Intermediate-Task Transfer Learning for Peer Review Score Prediction

Panitan Muangkammuen[1], Fumiyo Fukumoto[2], Jiyi Li[2], and Yoshimi Suzuki[2]
[1]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences
[2]Interdisciplinary Graduate School
University of Yamanashi, Kofu, Japan
{g21dts04,fukumoto,jyli,ysuzuki}@yamanashi.ac.jp

Abstract

Peer review is a fundamental component of the academic publishing process, ensuring the quality and validity of research findings. However, predicting peer-review aspect scores accurately can be challenging due to the small size of publically available datasets on the target aspect of scores. To address this issue, we propose an intermediate-task transfer learning method to further improve the performance of pre-trained models. The method assumes an intermediate task that is related to the target task to learn beneficial features before fine-tuning it on a target task. Our experiments demonstrate that intermediate-task transfer learning helps improve the performance of the pre-trained model on peer review score prediction. Our code is available at https://github.com/panitan-m/peerreview-intermediate-trans.

1 Introduction

In recent years, there has been a surge volume of submissions to AI-related international conferences and journals. This upsurge has consequently intensified the difficulties of the review process. To alleviate the burgeoning reviewers' workload, employing an approach to reject papers with evidently low quality serves as a practical strategy. On the other hand, constructive critique extended to authors about the shortcomings in their submissions can encourage refinement and enhancement of their work. In response to this challenge, the development of automatic Peer Review Score Prediction systems has emerged. These systems score a numerical evaluation of academic papers, assessing a spectrum of aspects like "*clarity*" and "*originality*".

A pioneering contribution to the field comes in the form of the PeerRead dataset. This publicly accessible corpus of scientific peer reviews, introduced by Kang et al. (2018), serves as a valuable resource for researchers with diverse objectives. These objectives are ranging from classification of paper acceptance (Ghosal et al., 2019; Deng et al., 2020; Maillette de Buy Wenniger et al., 2020; Fytas et al., 2021), prediction of review aspect scores (Li et al., 2020; Wang et al., 2020; Muangkammuen et al., 2022), to citation recommendation (Jeong et al., 2019), and predicting citation counts (van Dongen et al., 2020). In this paper, we focus on review aspect score prediction.

Unsupervised pre-training SciBert (Beltagy et al., 2019) was utilized on various downstream scientific NLP tasks, including biomedical domain (Li et al., 2016; Nye et al., 2018), computer science domain (Luan et al., 2018; Jurgens et al., 2018), and multiple domains (Cohan et al., 2019). One promising approach for further enhancing pre-trained models that have been shown to be broadly helpful is to first fine-tune a pre-trained model on an intermediate task, before fine-tuning again on the target task, also referred to as *Supplementary Training on Intermediate Labeled-data Tasks* (STILTs) (Phang et al., 2019; Pruksachatkun et al., 2020). STILTs explore the potential of incorporating a secondary phase of pre-training using data-rich intermediate supervised tasks, with the aim of improving the effectiveness of the resulting target task model. In this work, we perform comprehensive experiments using the Aspect-enhanced Peer Review (ASAP-Review) dataset (Yuan et al., 2022) that we extract review aspect sentiments for our intermediate task training. The ASAP-Review dataset is a collection of peer-reviews with fine-grained annotations of review aspect information. For example, *"The paper is well-written and easy to follow"* shows a positive sentiment of *clarity* aspect and a high score of clarity aspect. These aspect sentiments can be beneficial for the review aspect score prediction. We extract the review aspect sentiment from the review texts of a paper and use it as a target label for that given paper. We ran our experiments on 6 intermediate tasks and 7 target tasks, resulting in a total of 42 intermediate-target task pairs.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 40–47
November 1–4, 2023. ©2023 Association for Computational Linguistics

