A Deep Neural Architecture for Decision-Aware Meta-Review Generation

Asheesh Kumar*
Department of CSE
IIT Patna, India
aseesnathhh@gmail.com

Tirthankar Ghosal*
ÚFAL, MFF
Charles University, CZ
ghosal@ufal.mff.cuni.cz

Asif Ekbal
Department of CSE
IIT Patna, India
asif@iitp.ac.in

***Abstract*—Automatically generating meta-reviews from peer-reviews is a new and challenging task. Although close, the task is not precisely summarizing the peer-reviews. Usually, a conference chair or a journal editor writes a meta-review after going through the reviews written by the appointed reviewers, rounds of discussions with them, finally arriving at a consensus on the paper's fate. In essence, the meta-review texts are decision-aware, i.e., the meta reviewer already forms the decision before writing the meta-review, and the corresponding text conforms to that decision. We leverage this seed idea and design a deep neural architecture to generate decision-aware meta-reviews in this work. We propose a multi-encoder transformer network for peer-review decision prediction and subsequent meta-review generation. We analyze our output quantitatively and qualitatively and argue that quantitative text summarization metrics are not suitable for evaluating the generated meta-reviews. Our proposed model performs comparably with the recent state-of-the-art text summarization approaches. Qualitative evaluation of our model-generated output is encouraging on an open access peer reviews dataset that we curate from the open review platform. We make our data and codes available[1].**

***Index Terms*—meta-review generation, decision prediction, deep learning**

I. Introduction

Peer review texts are an essential artifact for research validation. It involves some subject experts independently reviewing the submission and then providing their opinion on the paper, whether it should get accepted or rejected (especially in conference reviewing). After this, the area chair arrives at a decision and writes the meta-review considering all the reviews and recommendations, sometimes after rounds of discussions with the appointed reviewers to resolve conflicts. Meta-reviews are means to communicate the decision of the reviewing committee to the authors summarizing the reviewers' opinions, ensuing discussions, and the chairs' views. A meta-review is supposed to be a concise summary of the submission's primary content, an overview of the reviewers' discussions or the peer-reviews, and the decision of acceptance/rejection. Thus, meta-reviews focus only on those points that help justify the final decision. Hence, writing meta-reviews is a crucial task in the peer-review process and is usually done by the senior researchers in the community (journal editors, program chairs, area chairs). However, with the surge in submissions in major conferences, writing a meta-review following the reviews and peer review discussions could be a challenging task in the strict time frame for chairs who do this job voluntarily for the community. Hence an AI assistant to generate meta-reviews would help the meta-reviewer in their decision and craft the final meta-review quickly. Here in this work, we attempt this challenging task of automatically generating the meta-reviews while also predicting the peer review decision in the process. We present a decision-aware transformer-based multi-encoder architecture to generate a meta-review with the final acceptance decision predicted by a self-attentive multi-encoder. Multi-encoder gives three separate representations to the three peer-reviews for further input to the decoder. We also see how the recommendation polarity of the reviewers motivates the meta-reviewer(s) in the writing process. We evaluate our generated meta-reviews, both qualitatively and quantitatively.

To the best of our knowledge, automatic Meta-Review Generation (MRG) is fairly a new task. The only previous work we can find is MetaGen [1], a system for generating assistive meta-reviews from peer reviews. Authors first generate the extractive draft and then use a fine-tuned UniLM [2] (Unified Language Model) for the final decision prediction and generate the abstractive meta-reviews. Although our objectives are similar, we differ in the way we approach the problem. We could not compare our approach with MetaGen due to a lack of data and associated code. Hence, we compare with two state-of-the-art summarization models, PEGASUS [3] and a BART-based text summarizer [4]. We report our initial results on both the tasks, *viz:* decision prediction and meta-review generation.

