科学論文中の同じ貢献を説明しているイントロダクションの文と本文のパラグラフを判定するためのデータセット

田中翔平 [1] 牛久祥孝 [1]
[1] オムロンサイニックエックス株式会社
{shohei.tanaka, yoshitaka.ushiku}@sinicx.com

概要

本研究では論文中の同じ貢献を説明しているイントロダクションの文と本文のパラグラフを判定するという新しいタスクを提案する．よく書かれた科学論文はイントロダクション中で示された貢献について，以降の本文において詳細に説明している．しかし初学者や専門外の論文を読む研究者にとって，イントロダクション中で示された貢献と本文の対応を把握することは難しい．機械学習モデルがこうしたイントロダクションと本文の対応をハイライトすることで，人間の論文読解の助けになることが期待される．提案するタスクのために，本研究では自動アノテーションと人手アノテーションを組み合わせてデータセットを収集した．収集したデータセットを用いて構築したエンコーダーモデルをパラグラフ検索タスクにおいて評価したところ，科学ドメインデータで事前学習されたモデルが最も性能が高く，38%の精度でイントロダクション中の貢献を説明する文に対応する本文のパラグラフを選択できることがわかった．

1 はじめに

論文読解は機械学習モデルによって論文の内容を読み解くタスクである．従来の論文読解の研究は論文の構造理解 [1]，論文の内容についての質問応答 [2]，論文に含まれるエンティティの関係抽出 [3, 4]，論文中の問題-解決方法認識 [5] といったタスクを扱ってきた．

本研究では論文中の同じ貢献を説明しているイントロダクションの文と本文のパラグラフを判定するという新しいタスクを提案する．ここで論文の貢献を説明している文とは新しい手法の提案や新しい実験結果の発見など，その論文の新規性や有効性を表す文のことを意味する．また本文のパラグラフはイントロダクション以降のすべてのパラグラフを意味する．従来の論文の貢献に着目したタスク [6] は貢献を表す文の抽出や貢献を説明している文に基づく研究概要の知識グラフの構築を取り扱ってきた．これに対して本研究で提案するタスクは論文の中で同じ貢献を表す文とパラグラフを結びつけることをねらう．よく書かれた科学論文はデータセットの収集，タスクの定義，提案手法や実験結果といったそのイントロダクションで述べられた貢献の具体的な内容がそれ以降の本文において正確かつ詳細に説明されている．しかしよく書かれた論文であっても，経験が浅い学生や専門外の論文を読む研究者にとって，その論文のイントロダクションに書かれた貢献と本文との対応を正確に把握することは難しい．もし機械学習モデルが論文のイントロダクション中で貢献を説明している文と本文との対応をハイライトすることができれば，こうした初学者の論文読解の大きな助けになると考えられる．図 1 に本研究で提案するタスクの概要を示す．ここでイントロダクション中の貢献を説明している文 “4. For small to medium training data, changing only these parameters reaches the same task accuracy as full fine-tuning, and sometimes even improves results.” に対しては 3 ページ目の “On validation set, BitFit outperforms...” というパラグラフがこの貢献を説明している文に対応するパラグラフとなる [7]．よってこの例の場合，機械学習モデルは “4. For small...” というイントロダクションの文と “On valdation set...” という本文のパラグラフが同じ貢献であると判定する必要がある．

このタスクのためのデータセットを構築するためには，論文のテキストデータに対して (1) イントロダクション中の貢献を説明している文の選択 (2) イントロダクション中の貢献を説明している文に対応するパラグラフの選択，という二段階のアノテー