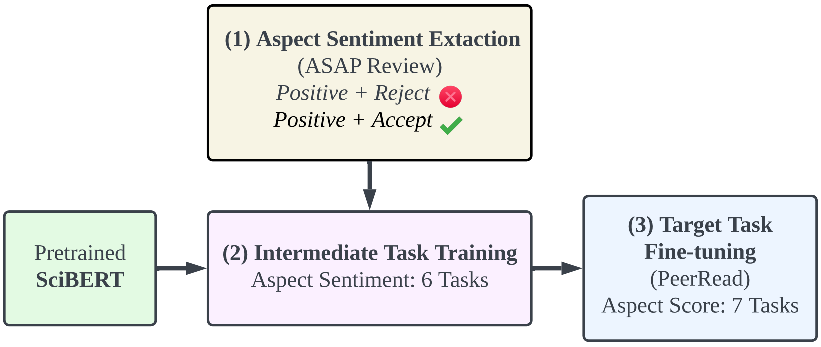


Figure 1: Overview of our pipeline framework. It comprises aspect sentiment extraction, intermediate-task training, and fine-tuning on the target task.

In summary, our main contributions are:

- This work is the first to introduce an intermediate-task transfer learning method to peer-review score prediction.
- We propose a method to extract aspect sentiments for intermediate-task training for peer-review score prediction.
- We conduct experiments to demonstrate the efficacy of each intermediate task, resulting in performance gains across every review aspect score prediction.

2 Related Work

Artificial Intelligence is a crucial tool for academic peer review, and it is a rapidly growing field that demands more attention from the academic community. The renowned Toronto Paper Matching system, developed by Charlin and Zemel (2013), was designed to match papers with appropriate reviewers. Notably, Price and Flach (2017) conducted an in-depth examination of the diverse methods for harnessing computational support in the peer review system. Mrowinski et al. (2017) explored the application of evolutionary algorithms to enhance editorial strategies within the peer review process. Ghosal et al. (2018a,b) delved into an investigation of the impact of various features in the editorial pre-screening process. Wang and Wan (2018) explored a multi-instance learning framework for conducting sentiment analysis on peer review texts. Ghosal et al. (2019) investigated the impact of reviewer sentiment expressed in peer review texts on the outcome of the review process. Li et al. (2020) proposed a multi-task learning approach that automatically selects shared structures and auxiliary resources for peer review prediction. More recently, Muangkammuen et al. (2022) explored a semi-supervised learning for improving peer review score prediction.

Our investigations are currently centered on a portion of the PeerRead dataset that has been made available to the public (Kang et al., 2018). Our approach achieves performance improvement on the peer review aspect score prediction task compared to Kang et al. (2018). We attribute this to the use of intermediate task training and the extraction of aspect sentiment in our approach.

3 Methods

We present a simple intermediate-task transfer learning for peer review score prediction. Figure 1 illustrates the method pipeline that consists of the following steps: *aspect sentiment extraction*, *intermediate-task training*, and *fine-tuning on the target task*.

3.1 Aspect Sentiment Extraction

To further train the pre-trained model SciBERT on the intermediate tasks, we extract aspect sentiments from the ASAP-Review dataset (Yuan et al., 2022) to utilize them for our intermediate-task training. The ASAP-Review dataset comprises peer-review data from ICLR and NeurIPS. We use only ICLR data as it contains both accepted and rejected papers which are the same as the target task dataset, PeerRead.

Originally, this dataset contained review texts with sequence labels of fine-grained annotation of aspect information. An example of the review annotations is shown in Table 1. We utilize 6 aspects in the dataset, which are Clarity (CLA-*i*), Meaningful Comparison (COM-*i*), Motivation/Impact (MOT-*i*), Originality (ORI-*i*), Soundness/Correctness (SOU-

🟥 Summary	🟨 Soundness +	🟩 Motivation +	🟦 Clarity +
The authors prove a generalization guarantee for deep neural networks with ReLU activations, in terms of margins of the classifications and norms of the weight matrices. They compare this bound with a similar recent bound proved by Bartlett, et al. While strictly speaking, the bounds are incomparable in strength, the authors of the submission make a convincing case that their new bound makes stronger guarantees under some interesting conditions. The analysis is elegant. It uses some existing tools but brings them to bear in an important new context, with substantive new ideas needed. The mathematical writing is excellent. Very nice paper. I guess that networks including convolutional layers are covered by their analysis. It feels to me that these tend to be sparse, but that their analysis still my provides some additional leverage for such layers. Some explicit discussion of convolutional layers may be helpful.			

