Can We Automate Scientific Reviewing?

Weizhe Yuan
Carnegie Mellon University
weizhey@cs.cmu.edu

Pengfei Liu *
Carnegie Mellon University
pliu3@cs.cmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

TL;QR

This paper proposes to use NLP models to generate reviews for scientific papers. The model is trained on the ASAP-Review dataset and evaluated on a set of metrics to evaluate the quality of the generated reviews. It is found that the model is not very good at summarizing the paper, but it is able to generate more detailed reviews that cover more aspects of the paper than those created by humans. The paper also finds that both human and automatic reviewers exhibit varying degrees of bias and biases, and that the system generate more biased reviews than human reviewers. ("Too Long; Quick Read", this paragraph, is generated by our system.)

Abstract

The rapid development of science and technology has been accompanied by an exponential growth in peer-reviewed scientific publications. At the same time, the review of each paper is a laborious process that must be carried out by subject matter experts. Thus, providing high-quality reviews of this growing number of papers is a significant challenge. In this work, we ask the question "*can we automate scientific reviewing?*", discussing the possibility of using state-of-the-art natural language processing (NLP) models to generate first-pass peer reviews for scientific papers. Arguably the most difficult part of this is defining what a "good" review is in the first place, so we first discuss possible evaluation measures for such reviews. We then collect a dataset of papers in the machine learning domain, annotate them with different aspects of content covered in each review, and train targeted summarization models that take in papers to generate reviews. Comprehensive experimental results show that system-generated reviews tend to touch upon more aspects of the paper than human-written reviews, but the generated text can suffer from lower constructiveness and factuality for all aspects except the explanation of the core ideas of the papers, which are largely factually correct. We finally summarize *eight* challenges in the pursuit of a good review generation system together with potential solutions, which, hopefully, will inspire more future research.

While our conclusion is that the technology is not yet ready for use in high-stakes review settings, to spur future research on this subject we make all code, and the dataset publicly available: https://github.com/neulab/ReviewAdvisor and a system demo: *ReviewAdvisor*: http://review.nlpedia.ai/ (See demo screenshot in A.2). The review of this paper (without the TL;QR section) written by the system proposed in this paper can be found in A.1

1 Introduction

The number of published papers is growing exponentially (Tabah, 1999; De Bellis, 2009; Bornmann and Mutz, 2015). While this may be positively viewed as indicating acceleration of scientific progress, it also poses great challenges for researchers, both in reading and synthesizing the relevant literature for one's own benefit, and for performing *peer review* of papers to vet their correctness and merit. With respect to the former, a large body of existing work explores automatic summarization of a paper or a set of papers for automatic survey generation (Mohammad et al., 2009; Jha et al., 2013, 2015b,a; Yasunaga et al., 2019b; Cohan et al., 2018b; Xing et al., 2020). However, despite the fact that peer review is an important, but laborious part of our scientific process, automatic systems to aid in the peer review process re-

*Corresponding author.

main relatively underexplored. Bartoli et al. (2016) investigated the feasibility of generating reviews by surface-level term replacement and sentence reordering, and Wang et al. (2020) (contemporaneously and independently) propose a two-stage information extraction and summarization pipeline to generate paper reviews. However, both do not extensively evaluate the quality or features of the generated review text.

In this work, we are concerned with providing at least a preliminary answer to the ambitious overarching question: *can we automate scientific reviewing?* Given the complexity of understanding and assessing the merit of scientific contributions, we do not expect an automated system to be able to match a well-qualified and meticulous human reviewer at this task any time soon. However, some degree of review automation may assist reviewers in their assessments, or provide guidance to junior reviewers who are just learning the ropes of the reviewing process. Towards this goal, we examine two concrete research questions, the answers to which are prerequisites to building a functioning review assistant:

Q1: What are the desiderata of a good automatic reviewing system, and how can we quantify them for evaluation? Before developing an automatic review system, we first must quantify what constitutes a good review in the first place. The challenge of answering this question is that a review commonly involves both objective (e.g. "lack of details necessary to replicate the experimental protocol") and subjective aspects (e.g. "lack of potential impact"). Due to this subjectivity, defining a "good" review is itself somewhat subjective.

As a step towards tackling this challenge, we argue that it is possible to view review generation as a task of *aspect-based scientific paper summarization*, where the summary not only tries to summarize the core idea of a paper, but also assesses specific aspects of that paper (e.g. novelty or potential impact). We evaluate review quality from multiple perspectives, in which we claim a good review not only should make a good summary of a paper but also consist of factually correct and fair comments from diverse aspects, together with informative evidence.

To operationalize these concepts, we build a dataset of reviews, named `ASAP-Review`[1] from machine learning domain, and make fine-grained annotations of aspect information for each review, which provides the possibility for a richer evaluation of generated reviews.

Q2: Using state-of-the-art NLP models, to what extent can we realize these desiderata? We provide an initial answer to this question by using the aforementioned dataset to train state-of-the-art summarization models to generate reviews from scientific papers, and evaluate the output according to our evaluation metrics described above. We propose different architectural designs for this model, which we dub *ReviewAdvisor* (§4), and comprehensively evaluate them, interpreting their relative advantages.

