Attend to Your Review: A Deep Neural Network to Extract Aspects from Peer Reviews

Rajeev Verma[1], Kartik Shinde[2](✉), Hardik Arora[2], and Tirthankar Ghosal[3]

[1] University of Amsterdam, Amsterdam, Netherlands
rajeev.verma@student.uva.nl
[2] Indian Institute of Technology Patna, Patna, India
{kartik_1901ce16,hardik_1901ce15}@iitp.ac.in
[3] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
ghosal@ufal.mff.cuni.cz

Abstract. Peer-review process is fraught with issues like bias, inconsistencies, arbitrariness, non-committal weak rejects, etc. However, it is anticipated that the peer reviews provide constructive feedback to the authors against some aspects of the paper such as *Motivation/Impact*, *Soundness/Correctness*, *Novelty*, *Substance*, etc. A good review is expected to evaluate a paper under the lens of these aspects. An automated system to extract these implicit aspects from the reviews would help determine the quality/goodness of the peer review. In this work, we propose a deep neural architecture to extract the aspects of the paper on which the reviewer commented in their review. Our automatic aspect-extraction model based on BERT and neural attention mechanism achieves superior performance over the standard baselines. We make our codes, analyses and other matrials available at https://github.com/cruxieu17/aspect-extraction-peer-reviews .

Keywords: Aspect extraction · Peer reviews · Deep neural networks

1 Introduction

Peer review is the central system of scientific research validation. However, several studies highlight the bias [1], inconsistencies [2,3], arbitrariness [4] of the peer-review process, thus, degrading the integrity and trust of the central system of research validation. Area chairs/Editors are responsible for mitigating such issues via assigning expert reviewers and evaluating reviewers' comments to generate informed decisions. However, the exponential rise in paper submissions has put the peer-review system under severe stress in recent years, leading to a dearth of experienced reviewers.

R. Verma and K. Shinde—Equal contribution.

T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1517, pp. 761–768, 2021.
https://doi.org/10.1007/978-3-030-92310-5_88

More recently, the peer-review system has been shown to be suffering from the problem of Collusion Rings and Non-Committal Weak Rejects. Non-Committal Weak rejects are defined as paper rejects with no substantial feedback for the authors to act upon and have conflicting scores compared with the reviews with a good evaluation of the said manuscript. Sometimes the competitive venues have vested interests in controlling their acceptance rates. Coupled with the increasing submission rate, this makes the senior area chairs reject as many papers as possible in a bid to accept only perfect papers. Thus, the Non-Committal Weak Rejects reviews increase the probability that the said paper will eventually be rejected as the low review score deems the paper imperfect.

We believe that an essential step in the direction of re-establishing trust in the peer-review process should be an attempt to establish confidence in the peer-review reports.

As per the rubrics defined in [5], we expect the review to evaluate the work for indicators like novelty, theoretical and empirical soundness of the research methodology, writing, and clarity of the work, impact of the work in a broader academic context, etc. We call these indicators review-level aspects.

In this work, we propose a deep neural network architecture that takes the review text as input and extracts the review-level aspects from it. Our model is based on the simple Attention mechanism that can decompose the review text into aspect-level representations, thus, aiding the extraction. We get superior performance compared with the baselines. Such a system can help check the quality of the review reports, thus maintaining the peer-review integrity of the peer-review system.

2 Related Work

Some studies analyze the reviewing process for major ML conferences like Tran et al. [7], Shah et al. [2]. Researchers have also argued for changes in reviewing practices and proposed changes. Sculley et al. [5] proposed a rubric to hold reviewers to an objective standard for review quality. Rogers and Augenstein et al. [6] identifies challenges and problems in the reviewing system and suggests ways to tackle them. Yuan et al. [8] investigated if the writing of reviews could be automated. They also propose a review dataset with phrase-level annotations for the aspects present and present preliminary results on sentence-level aspect prediction. We use this dataset in our experiments. Our objective to judge the review reports is somewhat similar to Xiong and Litman et al. [9] who predict the helpfulness of a peer-review using manual features.