Table 1: An example of review annotations of ASAP-Review dataset. “+" denotes positive sentiment. Negative sentiment does not occur in this example.

Aspects	**Negative**	**Positive**	**Total**
CLA-*i*	1,560	1,003	2,563
COM-*i*	1,738	180	1,918
MOT-*i*	525	1,453	1,978
ORI-*i*	1,257	1,186	2,443
SOU-*i*	1,789	933	2,722
SUB-*i*	1,726	505	2,231

Table 2: Statistics of the aspect sentiments of ASAP-Review dataset for the intermediate-task training.

Aspects	**Total**
Clarity (CLA)	136
Meaningful Comparison (COM)	132
Impact (IMP)	132
Originality (ORI)	136
Soundness/Correctness (SOU)	136
Substance (SUB)	136
Overall Recommendation (REC)	136

Table 3: Statistics of the PeerRead ACL 2017 dataset for the target tasks.

i), and Substance (SUB-*i*). Each aspect is also marked with a sentiment, *positive* or *negative*. We count the number of positives and negatives of each aspect in the reviews. We use the majority polarity as a label for the reviewed paper since one paper consists of multiple reviews. We further remove the samples having a positive aspect label with a reject decision and having a negative aspect label with an accept decision to amplify the characteristic in the data. The statistics of the ASAP-Review dataset after aspect sentiment extraction are shown in Table 2. To distinguish it from the target tasks, i.e., review aspect score predictions, we add "-*i*" to each intermediate task.

3.2 Intermediate Task Training

We fine-tune SciBert model on each intermediate task, following the standard procedure of fine-tuning a pre-trained model on a target task as described in Devlin et al. (2019). Instead of multi-task training (Liu et al., 2019), we use single intermediate-task training to examine the effect of each intermediate task independently. The objective of these intermediate tasks is to predict the sentiment for each review aspect. We train the model to minimize the *Binary Cross-Entropy* loss.

3.3 Target Task Fine-tuning

After intermediate-task training, we fine-tune our models on each target task individually. Our target task is peer-review score prediction, which consists of 7 aspects shown in Table 3. The PeerRead dataset contains peer-review datasets from several conferences. Among them, we chose the ACL 2017 dataset for our experiment as it includes aspect scores that are fully annotated. In this dataset, an input paper has multiple review scores, we use the rounded average score of each aspect as the target score ranging from 1 to 5. We fine-tune the models to minimize the *Categorical Cross-Entropy* loss of five classes.

4 Experiments

4.1 Experimental settings

We used the pre-trained model `scibert-scivocab-uncased` in all experiments. For each intermediate and target task, we used a peak learning rate at 5×10^{-5} and a dropout rate of 0.1. We used a batch size of 8 and a maximum se-

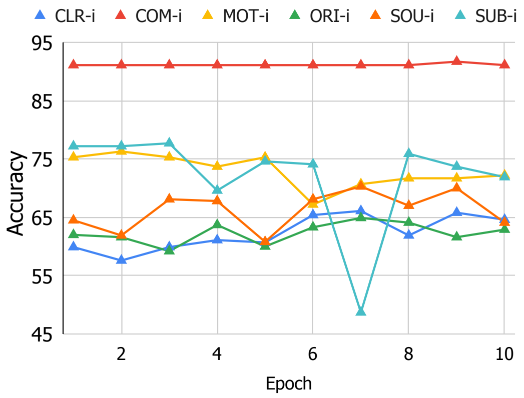


Figure 2: Performances on intermediate tasks in accuracy at each checkpoint.