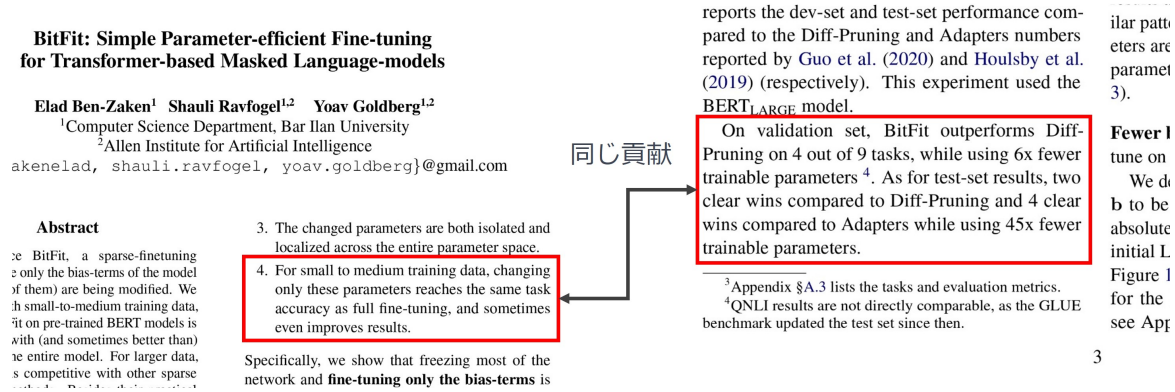

図 1 Ben-Zaken ら [7] の論文における同じ貢献を説明しているイントロダクションの文と本文のパラグラフの例

ションが必要となる．しかし論文読解は高い専門性を必要とするタスクであり，両方のアノテーションを人手で行うことは大きなコストがかかる．そこで本研究ではイントロダクション中の貢献を説明している文の選択は GPT-4 [8] を用いて自動で行うことでアノテーターの負担を約 70%軽減した．111 本の自然言語処理分野の論文についてアノテーションを行い，717 本のイントロダクションの貢献を説明している文と本文のパラグラフのペアを収集した．

本論文では，イントロダクションの貢献を説明している文をクエリ，本文のパラグラフをキーとみなし，それぞれを embedding に変換してパラグラフ検索を行うタスクとしてエンコーダーモデルを学習させた．エンコーダーモデルとしては BERT [9] ベースのモデルを用い，モデルの学習にはイントロダクションの貢献を説明している文と対応する本文のパラグラフのペアを正例，イントロダクションの文と対応していない本文のパラグラフのペアを負例とみなす supervised contrastive learning [10] を用いた．様々な BERT ベースモデルを比較した実験結果より，科学ドメインデータで事前学習を行ったモデル [11] が最も高い精度で正しいパラグラフを選択できることが判明した．

2 データセット収集

本章では提案するタスクのためのデータセットを収集する方法，および収集したデータセットの統計情報について述べる．

まず ACL 2023 [12] のメインカンファレンスの論文 PDF をダウンロードし，GROBID [1] を用いて論文 PDF を論文の章やパラグラフの構造情報も含んだ XML データへと変換した．この XML データを用いて本研究が提案するタスクのためのアノテーション付きデータセットを構築するためには，(1) イントロダクション中の貢献を説明している文の選択 (2) イントロダクションの貢献を説明している文に対応する本文のパラグラフの選択，という二段階のアノテーションが必要となる．しかし科学論文を読解しアノテーションを行うことは高い専門性を必要とするタスクであり，両方のアノテーションを人手で行うことはコストが大きい．そこで本研究ではイントロダクション中の貢献を説明している文の選択に GPT-4 [8] を用いて自動アノテーションを行った．そして自動アノテーションで貢献を説明している文だと判定された文に対応する本文のパラグラフを人手で選択した．本研究で収集したデータセットでは，イントロダクション中の貢献を説明している文と対応する本文のパラグラフのペアが一サンプルとして扱われる．論文は複数の貢献を含むこともあるため，同じ論文に対して複数のイントロダクション中の貢献を説明している文が存在することもある．なお，イントロダクション中の貢献を説明している文は自動アノテーションで選択されているため，実際には貢献を説明していない文が選択されている可能性がある．また論文 PDF も自動で XML データに変換されているため，本文のパラグラフが正しく抽出されていない可能性がある．よってイントロダクションの貢献を説明している文に対応する本文のパラグラフを選択する際には，こうした自動アノテーションの結果が正しいかも同時に判定する必要がある．

このアノテーションのために自然言語処理分野の研究室に所属する 4 人の大学院生を雇用した．アノテーターは以下の作業手順に従ってアノテーションを行った．

Step 1:論文 PDF のイントロダクションを読む

Step 2:アノテーション画面に表示されている，論文の貢献を説明しているイントロダクション中の文を読む

Step 3:表示されている文がイントロダクション中の貢献を説明している文かを判定する

(1)イントロダクション中の貢献を説明している文でない場合は次のアノテーションサンプルへと移行する

(2)イントロダクション中の貢献を説明している文の場合は Step 4 へ進む

Step 4:イントロダクション以降の本文をどこに何が書いてあるかわかる程度にざっと読む

Step 5:イントロダクション中の貢献を説明している文に対応する本文のパラグラフがアノテーション画面に表示されているかを判定する

(1)本文のパラグラフが表示されていない場合は次のアノテーションサンプルへと移行する

(2)本文のパラグラフが表示されている場合は Step 6 へ進む

Step 6:イントロダクション中の貢献を説明している文に対応する本文のパラグラフを選択する

ここで Step 3, 5 は GPT-4 による自動アノテーションと GROBID の構造解析の結果に対する判定である．また Step 6 において，イントロダクション中の貢献を説明している文に対応する本文のパラグラフが複数存在する場合，アノテーターは最初に対応するパラグラフを選択する．

