 [13], SciNCL [14], OpenAI の Embedding Model [20] と従来手法 [7] (「Specter+Attention」と記述する) を用いる。また、OpenAI の Embedding Model の text-embedding-ada-002 も比較対象とする。このモデルは OpenAI の API を介して埋め込みを生成でき、情報検索 (IR) タスクによる評価で最先端に近い性能を示すと報告されている。これらの手法では、指定された観点到に属する文章を入力とし、その出力を観点埋め込みとする。表 1 に示すように、「手法」が類似している論文の推薦精度がベースラインを上回った。なお、重み調整係数による精度の変化を表 2 に示す。結果、重み調整係数 α を 0.800 0.700 倍付近に調整したときに、モデルの精度が向上していることが確認できる。重み調整係数 α を 0.671 倍に調整した

1) <https://github.com/iesl/CSFCube>

表 1: CSFCube による手法観点の評価結果

Model	Recall@20
SPECTER [13]	0.332
SciNCL [14]	0.407
OpenAI Emb [20]	0.510
SPECTER+Attention [7]	0.500
提案手法	0.524

表 2: CSFCube による手法観点の評価結果 (重み係数 α を調整した結果)

重み調整係数 α	Recall@20
1.000	0.500
0.900	0.506
0.800	0.511
0.700	0.520
0.671	0.524
0.600	0.508
0.500	0.405

ときに、精度が最も高い。この結果は、分野情報に関連する単語の重みを適切に調整することで、モデルが異なる分野間での判断誤りを効果的に低減できることを示唆している。

6 おわりに

本稿では、論文推薦において、要旨の観点に着目し、「手法」同士が類似している論文の推薦精度を向上させるため手法を提案した。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 を用いて評価した結果、従来手法を超える精度を示した。今後の展望としては、CSFCube のような観点に基づくアノテーションがなされているデータセットでの fine-tuning 学習によって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、分野情報の抽出手法を改善し、アルゴリズムが分野情報をより正確に抽出することでモデルの精度向上を目指したい。

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、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] B. Xiaomei, W. Mengyang, L. Ivan, Y. Zhuo, K. Xiangjie, and X. Feng. Scientific paper recommendation: A survey. IEEE Access, 2019.
- [3] K. C. Katharina and S. Ralf. Scientific paper recommendation systems: A literature review of recent publications. Int. J. Digit. Libr., 2022.
- [4] N. Hiromi, K. Kohei, D. Kohji, H. Ryuichiro, M. Yasuhiro, and T. Hirotooshi. Task definition and integration for scientific-document writing support. In Proc. 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] S. Mysore, T. O’Gorman, A. McCallum, and H. Zamani. CSFCube – a test collection of computer science research articles for faceted query by example. arXiv preprint arXiv:2103.12906, 2021.
- [7] 小林恵大, 小山康平, 成松宏美, 南泰浩. アブストラクトの観点に基づく学術論文推薦手法の検証. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 4Xin105–4Xin105, 2023.
- [8] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020.
- [9] Quoc Le and M. Tomas. Distributed representations of sentences and documents. In X. Eric P. and J. Tony, editors, Proc. of the 31st International Conference on Machine Learning, pp. 1188–1196, 2014.
- [10] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL2019: Human Language Technologies, Volume 1, 2019.
- [11] Z. Ali, G. Qi, K. Muhammad, B. Ali, and W. A. Abro. Paper recommendation based on heterogeneous network embedding. Knowledge-Based Systems, 2020.
- [12] C. Andrew and B. Joeran. Document embeddings vs. keyphrases vs. terms for recommender systems: A large-scale online evaluation. In 2019 ACM/IEEE Joint Conference on Digital Libraries, 2019.
- [13] C. Arman, F. Sergey, B. Iz, D. Doug, and W. Daniel. SPECTER: Document-level representation learning using citation-informed transformers. In Proc. of ACL2020, 2020.
- [14] O. Malte, R. Nils, A. Isabelle, G. Bela, and R. Georg. Neighborhood contrastive learning for scientific document representations with citation embeddings. In Proc. of EMNLP2022, 2022.
- [15] C. Arman, B. Iz, K. Daniel, D. Bhavana, and W. Dan. Pretrained language models for sequential sentence classification. In Proc. of EMNLP2019 and 9th IJCNLP, 2019.
- [16] V. Ashish, S. Noam, P. Niki, U. Jakob, J. Llion, G. Aidan N, K. Lukasz, and P. Illia. Attention is all you need. In NeurIPS, 2017.
- [17] S. Ravfogel, Y. Goldberg, and T. Linzen. Studying the inductive biases of rnns with synthetic variations of natural languages, 2019.
- [18] R. Marvin and T. Linzen. Targeted syntactic evaluation of language models, 2018.
- [19] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In International Workshop on Similarity-Based Pattern Recognition, 2014.
- [20] R. Greene, T. Sanders, L. Weng, and A. Nee-lakantan. New and improved embedding model, 2022. <https://openai.com/blog/new-and-improved-embedding-model>, Accessed: 15 December, 2023.