Aspects	PeerRead	Ours
CLA	67.4 (22.5)	**69.3 (27.4)**
COM	55.0 (20.4)	**62.1 (33.9)**
IMP	80.2 (30.3)	**82.0 (37.2)**
ORI	47.8 (21.5)	**56.9 (50.7)**
SOU	50.2 (21.6)	**60.5 (41.9)**
SUB	67.1 (21.1)	**68.6 (31.2)**
REC	58.8 (23.5)	**64.0 (36.4)**
Avg.	60.9 (23.0)	**66.2 (37.0)**

Table 4: Results compared with the method in PeerRead (Kang et al., 2018). Each cell indicates accuracy (macro F1). **Bold** indicates the best result.

quence length of 512. We trained our models using the AdamW (Loshchilov and Hutter, 2019) with linear decay and 0.2 warm-up ratio. We performed our experiments on NVIDIA GeForce RTX 3090 GPUs.

A pipeline with one intermediate task works as follows: First, we split the extracted ASAP-Review data into training and validation sets with a 9:1 ratio. We fine-tuned SciBERT on the intermediate task for 10 epochs and saved a checkpoint at the end of each epoch, resulting in 10 checkpoints. The performance of each intermediate task evaluated on the validation set is shown in Figure 2. The performances were quite stable during fine-tuning, except for SUB-*i*. We then fine-tuned copies of the resulting models separately on each of the 7 target tasks. We chose the result of the checkpoint that performs best on the target task. Because the test set of the PeerRead dataset is very small, i.e., only 7 samples, most of the results reported by Wang et al. (2020) can be obtained by just using the majority score as a prediction, and it could lead to inappropriate evaluation. Instead of using the original sets to perform the experiments, we ran the same pipeline on 5-fold cross-validation three times. This gave us 15 observations for each result in our experiments.

We compared our method to the PeerRead (Kang et al., 2018). We re-implemented their model based on CNN and kept the same hyperparameters. GloVe 840B embeddings (Pennington et al., 2014) were utilized as input word representations, without tuning. The outputs from the CNN model are fed into a max pooling layer and the final linear layer. We evaluated their model in our experimental settings.