3 Dataset Statistics

We use the ASAP-Review dataset recently propose by [8]. ASAP-Review dataset is composed of review text, their annotated review-level annotations, and review meta-data. It contains ICLR (International Conference on Learning Representations (2017–2020) and NeurIPS (Neural Information Processing Systems

Table 1. Dataset statistics. Along with the standard review text, the dataset also includes as reviews the replies to the reviews and discussion on the OpenReview platform.

Venues	Papers	Acc/Rej	# Reviews
ICLR	5,192	1,859/3,333	15,728
NeurIPS	3,685	3,685/0	12,391
Total	8,877	5,244/3,333	28,119

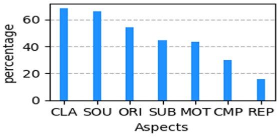


Fig. 1. Proportion of the present aspects in the dataset. As can be seen, the data is unbalanced and aspects like CMP and REP are very scarce.

(2016–2019) reviews from the open-access platform OpenReview[1] platform. We show the detailed dataset statistic in Table 1 and refer the reader to the original paper for more details. We further note that the dataset is unbalanced and show the percentage of each aspect in the whole dataset in Fig. 1.

4 Methodology

A visual schema of the architecture can be seen in Fig. 2. The input to the model is review text. We do not employ any special preprocessing so as to preserve the review structure. We describe the main components of the proposed model in the following subsections.

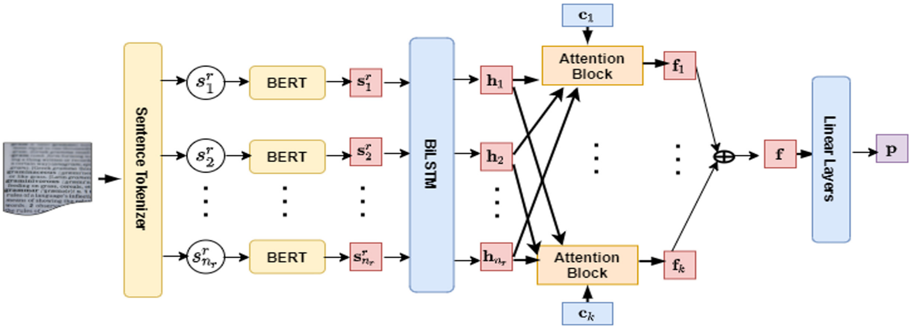


Fig. 2. Proposed Architecture for the review-level aspect extraction. As shown, we use the pretrained BERT model (specifically, RoBERTa) to get sentence representations. We then use BiLSTM to get the review representation. We use the simple attention mechanism using trainable codes $C_a = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$ to decompose the review representation into aspect-level representation. We then concatenate them and pass into feedforward linear layers for final classification.

[1] https://openreview.net/.

4.1 Review Representation and Encoder

The input to our model are review sentences. Denote the review $R = (s_1^r, s_2^r, ..., s_{n_r}^r)$ as the sequence of their respective sentences. For a sentence s_i^p we get a d dimensional embedding vector $\mathbf{s}_i^p \in \mathbb{R}^d$ using RoBERTa [11] model. We specifically use Sentence Transformers [10] with stsb-RoBERTa-base as the encoder. We use the Sentence Embeddings with Sentence Transformers using pretrained model to get meaningful sentence embeddings. Pre-trained language models incorporate rich representation power, as they are trained on large corpora. We get the review representation $\mathbf{R} \in \mathbb{R}^{n_r \times d}$ as $\mathbf{R} = \mathbf{s}_1^r \oplus \mathbf{s}_2^r \oplus ... \oplus \mathbf{s}_{n_r}^r, \quad \mathbf{R} \in \mathbb{R}^{n_r \times d}$ We then pass the review representation $\mathbf{R}$ to the BiLSTM layers [12] $g_{\theta_1}(.)$ to model contextual inter dependencies between individual sentences, and we get output hidden representation $\mathbf{h}_i$ at each step i, i.e. $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n_r}) = g_{\theta_1}(\mathbf{R})$

In this way, our encoding model is a hierarchical module which first models contextual dependencies between tokens to get sentence representations using Transformers based attention mechanism [13], and then uses BiLSTM to model sentence level dependencies.