収集したデータセットの統計情報を表 1 に示す．本研究では 111 本の論文に含まれる，GPT-4 によってイントロダクション中の貢献を説明している文だと判定された 1,006 文に対してアノテーションを行った．111 本の論文に含まれるイントロダクション中のすべての文の数は 2,996 文である．よってイントロダクション中の貢献を説明している文の選択について，アノテーターの負担は約 70% (1-1,006/2,996) 軽減されたとみなすことができる．111 本中 7 本の論文については agreement を計算するためにアノテーター全員でアノテーションを行った．7 本の論文に対するアノテーションの Free-marginal multirater kappa [13] は 0.43 であり，これは moderate agreement である．GPT-4 によってイントロダクション中の貢献を説明している文だと判定された文のうち，アノテーターによって実際に貢献を説明している文だと判定された文は 782 文 (77.73%) であった．すなわち GPT-4 によるイントロダクション中の貢献を説明している文の選択の精度は 77.73% であった．また貢献を説明している文に対応する本文のパラグラフが表示されたサンプルは 720 本 (92.07%) であった．残された 720 本のうち，アノテーターの過半数がイントロダクション中の貢献を説明している文が含まれていない，または対応する本文のパラグラフが表示されていないと判定した 3 本のサンプルを取り除いた．結果的に 717 本のイントロダクション中の貢献を説明している文と対応する本文のパラグラフのペアが得られた．

論文数	111
アノテーションサンプル数	1,006
Free-marginal multirater kappa	0.43
貢献を説明しているイントロダクションの文が表示されていた数	782/1006 (77.73%)
対応するパラグラフが画面に表示されていた数	720/782 (92.07%)
得られたサンプル数	717

表 1 収集したデータセットの統計情報

図 2 にイントロダクション中の貢献を説明している文と対応する本文のパラグラフの位置についてのヒートマップを示す．イントロダクション中の文の数や本文のパラグラフの数は論文によって異なるため，それぞれの位置は 5 パーセント刻みの百分率で表現されている．まずイントロダクション中の貢献を説明している文の位置に着目すると，貢献を説明している文はイントロダクションの後半に多く出現していることがわかる．これはイントロダクションの前半では先行研究やその論文が取り扱う問題の背景を説明することが多いからだと考えられる．またイントロダクション中の貢献を説明している文に対応する本文のパラグラフは本文全体の前方 80%以内に含まれることが多いことがわかる．これは後方 20%には conclusion や appendix といった本文ですでに述べられた貢献に再度言及している章や補足情報を説明する章が含まれるからだと考えられる．

3 エンコーダーモデル

本研究では提案するタスクを解くために，イントロダクション中の貢献を説明する文をクエリ，本文のパラグラフをキーとみなし paragprah retrieval を行うエンコーダーモデルを学習させる．エンコーダーモデルを学習させる方法として，本研究では supervised contrastive learning [10] を用いた．Contrastive learning [14] は同じ意味を持つ embedding のペアである正例よりも違う意味を持つ embedding

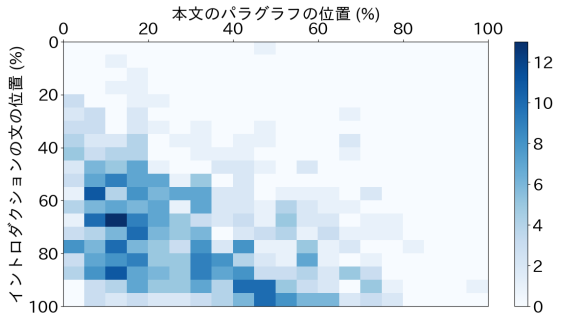

図 2　イントロダクションの文と本文のパラグラフの位置のヒートマップ

のペアである負例のスコアが低くなるようにモデルを学習させる．提案するタスクにおいて，正例はイントロダクション中の貢献を説明する文と対応する本文のパラグラフのペアである．また負例はイントロダクション中の貢献を説明する文と対応しない本文のパラグラフのペアである．本研究における supervised contrastive learning の詳細を付録 A に示す．