Lastly, we highlight our main observations and conclusions:

(1) What are review generation systems (not) good at? Most importantly, we find the constructed automatic review system *generates non-factual statements* regarding many aspects of the paper assessment, which is a serious flaw in a high-stakes setting such as reviewing. However, there are some bright points as well. For example, it *can often precisely summarize the core idea* of the input paper, which can be either used as a draft for human reviewers or help them (or general readers) quickly understand the main idea of the paper to be reviewed (or pre-print papers). It can also generate reviews that *cover more aspects* of the paper's quality than those created by humans, and provide evidence sentences from the paper. These could potentially provide a preliminary template for reviewers and help them quickly identify salient information in making their assessment.

(2) Will the system generate biased reviews? Yes. We present methods to identify and quantify potential biases in reviews (§5.3), and find that both human and automatic reviewers exhibit varying degrees of bias. (i) regarding native vs. non-native English speakers: papers of native English speakers tend to obtain higher scores on "`Clarity`" from human reviewers than non-native English ones,[2] but the automatic review generators narrow this gap. Additionally, system reviewers are harsher than human reviewers when commenting regarding the paper's "`Originality`" for non-native

[1] `AS`pect-enh`A`nced `P`eer Review dataset

[2] Whether this actually qualifies as "bias" is perhaps arguable. Papers written by native English speakers may be more clear due to lack of confusing grammatical errors, but the paper may actually be perfectly clear but give the impression of not being clear because of grammatical errors.

English speakers. (ii) regarding anonymous vs. non-anonymous submissions: both human reviewers and system reviewers favor non-anonymous papers, which have been posted on non-blind preprint servers such as arXiv[3] before the review period, more than anonymous papers in all aspects.

Based on above mentioned issues, we claim that a review generation system can not replace human reviewers at this time, instead, it may be helpful as part of a machine-assisted human review process. Our research also enlightens what's next in pursuing a better method for automatic review generation or assistance and we summarize eight challenges that can be explored for future directions in §7.2.

2 What Makes a Good Peer Review?

Although peer review has been adopted by most journals and conferences to identify important and relevant research, its effectiveness is being continuously questioned (Smith, 2006; Langford and Guzdial, 2015; Tomkins et al., 2017; Gao et al., 2019; Rogers and Augenstein, 2020).

As concluded by Jefferson et al. (2002b), "until we have properly defined the objectives of peer-review, it will remain almost impossible to assess or improve its effectiveness." Therefore we first discuss the possible objectives of peer review.

2.1 Peer Review for Scientific Research

A research paper is commonly first reviewed by several committee members who usually assign one or several scores and give detailed comments. The comments, and sometimes scores, cover diverse aspects of the paper (e.g. "clarity," "potential impact"; detailed in §3.2.1), and these aspects are often directly mentioned in review forms of scientific conferences or journals[4].

Then a senior reviewer will often make a final decision (i.e., "reject" or "accept") and provide comments summarizing the decision (i.e., a meta-review).

After going through many review guidelines[5] and resources about how to write a good review[6], we summarize some of the most frequently mentioned desiderata below:

1. Decisiveness: A good review should take a clear stance, selecting high-quality submissions for publication and suggesting others not be accepted (Jefferson et al., 2002a; Smith, 2006).

2. Comprehensiveness: A good review should be well-organized, typically starting with a brief summary of the paper's contributions, then following with opinions gauging the quality of a paper from different aspects. Many review forms explicitly require evaluation of different aspects to encourage comprehensiveness.

3. Justification: A good review should provide specific reasons for its assessment, particularly whenever it states that the paper is lacking in some aspect. This justification also makes the review more constructive (another oft-cited desiderata of reviews), as these justifications provide hints about how the authors could improve problematic aspects in the paper (Xiong and Litman, 2011).

4. Accuracy: A review should be factually correct, with the statements contained therein not being demonstrably false.

5. Kindness: A good review should be kind and polite in language use.

Based on above desiderata, we make a first step towards evaluation of reviews for scientific papers and characterize a "good" review from multiple perspectives.

2.2 Multi-Perspective Evaluation

Given input paper D and meta-review R^m, our goal is to evaluate the quality of review R, which can be either manually or automatically generated. We also introduce a function $\textsc{Dec}(D) \in \{1, -1\}$ that indicates the final decision of a given paper reached by the meta-review ("accept" or "reject"). Further, $\textsc{Rec}(R) \in \{1, 0, -1\}$ represents the acceptance recommendation of a particular review: "accept," "neutral," or "reject" (see Appendix A.3 for details).

[3] https://arxiv.org/

[4] For example, one example from ACL can be found at: https://acl2018.org/downloads/acl18-review_form.html

[5] https://icml.cc/Conferences/2020/ReviewerGuidelines https://NeurIPS.cc/Conferences/2020/PaperInformation/ReviewerGuidelines, https://iclr.cc/Conferences/2021/ReviewerGuide

[6] https://players.brightcove.net/3806881048001/rFXiCa5uY_default/index.html?videoId=4518165477001, https://soundcloud.com/nlp-highlights/77-on-writing-quality-peer-reviews-with-noah-a-smith, https://www.aclweb.org/anthology/2020.acl-tutorials.4.pdf, https://2020.emnlp.org/blog/2020-05-17-write-good-reviews

Desiderata	Metrics	Range	Automated
Decisiveness	RAcc	[-1, 1]	No
Comprehen.	ACov	[0, 1]	Yes
	ARec	[0, 1]	Yes
Justification	Info	[0, 1]	No
Accuracy	SAcc	[0, 1]	No
	ACon	[0, 1]	No
Others	ROUGE	[0, 1]	Yes
	BERTScore	[-1, 1]	Yes

Table 1: Evaluation metrics from different perspectives. "Range" represents the range value of each metric. "Automated" denotes if metrics can be obtained automatically.