4.2 Aspect-Level Feature Extractor

To extract aspect level information from the review representation, we use a simple attention mechanism. The goal is to decompose the review into aspect based representations. Given the review level aspects $R_a = \{a_1, a_2, \ldots, a_k\}$, we define codes $C_a = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$ where $|C_a| = |R_a|$ We extract the representation for the k^{th} aspect as follows:

$$\alpha_k^i = \frac{\exp(\mathbf{c}_k \cdot \mathbf{h}_i)}{\sum_{j=1}^{n_r} \exp(\mathbf{c}_k \cdot \mathbf{h}_j)}$$

$$\mathbf{f}_k = \sum_{i=1}^{n_r} \alpha_k^i \mathbf{h}_i$$

Here, α_k^i denote the attention weight for the hidden representation $\mathbf{h}_i$ and for code $\mathbf{c}_k$. And $\mathbf{f}_k$ denote the output representation for the code $\mathbf{c}_k$. The C_a is learned during the training procedure.

4.3 Feedforward Prediction Layers

We concatenate the outputs from the aspect-level feature extractor together for the final classification. Thus, we obtain $\mathbf{f}$ as $\mathbf{f} = \mathbf{f}_1 \oplus \mathbf{f}_2 \oplus \ldots \oplus \mathbf{f}_k$ We pass $\mathbf{f}$ to the feedforward linear layers $l_{\theta_2}(.)$ as get the prediction $\mathbf{p}$ as

$$\mathbf{p} = \frac{1}{1 + \exp(-l_{\theta_2}(\mathbf{f}))}$$

5 Experimental Setup and Evaluation

We split our dataset into 80% train,5% validation and 15% test sets. We use the loss on validation set to monitor the training and employ early stopping when the loss won't improve for 10 epochs. We use a learning rate of $1e-3$, a batch size of 32 to train our main model. We use fixed set $\{1e-1, 1e-2, 1e-3, 3e-3\}$ to tune the learning rate, and find $1e-3$ works best. We used Adam optimizer with a weight_decay=1e-3 (for avoiding overfitting) for training.

5.1 Comparison Systems

Our task of review-level aspect extraction is a multi-label classification problem. We compare our method with common multi-label classification techniques. The comparison systems are described next. To highlight the difference with our method, we follow the same notation from Sect. 4.

1. **Average Sentence Embeddings (ASE)**: We use average of the sentence embeddings as review representation and pass it to the feedforward linear layers for prediction.
2. **BiLSTM**: We use BiLSTM model to do the multi-label classification. More specifically, we do not use our aspect-level feature extractor, and instead get the prediction $\mathbf{p}$ as $\mathbf{h} = \mathbf{h}_1 \oplus \mathbf{h}_2 \oplus \ldots \oplus \mathbf{h}_{n_r}$, $\mathbf{p} = l_{\theta_2}(\mathbf{h})$
3. **BERT**: We finetune the pretrained BERT model [14], specifically bert-base-large using Huggingface [15]. The model takes full review text as input (although, the input gets truncated to 512 tokens as the model has the maximum limit of 512 tokens). The review representation is taken from the [CLS] token, which is then passed to the feedforward linear layers for prediction.

6 Results and Analysis

We show the results for the proposed architecture and the baselines in Table 2. The table shows that our simple attention-based model outperforms other baselines consistently for all aspects. One interesting thing to note is that even the simple baseline like Average Sentence Embeddings (ASE) has decent performance on this task. We attribute this to the use of pretrained RoBERTa embeddings. Since the RoBERTa model is trained on a huge corpus, this model employs all its inductive biases and superior language representation capabilities to give good performance. However, the performance is not consistently good for all aspects. As can be seen, the F1 score on aspect Meaningful Comparison (CMP) is a mere 0.570, while on aspect Replicability (REP), it is abysmally low. The lower performance for these two aspects is due to the imbalance in the dataset. Replicability (REP) and Meaningful Comparison (CMP) are present in a small percentage of the reviews in the dataset.