4　実験

エンコーダーモデルとして BERT [9], SciBERT [11], RoBERTa [15], ALBERT [16], DistilBERT [16], DeBERTa [17] といった BERT の派生形モデルを用いた．エンコーダーモデルが出力する embedding として，最終層の hidden states の平均ベクトルを用いた．

評価指標として，hit@k (k = 1 or 5) を用いた．hit@k は，イントロダクション中の貢献を説明する文に対応する本文のパラグラフがエンコーダーモデルが貢献を説明する文と類似度が高いと判定した上位 k 位以内のパラグラフに含まれている割合である．hit@1 は precision に等しい．アノテーションした 111 本の論文のうちランダムに選択した 10 本ずつを検証データ，テストデータとして用い，残りの 91 本の論文を学習データとして用いた．実験に用いる各モデルのパラメータは，検証データにおける損失関数の値が最小のものを使用した．実験はモデルごとに 10 回試行した．

各エンコーダーモデルの評価結果を表 2 に示す．SciBERT が hit@1 (precision), hit@5 ともに最も高い性能を示していることがわかる．これは SciBERT が科学ドメインのデータという，本研究で取り扱う論文データに最も近いデータで事前学習されているためだと考えられる．

モデル	hit@1 (%)	hit@5 (%)
BERT	25.93(±1.53)	56.10(±1.77)
SciBERT	**38.47**(±**0.78**)	**70.85**(±**1.27**)
RoBERTa	25.08(±2.25)	62.37(±1.83)
ALBERT	26.44(±1.73)	46.61(±2.04)
DistilBERT	24.07(±1.83)	55.93(±2.27)
DeBERTa	5.42(±1.27)	23.05(±1.89)

表 2　10 回の試行における平均性能．**太字**と下線はそれぞれ 1 位，2 位のスコアを示す．

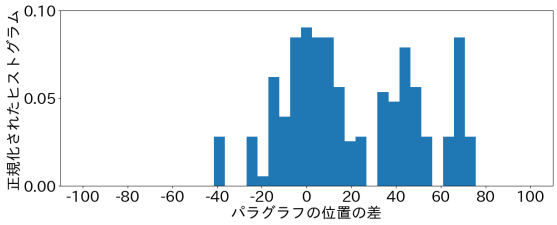

図 3　正しいパラグラフと誤って選択されたパラグラフの位置の差

図 3 に SciBERT が誤って選択したパラグラフと正解のパラグラフの位置の差の頻度をヒストグラムで示す．図 2 と同じく位置は論文における 5 パーセント刻みの百分率で表現されている．またヒストグラムは合計が 1 になるように正規化されている．誤って選択したパラグラフの方が後ろにある場合は位置の差は正となり，前にある場合は負となる．図を見ると，正解のパラグラフよりも後ろに位置するパラグラフを誤って選択する傾向があることがわかる．これは提案するタスクがイントロダクション中の貢献を説明する文に対応する本文の最初のパラグラフを選択するタスクであり，そのパラグラフ以降でより意味的に近い具体的な内容を説明している場合があるためだと考えられる．

5　おわりに

本研究では論文中の同じ貢献を説明しているイントロダクションの文と本文のパラグラフを判定するという新しいタスクを提案した．このタスクを解くモデルを実装するために，自然言語処理分野の論文 111 本を含むデータセットを構築した．このデータセット収集の際には，GPT-4 による自動アノテーションを活用することでアノテーターの負担を約 70%軽減した．収集したデータセットを用いて，イントロダクション中の貢献を説明する文をクエリ，本文のパラグラフをキーとみなすエンコーダーモデルを構築した．実験結果より，科学ドメインデータで事前学習を行った SciBERT が最も高い性能を示すことがわかった．

謝辞

本研究は，JST【ムーンショット型研究開発事業】【JPMJMS2236】の支援を受けたものです.

参考文献

[1] Grobid. https://github.com/kermitt2/grobid, 2008–2023.

[2] Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. A dataset of argumentative dialogues on scientific papers. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7684–7699, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, pp. 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[4] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In **Proceedings of the 12th International Workshop on Semantic Evaluation**, pp. 679–688, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Kevin Heffernan and Simone Teufel. Problem-solving recognition in scientific text. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6045–6058, Marseille, France, June 2022. European Language Resources Association.

[6] Jennifer D'Souza, Sören Auer, and Ted Pedersen. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In **Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)**, pp. 364–376, Online, August 2021. Association for Computational Linguistics.

[7] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] OpenAI. Gpt-4 technical report, 2023.

[9] 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.

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

[11] 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.

