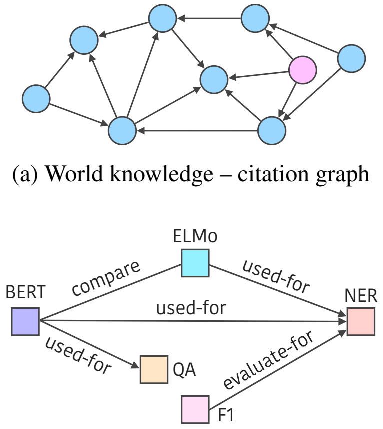
KID-Review: Knowledge-Guided Scientific Review Generation with Oracle Pre-training

Weizhe Yuan, Pengfei Liu

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{weizhey, pliu3}@cs.cmu.edu

Abstract

The surge in the number of scientific submissions has brought challenges to the work of peer review. In this paper, as a first step, we explore the possibility of designing an automated system, which is not meant to replace humans, but rather provide a first-pass draft for a machine-assisted human review process. Specifically, we present an *end-to-end* knowledge-guided review generation framework for scientific papers grounded in cognitive psychology research that a better understanding of text requires different types of knowledge. In practice, we found that this seemingly intuitive idea suffered from training difficulties. In order to solve this problem, we put forward an *oracle pre-training* strategy, which can not only make the KID-REVIEW better educated but also make the generated review cover more aspects. Experimentally, we perform a comprehensive evaluation (human and automatic) from different perspectives. Empirical results have shown the effectiveness of different types of knowledge as well as *oracle pre-training*. We make all code, relevant datasets available: https://github.com/yyy-Apple/KIDReview as well as the KID-REVIEW system: http://nlpeer.reviews.



(a) World knowledge – citation graph

(b) Temporary knowledge – concept graph

Figure 1: Two types of knowledge: citation graph and concept graph. Squares represent concepts, circles represent papers.

Introduction

The rapid growth of research publication not only requires scientists to devote more time to the literature review (Luu et al. 2020; Jha, Abu-Jbara, and Radev 2013; Mohammad et al. 2009; Xing, Fan, and Wan 2020), but brings difficulties to peer review (Yuan, Liu, and Neubig 2021). To address this problem, a small handful of works make a preliminary exploration towards automatic scientific review generation. Wang et al. (2020) perform template-based comment generation for each fine-grained aspect. Yuan, Liu, and Neubig (2021) first answer what the desiderata of a good automatic reviewing system are, and then design an end-to-end auto-review system using current state-of-the-art summarization models.

Despite making a good first step, it is still far from a well-qualified automated reviewing system that can match a human reviewer (Yuan, Liu, and Neubig 2021). Inspired by research in the context of cognitive psychology (Kintsch and Walter Kintsch 1998; Kamide, Altmann, and Haywood 2003; Mumper 2013; Chen et al. 2018), that human comprehend text from (i) *general world knowledge* (long-term memory) (ii) *temporary knowledge* (working memory). We claim that a better understanding of scientific papers also requires these two types of knowledge and operationalize this idea by proposing a knowledge-guided framework for scientific review generation (KID-REVIEW).

Specifically, as shown in Fig. 1, knowledge is incorporated by using diverse graphs, where *concept graph* carries the information of entities (e.g., `method`) associated with their relations (e.g., a method is `used for` a task) for a given paper. By contrast, *citation graph* expresses the whole citation topology within a specific domain. Architecturally, we propose an *end-to-end* framework where a citation graph is first encoded using a large-scale node representation learning algorithm (Tang et al. 2015) and incorporated with the paper content itself. Then we use Graph Neural Networks (Veličković et al. 2017) to represent entities and their interactions within a paper to guide the review generation process.

Practically, to make KID-REVIEW better educated from training data, we propose an *oracle pre-training* strategy. The basic idea is that instead of directly training KID-REVIEW with the whole content of a paper as input, we pre-train it by feeding oracle texts (Nallapati, Zhai, and Zhou 2017), which are sentences from the paper that achieve large lexical overlap with human reviews.[1] We then fine-tune pre-trained KID-REVIEW with different types of paper contents (e.g. introduction) so that during the inference stage, KID-REVIEW does not need to rely on information from human reviews.

Experimentally, we find the *oracle pre-training* strategy not only facilitates the optimization process but also makes generated reviews cover more aspects. Additionally, we observe that using different flavors of knowledge will bring diverse benefits. For example, using citation graphs will help distinguish the paper quality, while introducing concept graphs will lead to more detailed and critical reviews.