Below, we discuss evaluation metrics that can be used to approximate the desiderata of reviews described in the previous section. And we have summarized them in Tab. 1.

2.2.1 D1: Decisiveness

First, we tackle the decisiveness, as well as accuracy of the review's recommendation, through Recommendation Accuracy (RAcc). Here we use the final decision regarding a paper and measure whether the acceptance implied by the review is consistent with the actual accept/reject decision of the reviewed paper. It is calculated as:

$$\text{RAcc}(R) = \text{Dec}(D) \cdot \text{Rec}(R) \quad (1)$$

A higher score indicates that the review more decisively and accurately makes an acceptance recommendation.

2.2.2 D2: Comprehensiveness

A comprehensive review should touch on the quality of different aspects of the paper, which we measure using a metric dubbed Aspect Coverage (ACov). Specifically, given a review R, aspect coverage measures how many aspects (e.g. clarity) in a predefined aspect typology (in our case, §3.2.1) have been covered by R.

In addition, we propose another meta-Aspect Recall (ARec), which explicitly takes the meta-review R^m into account. Because the meta-review is an authoritative summary of all the reviews for a paper, it provides an approximation of which aspects, and with which sentiment polarity, should be covered in a review. Aspect recall counts how many aspects in meta-review R^m are covered by general review R, with higher aspect recall indicating better agreement with the meta-review.[7]

2.2.3 D3: Justification

As defined in §2.1, a good peer review should provide hints about how the author could improve problematic aspects. For example, when reviewers comment: "this paper lacks important references", they should also list these relevant works. To satisfy this justification desideratum, we define a metric called Informativeness (Info) to quantify how many negative comments[8] are accompanied by corresponding evidence.

First, let $n_{na}(R)$ denote the number of aspects in R with negative sentiment polarity. $n_{nae}(R)$ denotes the number of aspects with negative sentiment polarity that are supported by evidence. The judgement of supporting evidence is conducted manually (details in Appendix A.3). Info is calculated as:

$$\text{Info}(R) = \frac{n_{nae}(R)}{n_{na}(R)} \quad (2)$$

And we set it to be 1 when there are no negative aspects mentioned in a review.

2.3 D4: Accuracy

We use two measures to evaluate the accuracy of assessments. First, we use Summary Accuracy (SAcc) to measure how well a review summarizes contributions of a paper. It takes value of 0, 0.5, or 1, which evaluates the summary part of the review as incorrect/absent, partially correct, and correct. The correctness judgement is performed manually, with details listed in Appendix A.3.

Info implicitly requires that negative aspects should be supported with evidence, ignoring the quality of this evidence. However, to truly help to improve the quality of a paper, the evidence for negative aspects should be factual as well. Here we propose Aspect-level Constructiveness (ACon), the percentage of the supporting statements ($n_{nae}(R)$) that are judged as valid support by human annotators. If $n_{nae}(R)$ is 0, we set its ACon as 1. This

[7] Notably, this metric potentially biases towards high scores for reviews that were considered in the writing of the meta-review. Therefore, higher aspect recall is not the only goal, and should be taken together with other evaluation metrics.

[8] We only consider whether the reviewer has provided enough evidence for negative opinions since we find that most human reviewers rarely provide evidence for their positive comments.

metric will implicitly favor reviews that do not provide enough evidence for negative aspects. However, in this case, the INFO of those reviews will be rather low. The details of evaluating "validity" are also described in Appendix A.3.

2.4 D5: Kindness

While kindness is very important in maintaining a positive research community, accurately measuring it computationally in a nuanced setting such as peer review is non-trivial. Thus, we leave the capturing of kindness in evaluation to future work.

2.5 Similarity to Human Reviews

For automatically generated reviews, we also use Semantic Equivalence metrics to measure the similarity between generated reviews and reference reviews. The intuition is that while human reviewers are certainly not perfect, knowing how close our generated reviews are to existing human experts may be informative. Here, we investigate two specific metrics: ROUGE (Lin and Hovy, 2003) and BERTScore (Zhang et al., 2019). The former measures the surface-level word match while the latter measures the distance in embedding space. Notably, for each source input, there are multiple reference reviews. When aggregating ROUGE and BERTScore, we take the maximum instead of average since it is not necessary for generated reviews to be close to all references.

3 Dataset

Next, in this section we introduce how we construct a review dataset with more fine-grained metadata, which can be used for system training and the multiple perspective evaluation of reviews.

3.1 Data Collection

The advent of the Open Peer Review system[9] makes it possible to access review data for analysis or model training/testing. One previous work (Kang et al., 2018) attempts to collect reviews from several prestigious publication venues including the Conference of the Association of Computational Linguistics (ACL) and the International Conference on Learning Representations (ICLR). However, there were not nearly as many reviews accumulated in OpenReview at that time[10] and other private reviews only accounted for a few hundred. Therefore we decided to collect our own dataset Aspect-enhanced Peer Review (ASAP-Review).