II. Dataset Description

We collect the required peer review data (reviews, meta-reviews) from the OpenReview[2] platform along with the decision of acceptance/rejection in the top-tier ML conference ICLR for the years 2018, 2019, 2020. After pre-processing and eliminating some unusable reviews/meta-reviews, we arrive at a total of 7,478 instances for our experiments. We use 15 %

*equal contribution

[1]https://www.iitp.ac.in/~ai-nlp-ml/resources.html#decision-aware-meta-review

[2]https://openreview.net/

2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) | 978-1-6654-1770-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/JCDL52503.2021.00064

978-1-6654-1770-9/21/$31.00 ©2021 IEEE
DOI 10.1109/JCDL52503.2021.00064

of the data as the test set (1121), 75% as the training set (5683), and the remaining 10% as the validation set (674). Our proposed model treats each review individually (does not concatenate), so for training, we create a permutation in ordering the three reviews to have a training set of 34,098 reviews. We provide the total number of reviews, meta-reviews, and length in Table I. Table II shows the distribution of reviews across paper-categories (Accepted or Rejected).

TABLE I
DETAILS OF OUR REVIEW DATASET. LENGTH IN NUMBER OF WORDS

Data	# Data	Max Length	Min Length	Avg Length
Peer-reviews	22,434	4051	11	∼420
Meta-review	7,478	2117	6	∼131

TABLE II
PAPER DISTRIBUTION FOR DECISION PREDICTION TASK

Decision	Train	Test	Validation
Accept	1979	232	373
Reject	3704	442	748

III. METHODOLOGY

As most papers have three reviews in our data, we propose three encoders and a decoder. To make our proposed multi-source transformer model (Fig. 1) decision-aware, we use the last layer before the prediction using three encoders and pass it in the decoder layers to provide the context.

A. *Encoder-Decoder for Meta Review Generation:*

The encoder acts as a feature extractor that maps input vectors to a high-level representation. With this representation, the decoder predicts the sequence one at a time auto-regressively. Our proposed model has three encoders, each consisting of N layers. With each encoder layer performing multi-head attention, we add the previous residual connections for normalization, which goes in a feed-forward network followed by a new residual for the final normalization. The decoder consists of M layers, each having representations with cross-attention applied parallelly to each encoder's key-value pair that further goes to normalization and feed-forward network. Let us consider the input sequence of reviews R1, R2, R3 are as follows: $R1 = (r_1^1, r_2^1, r_3^1....r_n^1), R2 = (r_1^2, r_2^2, r_3^2....r_n^2)$ and $R3 = (r_1^3, r_2^3, r_3^3....r_n^3)$ These are then mapped to the representation of $Z = [z_1, z_2, z_3]$, where each z_1, z_2, z_3 have two components, *viz.* the hidden states $[h_1, h_2, h_3]$ and key-value pairs $[kp_1, kp_2, kp_3]$. Given Z, the decoder then generates the output meta-review sequence $Y = (Y_1, Y_2, Y_3....Y_n)$

B. *Decision Aware:*

We concatenate the three encoders' hidden states $[h_1, h_2, h_3]$ after average pooling and pass to the fully connected linear layer. We use these outputs $[\hat{h}]$ as the context of the decision in every decoder layer (at the last hidden state before normalization).

C. *Decision Prediction*

To predict a decision from the encoded representation, we use hidden states $[\hat{h}]$ and pass through ReLU activation then feed to new linear layer for decision predicton output $\hat{y}_i$ *(Accept / Reject)*.

D. *Loss Function*

Loss= $\alpha(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - y_i))_{decision} +$
$\beta(-\sum p([R_1, R_2, R_3]) * \log(q([R_1, R_2, R_3])))_{generation}$

where $p([R_1, R_2, R_3])$ is the true distribution and $q([R_1, R_2, R_3])$ is the estimated distribution; $\hat{y}_i$ is the i-th scalar value in the model output, y_i is the corresponding target value. For our experiments, loss is binary cross entropy for decision prediction and later is cross entropy for MRG, we chose α, β to be 1, M=2, and N=2.