BiLSTM baseline improves the numbers for all the aspects compared with the ASE baseline. However, the F1 for aspect Replicability (REP) is still low. BiLSTM improves over the ASE baseline as BiLSTM can better model the review

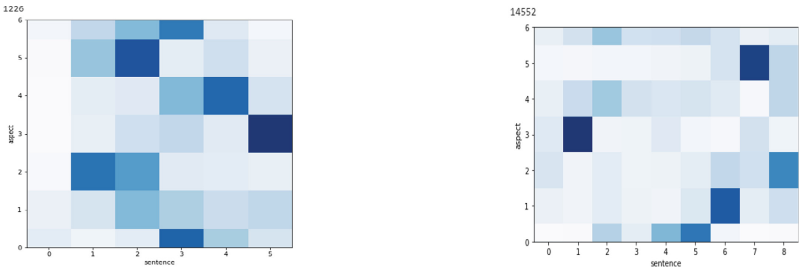
Table 2. Results for the review-level aspect extraction task for each of the aspect corresponding to the proposed model and the baselines. The task is multi-label classification. We report Accuracy(**ACC**) and F1 (**F1**) score.

Aspect	ASE		BiLSTM		BERT		Ours	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Motivation (MOT)	0.712	0.714	**0.835**	0.809	0.72	0.716	**0.835**	**0.827**
Clarity (CLA)	0.878	0.907	0.913	0.935	0.838	0.889	**0.933**	**0.95**
Soundness/Correctness (SOU)	0.749	0.830	0.800	0.860	0.753	0.837	**0.829**	**0.877**
Substance (SUB)	0.720	0.720	0.747	0.749	0.748	0.739	**0.817**	**0.803**
Meaningful Comparison (CMP)	0.752	0.570	0.872	0.794	0.825	0.685	**0.893**	**0.824**
Originality (ORI)	0.808	0.839	**0.876**	**0.885**	0.741	0.787	**0.876**	**0.886**
Replicability (REP)	0.852	0.248	0.875	0.291	0.872	0.422	**0.916**	**0.691**

representation than the less informative average operation of the ASE baseline. Surprisingly, the BERT baseline performs poorer than the BiLSTM and the simple ASE baseline except for the Replicability (REP) aspect. This is contrary to the general knowledge in the NLP community where BERT models are state-of-the-art across many NLP tasks. However, we remark that the ASE and BiLSTM baselines are not independent of the significant performance gains that come with BERT. Both the ASE and BiLSTM model uses pretrained RoBERTa sentence embeddings and then build a review representation. As stated before, the representation obtained is hierarchical with utilizing token-level contextual information in a sentence to get sentence representation and then use sentence-level contextual information to get the review representation. We assert that this hierarchy is missing in the BERT baseline. The BERT model gets the review representation by using only token-level information. However, it models a longer-range context (by self-attention across the complete review), and we get the review representation from the [CLS] token. We hypothesize that the review representation obtained from BERT is not that richer as the reviews can be arbitrarily large. This can affect getting good informative representations.

Finally, we note that our proposed architecture gives good performance even for scarce aspects like Meaningful Comparison (CMP) and Replicability (REP). As stated before, our model is hierarchical in nature, and the aspect-level attention makes it more precise. The attention mechanism is designed to get fine-grained aspect-level representations that can better aid the classification. This lowers false positives and false negatives, thus, improving the F1 score.