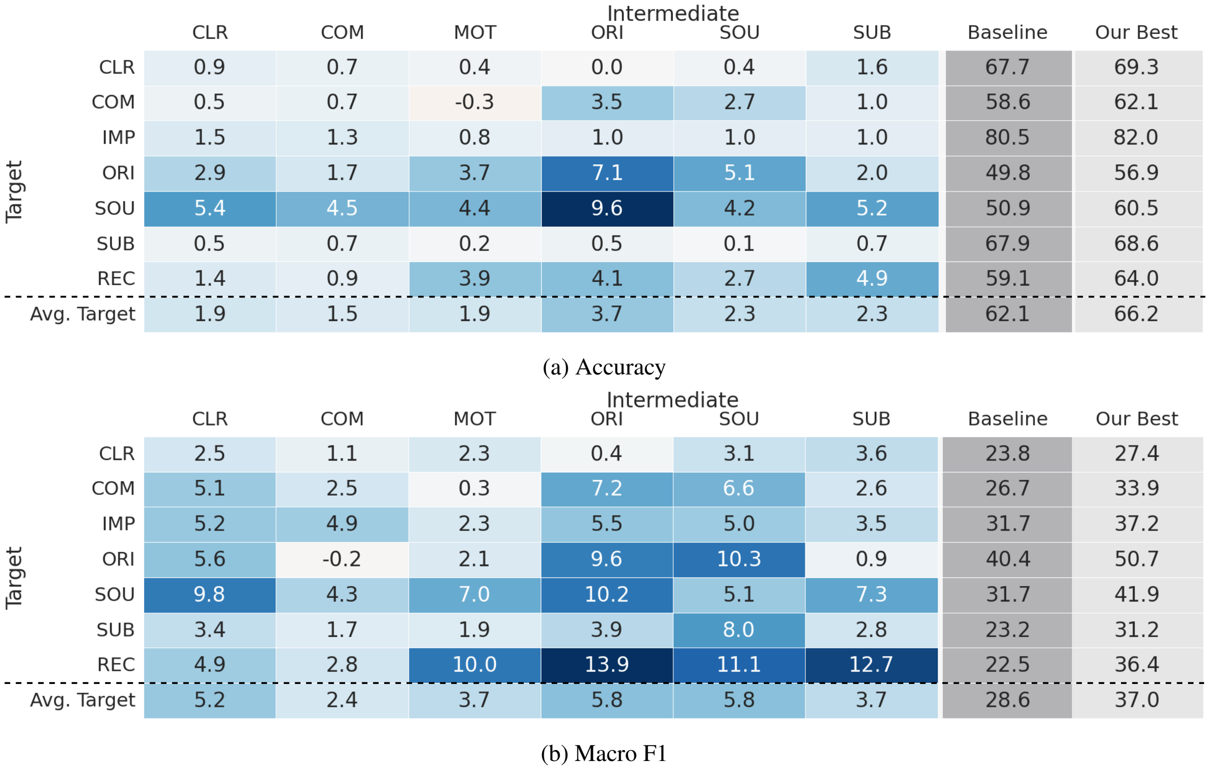
4.2 Results and Discussion

Figure 3 shows the differences in target task performances between the baselines and models trained with intermediate-task training, each averaged across three 5-fold cross-validations. A positive result indicates a successful transfer.

We observed that transfer learning, almost every intermediate-task training, helps improve the performance of the target task. The *Soundness/Correctness* score prediction gains more performance from intermediate-task training with around 10% on both accuracy and macro F1. Overall our best results are better than those of the baselines around 4.1% and 8.4% on average, in accuracy and macro F1, respectively. The best improvements in accuracy are from ORI-*i* on *Soundness/Correctness* at 9.6%. The best improvement in macro F1 score is up to 13.9% from ORI-*i* on *Overall Recommendation*. On average across every target task, the ORI-*i* is the most successful intermediate task that increases 3.7% and 5.8% in accuracy and macro F1, respectively.

Interestingly, we did not find the largest improvement from the same aspect of the intermediate task (sentiment prediction) and the target task (score prediction), except for the *Originality* on the accuracy metric. Instead, the score prediction task gains more performance from other aspects of the intermediate task.

We also compared our method to the PeerRead (Kang et al., 2018) which is shown in Table 4. Our method performed better than the PeerRead model on every task and increased 5.3% and 14% on average, in accuracy and macro F1, respectively. It outperformed the PeerRead model by 10.3% on *Soundness/Correctness* in term of accuracy and by 29.2% on *Originality* in term of macro F1.



(a) Accuracy

(b) Macro F1

Figure 3: Transfer learning results between intermediate and target tasks. Baselines on the second rightmost column are models that are fine-tuned without intermediate-task training. Our best results from the models with intermediate-task training are on the rightmost column. Each cell shows the difference in performance between the baseline and model with intermediate-task training. The cool and warm tone colors indicate improvement, and deterioration, respectively.

4.3 Ablation Study

Our approach to extracting the ASAP-Review dataset for intermediate-task training contains two strategies, i.e., aspect sentiment extraction from review text and removing a sample that has a positive label with a reject decision and vice versa. To examine how each strategy contributes to the performance of the target task, we consider the following variants of our intermediate task:

a) **Decision** - Using decision prediction as an intermediate task. Here, the decision prediction task predicts whether a paper gets *accepted* or *rejected*. The statistics of decision data are shown in Table 5.

b) **Aspect** - Using aspect sentiment data without removing a sample. Here, the sample has a positive label with a reject decision and vice versa. The statistics of the data are shown in Table 6.

c) **Aspect + Decision** - Our full method using two strategies altogether. By incorporating two strategies, the quantity of data is decreased by over 30% from the **Aspect**.