IV. EXPERIMENTS

We choose to train a Byte-pair encoding tokenizer with the same special tokens as RoBERTa [5] and pick its size to be 52,000. We now have a vocabulary, a list of the most frequent tokens ranked by the frequency. Furthermore, to take the count of the word's order in the input sequence, we give additional positional encoding information to the input. In a multi-source transformer the cross attention mechanism can be modeled in several ways [6] We use the parallel strategy for our experiments to produce a rich representation from our three encoders for the task. We keep the learning rate as 5e-05, the number of beams for beam search=4, and optimizer=Adam. We train the different models for 100 epochs with learning rate scheduler=linear and choose the best variant in terms of validation loss for generation and prediction. Our initial experiments include PEGASUS, a BART-based summarizer which we treat as the baselines for comparison, and three variants of our model.

PEGASUS [3] is an abstractive summarization algorithm which uses self-supervised objective Gap Sentences Generation (GSG) to train a transformer-based encoder-decoder model. In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary.

BART [4] uses a standard seq2seq architecture with a bidirectional encoder and a unidirectional decoder. The pre-training task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token. We use the Hugging Face implementation of 12 encoder layers and 12 decoder layers with pre-trained weights[3] and fine-tune on our dataset to generate the meta-review.

Simple Meta-Review Generator (S-MRG)

This is a simple transformer model with only three encoders, each with two encoder layers and a decoder of two decoder layers for generating the sequence, one at a time auto-regressively.

[3]https://huggingface.co/transformers/model_doc/bart.html

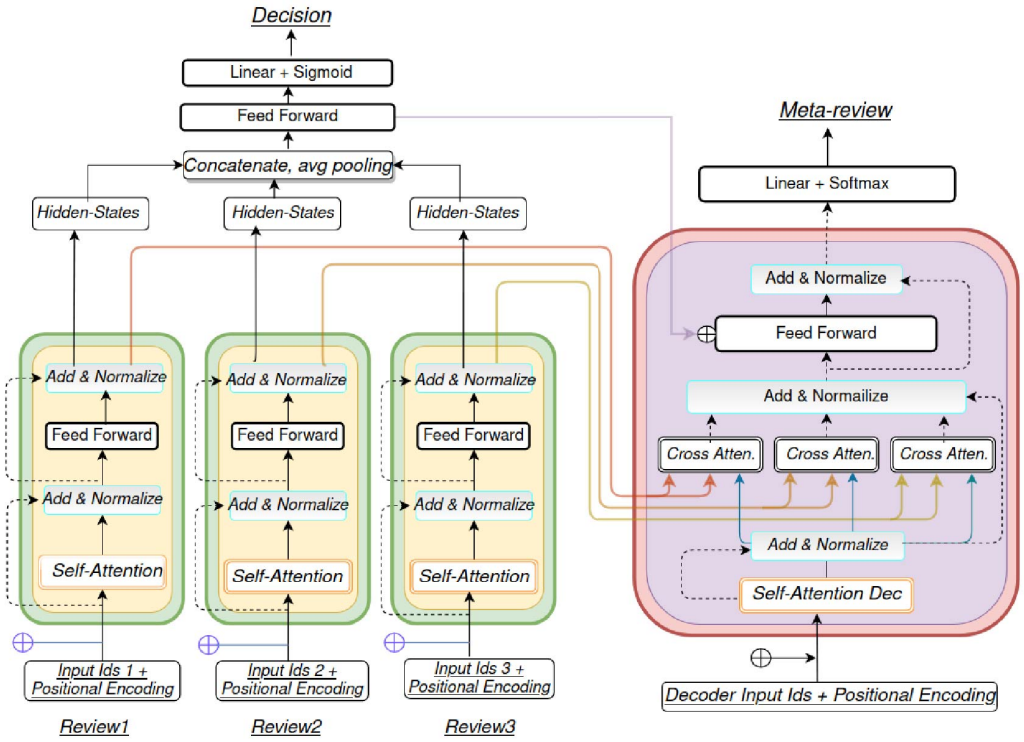

Fig. 1. The overall architecture of decision-aware multi-encoder transformer network for meta-review generation and decision prediction in peer review.