We can see the effect of the attention mechanism for the representative review examples - Review 1[2] and Review 2[3] have their attention heatmaps in Fig. 3. We can see that, for each sentence, attention weight shows significant attention weight for some of the aspect $\mathbf{c}_i$. For example, sentence 5 of Review 1, "*furthermore, the paper lacks in novelty aspect, as it uses mostly well-known techniques.*", which clearly has aspect "Originality (ORI)", shows the highest attention weight for aspect 3 ($\mathbf{c}_3$). This means that $\mathbf{c}_3$ is sensitive to the Originality aspect. This conclusion is drawn together with sentence 1 of Review 2, which again has the "Originality(ORI)" aspect and shows the highest attention weight for $\mathbf{c}_3$. The same conclusions can be drawn for other aspects as well. Another thing to note is that the heatmaps are sparse, showing the distinctiveness of the attention mechanism. This means that the aggregate representation $\mathbf{f}_k$ that we get aspect k is representative of the aspect level information. This shows that our model is able to decompose the review representation into aspect-level representations resulting in better prediction.



(a) Review 1

(b) Review 2

Fig. 3. Attention maps for each sentence (on x-axis) and $\mathbf{c}_k$ (k on y-axis) for Review-1:(https://openreview.net/forum?id=HkGJUXb0-&noteId=ry_11ijeM) and Review-2:(https://github.com/cruxieu17/aspect-extraction-peer-reviews/blob/main/review_2.txt) Darker color means high attention weight (α_k^i) for i_{th} review sentence. As can be observed, the attention weight is higher for the sentences having presence of the present aspect.

7 Conclusions

In this work, we propose attention-based deep-neural network architecture for aspect extraction from peer-review. The proposed model takes full review text as input. We also show the efficacy of the attention mechanism by demonstrating its distinctiveness quality. The attention mechanism works by decomposing the review representation into aspect-level representations. One major application of our work would be to detect the Non-Committal Weak rejects reviews which

[2] https://openreview.net/forum?id=HkGJUXb0-&noteId=ry_11ijeM.

[3] https://github.com/cruxieu17/aspect-extraction-peer-reviews/blob/main/review_2.txt.

do not provide actionable feedback. Having such an automated system can help detect the inferior quality reviews to penalize them, thereby alleviating some of the trust issues with the peer-review reports.

References

1. Tomkins, A., Zhang, M., Heavlin, W.D.: Reviewer bias in single-versus double-blind peer review. Proc. Natl. Acad. Sci. **114**(48), 12708–12713 (2017)
2. Shah, N.B., Tabibian, B., Muandet, K., Guyon, I., Von Luxburg, U.: Design and analysis of the NIPS 2016 review process. J. Mach. Learn. Res. **19**(49), 1–34 (2018)
3. Langford, J., Guzdial, M.: The arbitrariness of reviews, and advice for school administrators. Commun. ACM **58**(4), 12–13 (2015)
4. Brezis, E.S., Birukou, A.: Arbitrariness in the peer review process. Scientometrics **123**(1), 393–411 (2020). https://doi.org/10.1007/s11192-020-03348-1
5. Sculley, D., Snoek, J., Wiltschko, A.: Avoiding a tragedy of the commons in the peer review process. arXiv preprint arXiv:1901.06246 (2018)
6. Rogers, A., Augenstein, I.: What can we do to improve peer review in NLP?. arXiv preprint arXiv:2010.03863 (2020)
7. Tran, D., et al.: An open review of OpenReview: a critical analysis of the machine learning conference review process. arXiv preprint arXiv:2010.05137 (2020)
8. Yuan, W., Liu, P., Neubig, G.: Can we automate scientific reviewing?. arXiv preprint arXiv:2102.00176 (2021)
9. Xiong, W., Litman, D.: Automatically predicting peer-review helpfulness. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 502–507, June 2011
10. Reimers, N., Gurevych, I.: Sentence-HERT: sentence embeddings using Siamese HERT-networks. arXiv preprint arXiv:1908.10084 (2019)
11. Liu, Y., et al.: Roberta: a robustly optimized HERT pretraining approach. arXiv preprint arXiv:1907.11692, (2019)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
15. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, October2020