Accept	Reject	Total
3,295	1,855	5,150

Table 5: Statistics of the decision data.

Aspects	Negative	Positive	Total
CLA-*i*	2,430	1,626	4,056
COM-*i*	2,889	264	3,153
MOT-*i*	773	2,655	3,428
ORI-*i*	1,837	1,984	3,821
SOU-*i*	2,700	1,357	4,057
SUB-*i*	2,901	760	3,661

Table 6: Statistics of the aspect polarity data without removing a sample that has a positive label with a reject decision and vice versa.

Table 7 shows the results of different strategies of the intermediate task training. We can see that **Decision** helps improve the pre-trained model performance in almost every target task except *Substance* on macro F1. **Aspect** further improves the pre-trained model compared to **Decision** in almost

Target Task	Baseline	Intermediate Task		
		Decision	Aspects	Aspects + Decision
CLR	66.7 (23.8)	+0.4 (+1.4)	+0.4 (+1.3)	**+1.6 (+3.6)**
COM	58.6 (26.7)	+1.2 (+4.8)	+2.3 (+6.4)	**+3.5 (+7.2)**
IMP	80.5 (31.7)	+1.3 (+5.8)	**+2.0 (+7.7)**	+1.5 (+5.5)
ORI	49.8 (40.4)	+4.2 (+5.1)	+3.0 (+3.7)	**+7.1 (+10.3)**
SOU	50.9 (31.7)	+5.2 (+6.8)	+4.2 (+6.4)	**+9.6 (+10.2)**
SUB	67.9 (23.2)	+0.2 (-0.2)	**+1.4** (+4.3)	+0.7 **(+8.0)**
REC	59.1 (22.5)	+1.9 (+7.1)	+3.2 (+9.2)	**+4.9 (+13.9)**
Avg.	62.1 (28.6)	+2.1 (+4.4)	+2.4 (+5.6)	**+4.1 (+8.4)**

Table 7: Results on the variants of the intermediate task. The baseline column indicates the results without intermediate-task training. The other columns show the difference in performance between the baseline and model with intermediate-task training. Each cell indicates an improvement in accuracy (macro F1 score) compared with the baseline. **Bold** indicates the best result.

every target task and has a better performance on accuracy and macro F1 on average. This indicates that the aspect sentiment data contains richer information for review aspect score prediction compared to the decision data. In contrast, the decision data shows more relevance on the *Originality* and *Soundness/Correctness* score predictions than aspect sentiment data. One possible reason for this is that they are the main aspect of the reviewer's judgment.

As we can see from Table 7 that combining aspect polarity data with a decision strategy leads to a better result on almost every target task and the best result on average in both accuracy and macro F1 score. Although the data size of **Aspect + Decision** is smaller than that of **Aspect**, the average result of **Aspect + Decision** is still better. This shows that the characteristic is more important than the quantity of the data for intermediate-task training.

4.4 Error Analysis

We plot the confusion matrix between truth and model prediction on test data in Figure 4, which shows that the prediction scores of our model tend to be close to the true values. The model tends to be biased to a score of 4, which is the most common score in the dataset. The model was able to classify some papers with a score of 2 or 3 correctly. In contrast, it was unable to correctly classify papers with a score of 1 or 5. However, it still rated papers with a score of 5 higher than a score of 1. The shortage of training samples for scores 1 and 5 (less than 5 samples) complicates its prediction. Incorporating techniques to handle imbalanced datasets is an interesting direction for future work.