MRG with Decision at last: $MRG\ Decision_{LAST}$ uses the last hidden states of the decoder for both tasks, one with a linear layer for generation and another separate linear layer combined with dropout and ReLU for decision prediction.

Proposed Approach/model: Decision-aware MRG: $MRG\ \ Decision_{ENCODED}$ predicts the decision from encoders and carry the decision vector encoded from the encoder-hidden state output to the decoder layer, to provide the context to the generator module.

Please note that we do not use the baselines for the decision prediction task. Our proposed approach takes input from the decision-prediction module (hence *decision-aware* just as human chairs do) to generate the meta-reviews.

V. Evaluation

A. *Decision Prediction*

We show the evaluation results in Table III. The results are not surprising as predicting the acceptance of a paper is a very complicated task, and ICLR being a premier machine learning venue, the acceptance rate is very low. Hence, our review dataset has rejection comments in plenty as compared to the accepted ones.

TABLE III
Results with respect to F1 score and overall accuracy for decision prediction

Model	Accept	Reject	Accuracy
MRG Decision $_{LAST}$	0.28	0.75	0.6306
MRG Decision $_{ENCODED}$	0.29	0.75	0.6324

B. *Meta-review Generation*

*1) **Quantitative Analysis**:* To evaluate the model-generated meta-reviews, we use some popular automatic evaluation metrics for text generation and summarization. Since a single metric does not give the best evaluation for a generated summary, we use ROUGE-1, ROUGE-2, ROUGE-3 [7], BERTScore [8], S3 [9] and BLEU [10] metrics (Table IV). We achieve comparable results with the BART-based summarization model, although we argue that the evaluation is unfair as MRG and summarization are not the same tasks.

*2) **Qualitative Analysis**:* Table V shows the MRG output of the different techniques. We use the pre-trained models for PEGASUS and BART but fine-tune on our review dataset. Our custom architectures with three different setups are entirely trained on our dataset. We find that although PEGASUS generated meta-review manifests sentences with polarity, the output is not detailed. The significant aspects of concern in the human-generated review are not prominent in the generated meta-review. The overall polarity and also the decision does not match with the original meta-review. BART, on the other hand, is a large language model with 406 million parameters. We found that outputs are detailed, generated meta-reviews manifest polarity, and highlight merits and demerits. *Our model MRG with Decision at last,* does a reasonable job to capture the polarity (see Table III), and also the generated meta-review is in the third person. However, we notice that some irrelevant text from other paper's common primary

TABLE IV
MODEL SCORES FOR AUTOMATIC EVALUATION METRICS. THE OUTPUT IS THE AVERAGE OF ALL THE SCORES IN THE TEST SET.

Model	S-MRG	MRG Decision $_{ENCODED}$	MRG Decision $_{LAST}$	Pegasus	BART
ROUGE-1	0.27601	0.30381	0.2738	0.18842	0.31544
ROUGE-2	0.04805	0.05744	0.04612	0.03775	0.0763
ROUGE-3	0.00642	0.00853	0.0063	0.00812	0.01941
S3(pyr/resp)	0.1592/0.3330	0.1933/0.3503	0.1578/0.3312	0.1046/0.3135	0.23806/ 0.384974
BertScore(f1)	0.5450	0.5539	0.5434	0.5350	0.573946
BLEU	1.4977	1.7529	1.3568	2.1358	2.8543