Our contributions can be summarized: (1) This is the first work that neuralizes (i.e., end-to-end system) scientific review generation task with different types of knowledge, and present an *oracle pre-training* method to make the parameter optimization more approachable. The work opens the door to this challenging task and connecting it with the latest neural techniques (e.g., BART (Lewis et al. 2020) , GNNs) so that it can enjoy the latest research success. (2) Our work not only shows the complementarity between pre-trained knowledge (e.g., BART) and diverse types of knowledge graphs (e.g., citation graph) for scientific review generation, which could provide a reference for other generation tasks, but also presents how different types of knowledge play different roles. (3) We release our systems and provide a demo service.

Preliminaries

Task Definition

Scientific review generation is conceptualized as an *aspect-based scientific paper summarization* task. Given input paper D, the aim is to generate a review whose high-level objectives are (1) selecting high-quality submissions for publication and (2) improving different aspects of a paper by providing detailed comments (Jefferson et al. 2002; Smith 2006).

Systems & Evaluation Metrics

Systems Existing best-performing systems approach scientific review generation as a two-stage (*extract-then-generate*) summarization problem. Specifically, the first step is to extract salient text pieces from source documents (papers), then generate reviews based on these extracted texts with a state-of-the-art pre-trained sequence-to-sequence model.

Metrics We follow the definition proposed by Yuan, Liu, and Neubig (2021) about what desiderata of a good peer review are: (1) A good review should take a clear stance, selecting high-quality submissions (2) well-organized (3) provide specific reasons for assessment (4) constructive. We brief the core idea of each metric we will use, and detailed formulation could refer to the original paper.

- *Recommendation Accuracy*: Whether the acceptance implied by the review is consistent with the reviewed paper.
- *Aspect Coverage*: How many aspects in a pre-defined typology have been covered in a review.
- *Aspect Recall*: How many aspects in the meta-review of a paper have been covered in a review.
- *Summary Accuracy*: How accurate a review can summarize the core idea of a paper.
- *Constructiveness*: How helpful a review is in terms of pointing out constructive suggestions for paper improvement. Different from the original definition, we use review-level constructiveness in order to rank different systems more conveniently.

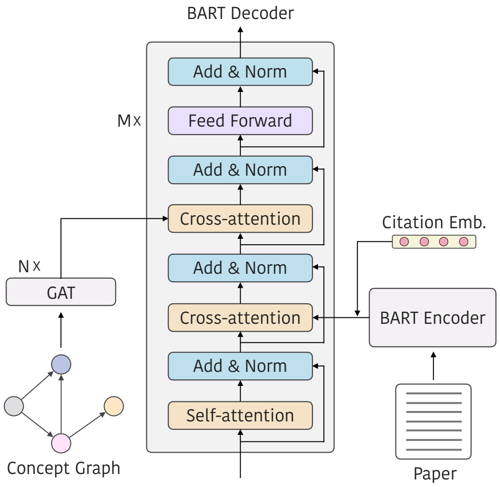


Figure 2: Architecture of our proposed model. N and M denote number of GAT layers and BART decoder layers respectively. “Emb.” is the abbreviation for “Embedding”.

Knowledge-guided Review Generation

Our proposed framework is illustrated in Fig. 2. The backbone of our model is a pre-trained sequence-to-sequence model BART (Lewis et al. 2020) due to its superior performance in text generation.[2] We introduce two types of knowledge into BART through different ways. Citation embeddings are learned through a large-scale node representation learning algorithm and are held fixed during training. Concept graph knowledge is encoded through Graph Attention Network (GAT) (Veličković et al. 2017) and is jointly trained with BART. We detail each knowledge component below.

Concept Graph

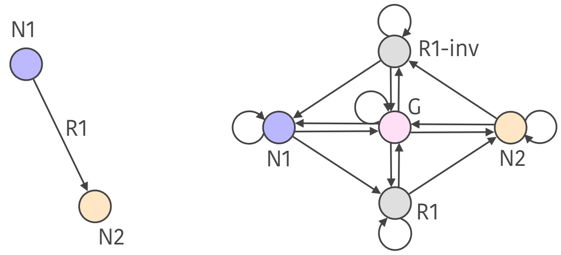
We first introduce how we construct a concept graph for each paper and then detail the graph propagation process.

Graph Construction We define concept graph as $G^{\mathrm{p}} = \{V^{\mathrm{p}}, E^{\mathrm{p}}\}$ where V^{p} stands for nodes one for each entity and E^{p} represents relation edges between entities.