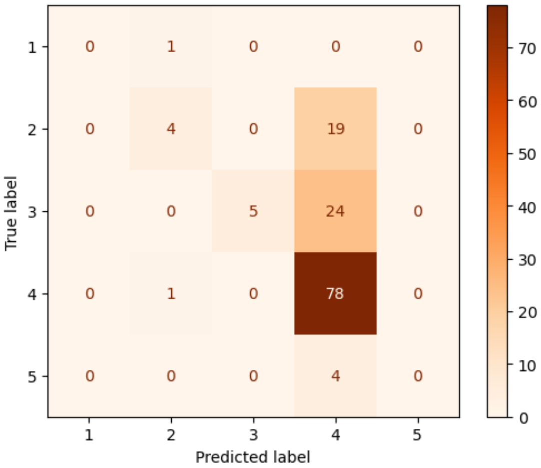


Figure 4: Confusion matrix of true and prediction of *Overall Recommendation* scores.

5 Conclusion

In this study, we investigated the impact of intermediate-task transfer learning on peer-review score prediction. Specifically, we fine-tuned a pre-trained model SciBERT on an intermediate task before fine-tuning again on the target task. We proposed a method to extract the ASAP-Review dataset for intermediate-task training to improve peer-review score prediction. The experimental results showed the effectiveness of the intermediate-task training as it attained a better result than the baseline on every target task in both accuracy and macro F1. Future work will include (1) extending the method to process longer sequences to cover the full length of the paper, and (2) incorporating multiple tasks for the intermediate-task training to exploit related information between intermediate tasks.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions. This work is supported by JST SPRING, Grant Number JPMJSP2133, Kajima Foundation's Support Program, and JSPS KAKENHI (No. 22K12146).

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Laurent Charlin and Richard S. Zemel. 2013. The toronto paper matching system: An automated paper-reviewer assignment system.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhongfen Deng, Hao Peng, Congying Xia, Jianxin Li, Lifang He, and Philip Yu. 2020. Hierarchical bidirectional self-attention networks for paper review rating recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6302–6314, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Panagiotis Fytas, Georgios Rizos, and Lucia Specia. 2021. What makes a scientific paper be accepted for publication? In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 44–60, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tirthankar Ghosal, Ravi Sonam, Sriparna Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2018a. Investigating domain features for scope detection and classification of scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130, Florence, Italy. Association for Computational Linguistics.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018b. Investigating impact features in editorial pre-screening of research papers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 333–334, New York, NY, USA. Association for Computing Machinery.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Park, and Sungchul Choi. 2019. A context-aware citation recommendation model with bert and graph convolutional networks.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. Multi-task peer-review score prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 121–126, Online. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Gideon Maillette de Buy Wenniger, Thomas van Dongen, Eleri Aedmaa, Herbert Teun Kruitbosch, Edwin A. Valentijn, and Lambert Schomaker. 2020. Structure-tags improve text classification for scholarly document quality prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 158–167, Online. Association for Computational Linguistics.

Maciej J. Mrowinski, Piotr Fronczak, Agata Fronczak, Marcel Ausloos, and Olgica Nedic. 2017. Artificial intelligence in peer review: How can evolutionary computation support journal editors? *PLOS ONE*, 12(9):1–11.

Panitan Muangkammuen, Fumiyo Fukumoto, Jiyi Li, and Yoshimi Suzuki. 2022. Exploiting labeled and unlabeled data via transformer fine-tuning for peer-review score prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2233–2240, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

Simon Price and Peter A. Flach. 2017. Computational support for academic peer review: A perspective from artificial intelligence. *Commun. ACM*, 60(3):70–79.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Thomas van Dongen, Gideon Maillette de Buy Wenniger, and Lambert Schomaker. 2020. SChuBERT: Scholarly document chunks with BERT-encoding boost citation count prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 148–157, Online. Association for Computational Linguistics.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 175–184, New York, NY, USA. Association for Computing Machinery.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *J. Artif. Int. Res.*, 75.