TABLE V
GROUND TRUTHS AND AUTOMATICALLY GENERATED META-REVIEW FOR A GIVEN PAPER

Original Meta-Review⟶ https://openreview.net/forum?id=HyenWc5gx
The proposed approach is not consistently applied for the different experiments; this significantly harms the overall value of the research. The results are also quite domain-specific, and it is not clear if the findings would hold more generally. The paper is not clearly organised or written and does not give a specific enough introduction to the field of transfer learning.
Decision-aware MRG
This paper proposes a method for modeling higher-order interactions in Poisson processes. Unfortunately, the reviewers do not feel that the paper, in its current state, meets the bar for ICLR. In particular, reviewers found the descriptions unclear and the justifications lacking. While the responses did aid the reviewers understanding, the paper would benefit from rewriting and more careful thought given to the experimental design.
MRG decision at last
The authors propose a multi-scale architecture for learning representations that can learn which dimensions to pass through more flow layers based on a heuristic that judges the contribution to the likelihood. The authors compare the technique to some other flow based approaches. The reviewers asked for more experiments, which the authors delivered. However, the reviewers noted that a comparison to the SOTA for CIFAR in this setting was missing. Several reviewers raised their scores, but none were willing to argue for acceptance.
Simple MRG
This paper proposes to use hypernetwork to prevent catastrophic forgetting. Overall, the paper is well-written, well-motivated, and the idea is novel. Experimentally, the proposed approach achieves SOTA on various (well-chosen) standard CL benchmarks (notably P-MNIST for CL, Split MNIST) and also does reasonably well on Split CIFAR-10/100 benchmark. The authors are suggested to investigate alternative penalties in the rehearsal objective, and also add comparison with methods like HAT and PackNet.
PEGASUS
ICLR 2018 Conference Acceptance Decision. The reviewers have unanimously expressed strong concerns about the novelty of the paper.. Reject
BART
This paper proposes a method for transfer learning, i.e. leveraging a network trained on some original task A in learning a new task B, which not only improves performance on the new task A, but also tries to avoid degradation in performance on A. The general idea is based on encouraging a model trained on A, while training on the target task B to match fake targets produced by the model itself but when it is trained only on the original task. Experiments show that this method can help in improving the result on task B and is better than other baselines, including standard fine-tuning. However, the details of

keywords is present in generated review, which is eventually noise in the output. It performs inferior compared to the custom architecture without decision. This can be because the shared parameters attach equal weights to the decision prediction and the generation tasks. **Our proposed decision-aware MRG model** writes the meta-review in the third person/as meta-reviewer in coherence with the existing peer reviews and brings out the merits and demerits of the paper. The decision prediction module has an accuracy of 63%, which can be further improved by augmenting reviewer sentiment and review-paper interaction as additional channels of information to the model.

OBSERVATION AND CONCLUSION

Based on our evaluation of the outputs, we found that the decision plays a vital role in generating the meta-review. Our architecture with decision awareness performs better in the generation and also has a higher F1-score for decision prediction. However, given the complex nature of this task, instead of predicting, it would be more sensible to directly use the ground-truth decisions to the generation model and let the model focus only on meta-review generation. We also see that the BART-based summarizer [4] with 12 encoders, and 12 decoder layers perform slightly better w.r.t. the quantitative evaluation metrics as it is trained on a huge corpus consisting of 160 GB of news, books, stories, and web text, giving a rich representation of the textual features when fine-tuned for the specific downstream task. Next we would like to investigate appropriate metrics for MRG evaluation and incorporate paper-reviewer interaction, the reviewer's sentiment, reviewer recommendation score, reviewer confidence for decision prediction, and meta-review generation.

REFERENCES

[1] C. Bhatia, T. Pradhan, and S. Pal, "Metagen: An academic meta-review generation system," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1653–1656.

[2] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," *arXiv preprint arXiv:1905.03197*, 2019.

[3] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] J. Libovickỳ, J. Helcl, and D. Mareček, "Input combination strategies for multi-source transformer decoder," *arXiv preprint arXiv:1811.04716*, 2018.

[7] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[9] M. Peyrard, T. Botschen, and I. Gurevych, "Learning to score system summaries for better content selection evaluation." in *Proceedings of the Workshop on New Frontiers in Summarization*, 2017, pp. 74–84.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.