Attributes of Nodes and Edges Specifically, we follow the entity types (*task*, *material*, *method*, *metric*, *generic*, *other scientific term*) and relation types (*part of*, *used for*, *compare*, *feature of*, *hyponym of*, *evaluate for*, *conjunction*) defined in SciERC (Luan et al. 2018) for concept graph construction.

[1] We use the greedy method to get oracle texts as described in (Nallapati, Zhai, and Zhou 2017).

[2] We also explored other pre-trained models like T5 (Raffel et al. 2019) while the performance is worse.



(a) Before transformation.

(b) After transformation.

Figure 3: Restruction of the original concept graph. N denotes an entity node, R denotes a relation node, R-inv denotes an inverse relation node, G denotes a global node.

Edges as Graph Nodes Since the raw entity nodes and relation edges typically cannot form a connected graph, we further adopt the method introduced in Koncel-Kedziorski et al. (2019) to restructure the graph where we convert relation edges into nodes and introduce a global node to connect all nodes. This transformation can be visualized in Fig. 3-(a) and Fig. 3-(b).

Graph Initialization The initial representation for an entity node is obtained using the l lower layers (l is a hyperparameter) of BART encoder as shown in Fig. 4.[3] Specifically, given an entity, we first tokenize it and add a `[BOS]` token as well as a `[EOS]` token, which results in a sequence of tokens $\{t_1, \cdots, t_n\}$ where t_1 is the `[BOS]` token and t_n is the `[EOS]` token. We then use the l lower layers of BART encoder to get the contextualized representations for each token therefore obtaining $\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$, which are the rectangles above BART encoder layers in Fig. 4. Finally, we take $\mathbf{e}_n$ (the rectangle inside a circle), which is the representation learned for `[EOS]` token as the initial entity embedding.

The initializations for relation nodes and global nodes are similar. For a relation node, we encode the descriptive text (Chai et al. 2020) for that specific relation to get its initial representation (e.g. "*is used to evaluate for*" for "*evaluate for*"). For a global node, we encode the title of its associated paper to get the initial representation. Detailed descriptive texts for each relation can be found in Appendix.

Graph Propagation Layer We learn the concept graph representations using Graph Attention Network (GAT). We refer to $\mathbf{e}_i \in \mathbb{R}^d$, $i \in \{1, \cdots, m\}$ as the initial node embeddings in a graph containing m nodes, d is the embedding dimension. We use a multi-head self-attention setup with N attention heads. The updated embedding for node i after going through a GAT layer can be calculated as:

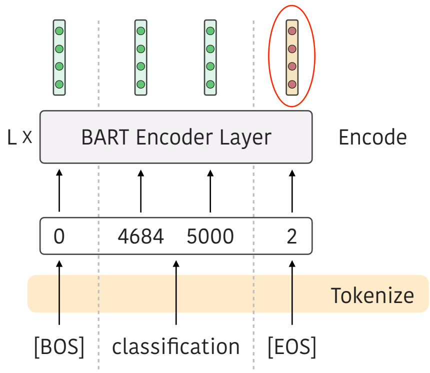


Figure 4: Illustration of an entity node embedding initialization. L denotes the number of BART encoder layers we use. We take the final representation for `[EOS]` token as the entity embedding.

$$\tilde{\mathbf{e}}_i = \mathbf{e}_i + \Big\Vert_{n=1}^{N} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^n \mathrm{W}_v^n \mathbf{e}_j \tag{1}$$

$$\alpha_{ij}^n = \frac{\exp\left(z_{ij}^n\right)}{\sum_{l\in\mathcal{N}(i)} \exp\left(z_{il}^n\right)} \tag{2}$$

$$z_{ij}^n = (\mathbf{W}_q^n \mathbf{e}_i)^\top (\mathbf{W}_k^n \mathbf{e}_j)/\sqrt{d} \tag{3}$$

Where $\Vert$ denotes the concatenation of N attention heads, $\mathcal{N}(\cdot)$ denotes the neighbor nodes of a given node, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are trainable parameters. Following the Transformer architecture (Vaswani et al. 2017), we add a feed-forward network to further enrich the graph representations. The final representation for node i is calculated using Eq. 4.

$$\mathbf{e}_i' = \mathrm{LN}(\mathrm{FFN}(\tilde{\mathbf{e}}_i) + \tilde{\mathbf{e}}_i) \tag{4}$$

$\mathrm{LN}(\cdot)$ denotes layer normalization and $\mathrm{FFN}(\cdot)$ represents feed-forward neural network.