We crawled ICLR papers from 2017-2020 through OpenReview[11] and NeurIPS papers from 2016-2019 through NeurIPS Proceedings[12]. For each paper's review, we keep as much metadata information as possible. Specifically, for each paper, we include following metadata information that we can obtain from the review web page:

- Reference reviews, which are written by a committee member.
- Meta reviews, which are commonly written by an area chair (senior committee member).
- Decision, which denotes a paper's final "accept" or "reject" decision.
- Other information like url, title, author, etc.

We used Allenai Science-parse[13] to parse the pdf of each paper and keep the structured textual information (e.g., titles, authors, section content and references). The basic statistics of our ASAP-Review dataset is shown in Tab. 2.

	ICLR	NeurIPS	Both
Accept	1,859	3,685	5,544
Reject	3,333	0	3,333
Total	5,192	3,685	8877
Avg. Full Text Length	7,398	5,916	6782
Avg. Review Length	445	411	430
# of Reviews	15,728	12,391	28,119
# of Reviews per Paper	3.03	3.36	3.17

Table 2: Basic statistics of ASAP-Review dataset. Note that NeurIPS only provide reviews for accepted papers to the public.

3.2 Aspect-enhanced Review Dataset

Although reviews exhibit internal structure, for example, as shown in Fig. 3, reviews commonly start with a paper summary, followed by different aspects of opinions, together with evidence. In practice, this useful structural information cannot be obtained directly. Considering that fine-grained information about the various aspects touched on by the review plays an essential role in review evaluation, we conduct aspect annotation of those reviews. To this end, we first (i) introducing an aspect typology and (ii) perform human annotation.

[9] https://openreview.net/

[10] During that time, there are no reviews of ICLR from 2018 to 2020 nor reviews of NeurIPS from 2018 to 2019.

[11] https://openreview.net

[12] http://papers.NeurIPS.cc

[13] https://github.com/allenai/science-parse

Figure 1: Data annotation pipeline.

3.2.1 Aspect Typology and Polarity

We define a typology that contains 8 aspects, which follows the ACL review guidelines[14] with small modifications, which are Summary (SUM), Motivation/Impact (MOT), Originality (ORI), Soundness/Correctness (SOU), Substance (SUB), Replicability (REP), Meaningful Comparison (CMP) and Clarity (CLA). The detailed elaborations of each aspect can be found in Supplemental Material B.1. Inside the parentheses are what we will refer to each aspect for brevity. To take into account whether the comments regarding each aspect are positive or negative, we also mark whether the comment is positive or negative for every aspect (except summary).

3.2.2 Aspect Annotation

Overall, the data annotation involves four steps that are shown in Fig. 1.

Step 1: Manual Annotation To manually annotate aspects in reviews, we first set up a data annotation platform using Doccano[15]. We asked 6 students from ML/NLP backgrounds to annotate the dataset. We asked them to tag an appropriate text span that indicates a specific aspect. For example, "The results are new [Positive Originality] and important to this field [Positive Motivation]". The detailed annotation guideline can be found in Supplemental Material B.1. Each review is annotated by two annotators and the lowest pair-wise Cohen kappa is 0.653, which stands for substantial agreement. In the end, we obtained 1,000 human-annotated reviews in total. The aspect statistics in this dataset are shown in Fig. 2-(a).

Step 2: Training an Aspect Tagger Since there are over 20,000 reviews in our dataset, using human labor to annotate them all is unrealistic. Therefore, we use the annotated data we do have to train an aspect tagger and use it to annotate the remaining reviews. The basic architecture of our aspect tagger contains a pre-trained model BERT (Devlin et al., 2019) and a multi-layer perceptron. The training details can be found in Appendix A.4.

Figure 2: (a) and (b) represent distributions over seven aspects obtained by human and BERT-based tagger respectively. Red bins represent positive sentiment while green ones suggest negative sentiment. We omit "Sum" aspect since there is no polarity definition of it.

Step 3: Post-processing However, after inspecting the automatically labeled dataset, we found that there appears to be some common problems such as interleaving different aspects and inappropriate boundaries. To address those problems, we used seven heuristic rules to refine the prediction results and they were executed sequentially. The detailed heuristics can be found in Appendix A.5. An example of our model prediction after applying heuristic rules is shown in Appendix A.6. Fig. 2-(b) shows the distribution of all reviews over different aspects. As can be seen, the relative number of different aspects and the ratio of positive to negative are very similar across human and automatic annotation.

Step 4: Human Evaluation To evaluate the data quality of reviews' aspects, we conduct human evaluation. Specifically, we measure both aspect precision and aspect recall for our defined 15 aspects.

We randomly chose 300 samples from our au-

[14] https://acl2018.org/downloads/acl2018-review-form.html. We manually inspected several review guidelines from ML conferenecs and found the typology in ACL review guideline both general and comprehensive.

[15] https://github.com/doccano/doccano

Aspect	Polarity	Precision	Recall
Summary		95%	100%
Motivation	+	94%	89%
	–	72%	71%
Originality	+	95%	87%
	–	94%	80%
Soundness	+	95%	98%
	–	92%	79%
Substance	+	90%	94%
	–	90%	78%
Replicability	+	100%	50%
	–	77%	71%
Clarity	+	97%	92%
	–	92%	73%
Comparison	+	85%	100%
	–	94%	94%

Table 3: Fine-grained aspect precision and recall for each aspect. + denotes positive and – denotes negative.

tomatically annotated dataset and assigned each sample to three different annotators to judge the annotation quality. As before, these annotators are all from ML/NLP backgrounds.