[12] Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors. **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] Justus Randolph. Free-marginal multirater kappa (multirater κ free): An alternative to fleiss fixed-marginal multirater kappa. 第 4 巻, 01 2010.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. **arXiv preprint arXiv:2002.05709**, 2020.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[17] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

A　学習方法の詳細

論文 $D_i(1 \leq i \leq N_D)$ のイントロダクション中の貢献を説明する文を $q_j(1 \leq j \leq N_{Qi})$, 本文のパラグラフを $p_k(1 \leq k \leq N_{Pi})$ と定義する．ここで N_D, N_{Qi}, N_{Pi} はデータセットに含まれる論文の数，論文 D_i に含まれるイントロダクション中の貢献を説明する文の数，論文 D_i に含まれる本文のパラグラフの数である．またイントロダクション中の貢献を説明する文 q_j に対応する本文のパラグラフのインデックスを k_+ として定義し，対応しないパラグラフのインデックスを $k_-(1 \leq k \leq N_{Pi}, k_- \neq k_+)$ と定義する．エンコーダーモデルはイントロダクション中の貢献を説明する一文 q_j をクエリとして与えられたとき，これを embedding $\mathrm{q_j}$ へと変換する．また本文のすべてのパラグラフ $p_k(1 \leq k \leq N_{Pi})$ もそれぞれ $\mathrm{p_k}$ へと変換する．そして $\mathrm{q_j}$ と最もコサイン類似度が高い embedding $\mathrm{q_k}$ を持つパラグラフがイントロダクション中の貢献を説明する文に対応する本文のパラグラフとして選択される．$k = k_+$ であれば，貢献を説明している文に対応する正しいパラグラフが選択されたとみなされる．

エンコーダーモデルを学習させる方法として，本研究では supervised contrastive learning [10] を用いる．Contrastive learning [14] はエンコーダーモデルの学習において広く用いられている手法であり，同じ意味を持つ embedding のペアよりも違う意味を持つ embedding のペアのスコアが低くなるようにモデルを学習させる．同じ意味を持つ embedding のペアは正例, 違う意味を持つ embedding のペアは負例として定義される．提案するタスクにおいて，正例はイントロダクション中の貢献を説明する文 q_j と対応する本文のパラグラフ p_{k_+} のペアである．また負例はイントロダクション中の貢献を説明する文 q_j と対応しない本文のパラグラフ p_{k_-} のペアである．ペアのスコアとしては文とパラグラフを表す embedding のコサイン類似度を用いる．これらの定義を用いて，本研究における supervised contrastive learning の loss を下記の通り定義する．

$$Loss = \max\left(0, -sim(\mathrm{q_j}, \mathrm{p_{k_+}}) + sim(\mathrm{q_j}, \mathrm{p_{k_-}}) + m\right), \quad (1)$$

ここで sim はコサイン類似度であり，m はマージンである．本研究では $m = 1$ とした．

B　実験設定の詳細

モデルの実装には PyTorch, HuggingFace, DeepSpeed を用いた．モデルの学習には V100 16GB x 4 枚を用いた．モデルの最適化には Adam を用い，学習率は 1e – 5 とした．

C　倫理的配慮

データセットを構築するために使用した論文はすべて ACL Anthology[1)] 上で Creative Commons Attribution 4.0 International License を付与されたものを使用している．本研究で構築したデータセットにはアノテーターの個人情報は一切含まれていない．

D　論文の限界

本研究では自然言語処理分野の 111 本の論文を対象として，イントロダクション中の貢献を説明している文と対応する本文のパラグラフのペアを含むデータセットを構築した．しかし本研究で構築したデータセットは少数データセットであり，それに伴ってモデルの精度も最大 38%程度に留まっている．大規模データセットの収集や少数データセットにおけるモデル学習方法の改善など，精度を向上させるための方法を模索する必要がある．また GPT-4 による自動アノテーションではイントロダクション中の貢献を説明している文を見落としている可能性がある．10 本の論文を用いた予備実験では GPT-4 によるイントロダクション中の貢献を説明している文の選択の再現率は 73.32% であったが，本研究ではアノテーターの負担を軽減することを優先した．また科学論文には有機化学やロボット工学など様々な研究ドメインがあり，本研究で使用したデータセット構築方法やモデルがほかのドメインでも有効であるかは不明である．今後は本研究で構築したデータセットを自然言語処理分野以外の論文も含めたデータセットへと拡張していき，様々なドメインにおけるモデルの性能を評価していく必要がある．

1)　https://aclanthology.org/