The detailed calculation for aspect precision and aspect recall can be found in Appendix A.7. Under these criteria, we achieved 82.75% aspect precision and 85.19% aspect recall. The ne-grained aspect precision and aspect recall for each aspect is shown in Tab. 3. The aspect recall for positive replicability is low. This is due to the fact that there are very few mentions of positive replicability. And in our human evaluation case, the system identi ed one out of two, which results in 50%. Other than that, the precision and recall are much higher.[16]

Besides, one thing to mention is that our evaluation criterion is very strict, and it thus acts as a lower bound for these two metrics.

4 Scienti c Review Generation

4.1 Task Formulation

The task of scienti c review generation can be conceptualized as a aspect-based scienti c paper summarization task but with a few important differences. Speci cally, most current works summarize a paper (i) either from an "author view" that only use content written by the author to form a summary (Cohan et al., 2018a; Xiao and Carenini, 2019; Erera et al., 2019; Cohan et al., 2018a; Cachola et al., 2020b), (ii) or from a "reader view" that argues a paper's summary should take into account the view of those in the research community (Qazvinian and Radev, 2008; Cohan and Goharian, 2017; Yasunaga et al., 2019a).

In this work, we extend the view of scienti c paper summarization from "author" or "reader" to "reviewer", and claim that a good summary of a scienti c paper can not only re ect the core idea but also contains critical comments from different aspects made by domain experts, which usually requires knowledge beyond the source paper itself. The advantages lie in: (i) authors helping them identify weak points in their paper and make it stronger. (ii) reviewers relieving them from some of the burden of reviewing process. (iii) readers helping them quickly grasp the main idea of the paper and letting them know what "domain experts" (our system) comments on the paper are. The three views of scienti c paper summarization are shown in Fig. 3.

Figure 3: Summarization from three different views for the paper "Attention Is All You Need" (Vaswani et al., 2017). Summareis from three views (author, reader, reviewer) comes from the paper's abstract, citance (i.e., a paper that cites this paper) and peer review respectively.

4.2 System Design

Despite the fact that our dataset contains fewer training samples compared with other benchmark summarization datasets, the few-shot learning ability of recent contextualized pre-trained models (Radford et al., 2019; Brown et al., 2020; Cachola

[16] The recall numbers for negative aspects are lower than positive aspects. However, we argue that this will not affect the delity of our analysis much because (i) we observe that the imperfect recall is mostly (over 85%) caused by partial recognition of the same negative aspect in a review instead of inability to recognize at least one. This will not affect our calculation of Aspect Coverage and Aspect Recall very much. (ii) The imperfect recall will slightly pull up Aspect Score (will discuss in x5.3.1), but the trend will remain the same.

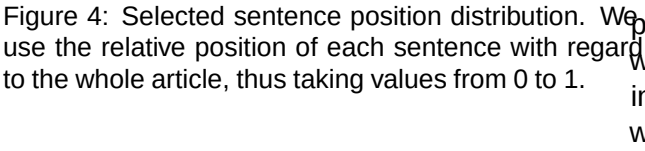

Figure 4: Selected sentence position distribution. We use the relative position of each sentence with regard to the whole article, thus taking values from 0 to 1.

et al., 2020a) still put training a passable review generation system from this dataset within grasp. We use BART (Lewis et al., 2019), which is a denoising autoencoder for pretraining sequence-to-sequence models, as our pre-trained model since it has shown superior performance on multiple generation tasks.

However, even if we can take the advantage of this pre-trained model, how to deal with lengthy text in the context of using a pre-trained model (BART, for example, has a standard length limit of 1024 since it was pre-trained on texts of this size) remains challenging. After multiple trials, we opted for a two-stage method detailed below, and describe other explorations that were less effective in Appendix A.8.

4.2.1 Two-stage Systems for Long Documents

Instead of regarding text generation as a holistic process, we decompose it into two steps, using an *extract-then-generate* paradigm (Chen and Bansal, 2018; Gehrmann et al., 2018; Subramanian et al., 2019; Dou et al., 2020). Specically, we first perform content selection, extracting salient text pieces from source documents (papers), then generate summaries based on these extracted texts.

To search for an effective way to select content that is most useful for constructing a review generation system, we operationalize the first extraction step in several ways. One thing to notice is that the extraction methods we use here mainly focus on heuristics. We leave more complicated selection methods for future work.

Oracle Extraction First, for comparison purposes, we construct an oracle for each paper which is the extraction that achieves highest average ROUGE scores with respect to reference reviews, specically using the greedy method described in Nallapati et al. (2017). Note that for each paper with multiple reviews, we construct multiple oracles for that paper. We assume that oracle extractions can reect where reviewers pay more attention to when they are writing reviews. The selected sentence position distribution in oracles is shown in Fig. 4.

Section-based Extraction Scientic papers are highly structured. As a convention, a scientic paper usually describes problem background, related work comparison, as well as its own contributions in the introduction part. Regarding this method, we only use the introduction section, which can be regarded as a baseline model.

Cross-entropy (CE) Method Extraction Here we select salient sentences from the full text range. The way we do so is through a two-step selection process:

1. Select sentences containing certain informative keywords (e.g. *propose*) which are detailed in Appendix A.9. Those selected sentences form a set S.
2. Select a subset $S^0 \subset S$ such that sentences in S^0 cover diverse content and satisfy a length constraint.

In the second step, we use the cross-entropy method introduced in Feigenblat et al. (2017) where we select diverse content by maximizing unigram entropy. The details of this two-step process can be found in Appendix A.9. The selected sentence position distribution using this method is shown in Fig. 4. We can see that the extractor tends to select sentences from the beginning of a paper as well as the ending part of a paper just as the oracle extractor does. This makes sense because the beginning part is the introduction part which talks about the essence of the whole paper and the ending part mostly contains the analysis of experimental results and conclusions etc.

Hybrid Extraction We combine the abstract of a paper and its CE extraction to form a hybrid of both.

4.2.2 Aspect-aware Summarization

Typically in the *extract-then-generate* paradigm, we can just use the extractions directly and build a sequence-to-sequence model to generate text. Here, in order to generate reviews with more diverse aspects and to make it possible to interpret the gener-

ated reviews through the lens of their internal structure, we make a step towards a generation framework involving extract-then-generate-and-predict. Specifically, instead of existing aspect-based summarization works that explicitly take aspects as input (Angelidis and Lapata, 2018; Frermann and Klementiev, 2019; Hayashi et al., 2020), we use our annotated aspects (§3.2) as additional information, and design an auxiliary task that aims to predict aspects of generated texts (reviews). Fig. 5 illustrates the general idea of this.

Figure 5: Aspect-aware summarization.

The loss of this model is shown in Eq. 3

$$L = L_{seq2seq} + L_{seqlab} \quad (3)$$

where $L_{seq2seq}$ denotes sequence to sequence loss which is the negative log likelihood of the correct next tokens, and L_{seqlab} denotes sequence labeling loss which is the negative log likelihood of the correct labels of next tokens. is a hyper-parameter (= 0:1) that is tuned to maximize aspect coverage on the development set.

5 Experiment

In this section, we investigate using our proposed review generation systems with state-of-the-art pre-trained models, to what extent can we realize desiderata of reviews that we de ned in §2.2. We approach this goal by two concrete questions: (1) What are review generation systems (not) good at? (2) Will systems generate biased reviews?

5.1 Settings

Here we consider three extraction strategies in §4.2.1 as well as two generation frameworks, one is the vanilla sequence to sequence model, the other is jointly sequence to sequence and sequence labeling.

Dataset We use our constructed dataset ASAP-Review described in §3 to conduct experiments. For each paper, we use full text (without Appendix) as source document.[17] And we ltered papers with full text fewer than 100 words since they don't contain enough information for models to learn. For reviews, we only use 100-1024 word reviews[18] for training due to computational ef ciency, which account for 92.57% of all the reviews. This results in 8,742 unique papers and 25,986 paper-review pairs in total, the split of our dataset is shown in Tab. 4.

	Train	Validation	Test
Unique papers	6,993	874	875
Paper-review pairs	20,757	2,571	2,658

Table 4: Data split of ASAP-Review .

Model As mentioned in §4.2, the pre-trained sequence-to-sequence model we used is BART. For all models, we initialized the model weights using the checkpoint: " bart-large-cnn " which is pre-trained on "CNN/DM" dataset (Hermann et al., 2015).[19] For extract-then-generate-and-predict framework, we add another multilayer perceptron on top of the BART decoder, and initialize it with 0.0 mean and 0.02 standard deviation. We use the Adam optimizer(Kingma and Ba, 2014) with a linear learning rate scheduler which increases the learning rate linearly from 0 to 4^{-5} in the rst 10% steps (the warmup period) and then decreases the learning rate linearly to 0 throughout the rest of training steps. We netuned our models on the whole dataset for 5 epochs. We set a checkpoint at the end of every epoch and nally took the one with the lowest validation loss.

During generation, we used beam search decoding with beam size 4. Similarly to training time, we set a minimum length of 100 and a maximum length of 1024. A length penalty of 2.0 and trigram blocking (Paulus et al., 2017) were used as well.

5.2 What are Systems Good and Bad at?

Based on the evaluation metrics we de ned in §2.2, we conduct both automatic evaluation and human

[17] If a paper has more than 250 sentences, we truncate it and take the rst 250 sentences when we do the extraction step.

[18] As measured by BART's subword tokenizer.

[19] We also tried " bart-large-xsum " checkpoint which is pre-trained on "XSUM" dataset (Narayan et al., 2018)", however that results in much shorter reviews, and sentences in it tend to be succinct.

Desiderata		Decisive.	Comprehen.		Justification	Accuracy		Others			
Metric		RACC	ACOV	AREC	INFO	ACON	SACC	R-1	R-2	R-L	BS
HUMAN		30.32	49.85	58.66	97.97	75.67	90.77	–	–	–	–
EXTRACTIVE											
INTRO		–	–	–	–	–	–	38.62	8.84	25.11	29.22
CE		–	–	–	–	–	–	38.56	7.81	25.94	29.11
ABSCE		–	–	–	–	–	–	37.55	8.53	25.85	31.99
EXTRACTIVE+ABSTRACTIVE											
	Aspect										
INTRO	p	-15.38†	50.37	55.52	100.00†	43.78†	83.93	41.39	11.53	38.52	42.29
		-11.54†	51.50	58.24	99.29	32.51	80.36†	41.31	11.41	38.38	42.33
CE	p	-23.08†	62.64†	60.73	99.29	39.17	78.57†	42.37	11.72	39.86	41.78
		-30.77†	63.96†	61.62†	100.00†	34.46†	69.64†	42.27	11.62	39.73	41.71
ABSCE	p	-30.77†	55.37†	58.31	98.21	34.75	92.86	43.11	12.24	40.18	42.90
		-38.46†	56.91†	57.56	98.21	35.21	87.50	42.99	12.19	40.12	42.63

Table 5: Results of the baseline models as well as different aspect-enhanced models under diverse automated evaluation metrics. "BS" represents BERTScore. † denotes that the difference between system generated reviews and human reviews are statistically significant (p-value < 0.05 using 10,000 paired bootstrap resampling (Efron, 1992) tests with 0.8 sample ratio).

evaluation to characterize both reference reviews and generated reviews, aiming to analyze what subtasks of review generation automatic systems can do passably at, and also where they fail. The aspect information in each review is obtained using aspect tagger we trained in §3.2.

Automatic Evaluation Automatic evaluation metrics include Aspect Coverage (ACOV), Aspect Recall (AREC) and Semantic Equivalence (ROUGE, BERTScore). Notably, for each source input, there are multiple reference reviews. When aggregating ROUGE and BERTScore,[20] we take the maximum instead of average. And when aggregating other metrics for human reviews, we take the average for each source document. The results are shown in Tab. 5.

Human Evaluation Metrics that require human labor include Recommendation Accuracy (RACC), Informativeness (INFO), Aspect-level Constructiveness (ACON) and Summary Accuracy (SACC). We select 28 papers from ML/NLP/CV/RL domains. None of these papers are in the training set. Details regarding human judgment are shown in Appendix A.3. The evaluation results are shown in Tab. 5.

Overall, we make the following observations:

5.2.1 Weaknesses

Review generation system will generate non-factual statements for many aspects of the paper assessment, which is a serious flaw in a high-stakes setting.

Lacking High-level Understanding Specifically, when using metrics that require higher level understanding of the source paper like Recommendation Accuracy and Aspect-level Constructiveness, proposed systems achieved much lower performance, with even the smallest gaps between systems and humans being 41.86% for Recommendation Accuracy and 31.89% for Aspect-level Constructiveness compared to reference reviews. This means our systems cannot precisely distinguish high-quality papers from low-quality papers and the evidence for negative aspects is not reliable most of the time.[21]

Imitating Style After careful inspection, we find that some of sentences will appear frequently in different generated results. For example, the sentence "The paper is well-written and easy to follow" appears in more than 90% of generated reviews due to the fact that in

[20] We have used our own custom baseline to rescale BERTScore, details can be found in Appendix A.3.

[21] Although there exist varying degrees of performance differences on RACC and ACon for different systems, we only find one pair of systems perform statistically different on ACon.

the training data, this exact sentence appears in more than 10% of papers. This suggests that the style of generated reviews tend to be influenced by high-frequency sentence patterns in training samples.

Lack of Questioning Generated reviews ask few questions about the paper content, which is an important component in peer reviewing. In the reference reviews, the average number of questions per review is 2.04, while it is only 0.32 in generated reviews.

5.2.2 Advantages

We find that review generation systems can often precisely summarize the core idea of the input paper, and generate reviews that cover more aspects of the paper's quality than those created by human reviewers. Systems with aspect information are also aspect-aware and evidence sensitive as we will discuss below.

Comprehensiveness In terms of Aspect Coverage and Informativeness, our systems can outperform human reviewers by at most 14.11% and 2.03% respectively, suggesting that even reviews from the reviewers may also fall short on our defined criterion regarding comprehensiveness.

Good Summarization Current systems can correctly summarize the contributions of papers most of the time as shown by Summary Accuracy. 4 out of 6 systems can achieve over 80% accuracy and statistical significance tests show that gaps between top 3 systems and human reviewers are not significant. This means that in terms of summarizing the paper content, current systems can achieve comparable performance to human reviewers.

5.2.3 System Comparisons

We also look into how systems with different settings are diverse in performance and make the following observations.

Summarization Paradigms By looking at ROUGE (R) and BERTScore (BS), we can see that "extractive + abstractive"-based methods can consistently outperform pure extractive methods, with the smallest gaps of 2.69, 2.57, 12.44, 9.72 for R-1, R-1, R-L and BS respectively. This demonstrates the necessity of using abstractive summarization which can generate reviews that are close both in meaning as well as language use to human reviews.

Extraction Strategies We can see that it is more effective to use extracted text from the full paper to aid the generation process, resulting in higher aspect coverage compared with solely using introduction information. This is reasonable since models can obtain more diverse input from the full text.

System Diagnosis Our fine-grained evaluation metrics enable us to compare different systems and interpret their relative merits. For example, as discussed before, our systems can achieve higher Informativeness than reference reviews while suffering from much lower Aspect-level Constructiveness. This means that if we want our systems to match the performance of real reviewers, we should focus on improving the constructiveness of our systems instead of aiming for methods that provide better evidence for negative aspects (which are not factually correct most of the time in the first place).

5.2.4 Case Study

To get an intuitive understanding of how aspect-enhanced review generation system worked, we perform analysis on a real case. (More analysis can be found in Appendix A.10.) Specifically, since our aspect-enhanced model is trained based on multi-task learning framework, it would be interesting to see how well the tokens are generated associated with corresponding aspects. We take our aspect-enhanced model using CE extraction to conduct this experiment. Tab. 6 shows an example review when we do so.

We can see that the model can not only generate fluent text but also be aware of what aspect it is going to generate as well as the correct polarity of that aspect. Although the generated aspects are often small segments and there are some minor alignment issues, the model is clearly aspect-aware.

5.3 Will System Generate Biased Reviews?

Biases in text are prevalent, but often challenging to detect (Manzoor and Shah, 2020; Stelmakh et al., 2019). For example, in natural language processing, researchers are trying to identify societal biases (e.g, gender) in data and learning systems on different tasks (Bolukbasi et al., 2016; Zhao et al., 2018; Stanovsky et al., 2019). However, previous works on analyzing bias in scientific peer review usually focus on disparities in numerical feedback instead of text. Manzoor and Shah (2020) recently uncover latent bias in peer review based on review

summary	clarity +	substance +	soundness -	substance -

This paper studies the problem of transfer learning from a single pre-trained network onto a new task . The authors propose a method of combining the outputs of multiple pre-trainable classi ers by training on their raw predictions and then ne-tuning on the target samples . The proposed method is based on the use of (non-linear) maximal correlation analysis that originated with Hirschfeld [9] to this problem . The paper is written and easy to follow . The experimental results show that the proposed method outperforms the state-of-the-art methods on the CIFAR-100 , Stanford Dogs , and Tiny ImageNet datasets . However , it is not clear to me how the performance of the method is by the number of target training samples . It would be better the authors can provide some theoretical analysis on the effect of the size of the target dataset .

Table 6: Illustration of generated tokens associated with corresponding aspects. + denotes positive sentiment. - denotes negative sentiment.

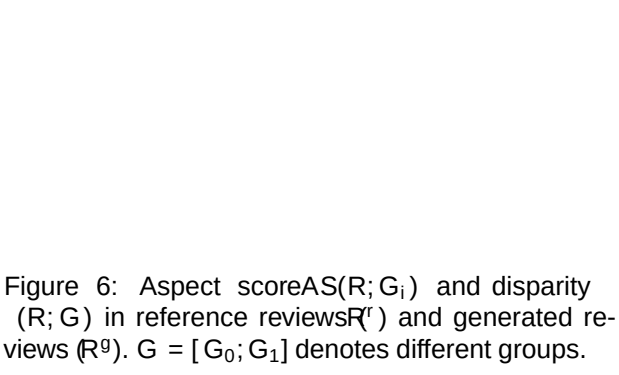

Figure 6: Aspect score $AS(R; G_i)$ and disparity $(R; G)$ in reference reviews (R^r) and generated reviews (R^g). $G = [G_0; G_1]$ denotes different groups.

text. In this work, besides designing a model to generate reviews, we also perform an analysis of bias, in which we propose a method to identify and quantify biases both in human-labeled and system-generated data in a more ne-grained fashion.

5.3.1 Measuring Bias in Reviews

To characterize potential biases existing in reviews, we (i) rst de ne an aspect score, which calculates the percentage of positive occurrences[22] of each aspect. The polarity of each aspect is obtained based on our learned tagger in 3.2; (ii) then we aim to observe if different groups G_i (e.g., groups whether the paper is anonymous during reviewing or is not anonymous) of reviews R would exhibit disparity $(R; G)$ in different aspects. The calculation of disparity can be visualized in Fig. 6.

Based on above two de nitions, we characterize bias in two ways respectively:
(1) spider chart, which directly visualizes aspect scores of different groups of reviews w.r.t each aspect.

[22] If an aspect does not appear in a review, then we count the score for that aspect 0.5 (stands for neutral). Details see Appendix A.11.

(2) disparity difference, which represents the difference between disparities in generated reviews R^g and reference reviews R^r and can be formally calculated as:

$$(R^g; R^r; G) = (R^g; G) \quad (R^r; G) \quad (4)$$

where $G = [G_0; G_1]$ denotes different groups based on a given partition criterion. Positive value means generated reviews favor group G_0 more compared to reference reviews, and vice versa.

In this work, we group reviews from two perspectives. The basic statistics are shown in Tab. 7.

	Native	Non-native	Anonym.	Non-anonym.
Total	651	224	613	217
Acc.%	66.51%	50.00%	57.59%	78.34%

Table 7: Test set statistics based on nativeness and anonymity.

Nativeness We categorize all papers in test set into "native" (G_0) and "non-native" (G_1) based on whether there is at least one native speaker in the author list as well as whether the institution is in an English-speaking country.[23]

Anonymity We categorize all papers in test set into "anonymous" (G_0) and "non-anonymous" (G_1) based on whether the paper has been released as a pre-print before a half month after the conference submission deadline.[24]

[23] We used https://www.familysearch.org/en/ to decide the nationality of an author. In cases where all authors are not from an english-speaking country, we look into the institution information to further decide the attribution of the paper based on whether the institution is from an english-speaking country.

[24] We discard papers from ICLR 2017 since the reviewing process was single blind.