Graph Into BART After getting graph representations, we need to infuse such knowledge into our BART decoder. As shown in Fig.2, the way we do so is to add another cross attention module inside each BART decoder layer to attend to entity representations in our constructed concept graph. We refer to $\mathbf{x}$ as encoded representations for input paper, $\mathbf{y}^l$ as the representations of output in l-th BART decoder layer, $\mathbf{e}$ as the entity representations got from GAT. The $(l+1)$-th decoder layer output is obtained as follows:

$$\tilde{\mathbf{y}}^{l+1} = \mathrm{LN}(\mathbf{y}^l + \mathrm{SelfAttn}(\mathbf{y}^l)) \tag{5}$$

$$\tilde{\mathbf{y}}^{l+1} = \mathrm{LN}(\tilde{\mathbf{y}}^{l+1} + \mathrm{CrossAttn}(\tilde{\mathbf{y}}^{l+1}, \mathbf{x})) \tag{6}$$

$$\tilde{\mathbf{y}}^{l+1} = \mathrm{LN}(\tilde{\mathbf{y}}^{l+1} + \mathrm{CrossAttn}(\tilde{\mathbf{y}}^{l+1}, \mathbf{e})) \tag{7}$$

$$\mathbf{y}^{l+1} = \mathrm{LN}(\tilde{\mathbf{y}}^{l+1} + \mathrm{FFN}(\tilde{\mathbf{y}}^{l+1})) \tag{8}$$

Where $\mathrm{LN}(\cdot)$ denotes layer normalization, $\mathrm{SelfAttn}(\cdot)$ and $\mathrm{CrossAttn}(\cdot)$ represent self-attention module and cross-attention module in BART decoder layer respectively, $\mathrm{FFN}(\cdot)$ denotes feed-forward neural network.

[3] Lower layers of pre-trained language models typically capture more lexical and syntactical information. (Jawahar, Sagot, and Seddah 2019)

	Accept	Reject	# of Reviews
ICLR	1,859	3,333	15,728
NeurIPS	3,685	0	12,391

Table 1: Basic statistics of `ASAP-Review` dataset.

Citation Graph

Graph Construction To construct a citation graph, we use `S2ORC` dataset (Lo et al. 2020) as our knowledge base. It is a large corpus consisting of 81.1M English-language academic papers spanning many academic disciplines.

Graph Representation Learning We select a subset that contains all computer science papers to construct an undirected graph. The citation embeddings for papers are learned using LINE (Tang et al. 2015), which is an efficient algorithm to embed large information networks into low-dimensional vector spaces. Once learned, the citation embedding for each paper is fixed afterward.

Graph Into BART We incorporate citation graph knowledge into BART to enrich the original BART encoder output with the citation embedding of a paper. Formally, we refer to $\mathbf{x}'$ as regular encoder output given a source paper, $\mathbf{c}$ as citation embedding of that paper. The final encoder output $\mathbf{x}$ is $[\mathbf{W}_c\mathbf{c}\|\mathbf{x}']$, where $\mathbf{W}_c$ is a trainable parameter and $\|$ denotes concatenation. The newly concatenated encoder output will be feed into the BART decoder to be further attended.

Oracle Pre-training

Although our proposed system can be directly optimized by feeding input texts and targeted reviews, in practice, we found it challenging to find a satisfying local optimum when training the newly initialized GAT and pre-trained BART together when feeding non-oracle texts. We speculate that this may be caused by the complicated mapping between lengthy input texts to targeted reviews, making it hard to train the knowledge graph component from scratch.

Inspired by the recent idea of oracle guided training (Dou et al. 2020), which has achieved the state-of-the-art performance on the task of summarization, we propose an *oracle pre-training* mechanism, which, (i) engineeringly, ensures a smoothing training process, (ii) experimentally, provides better results w.r.t some evaluation metrics. The basic idea is first to pre-train KID-REVIEW by feeding it with oracle texts (Nallapati, Zhai, and Zhou 2017), which are sentences from the paper with large lexical overlap with human reviews, and then fine-tune systems using different paper contents extracted by diverse strategies (e.g., cross-entropy based methods).

Experiment

Dataset

Peer Review Dataset We use `ASAP-Review` dataset introduced by Yuan, Liu, and Neubig (2021) for our experiment. It consists of ICLR papers from 2017-2020 and NeurIPS papers from 2016-2019, together with their aligned reviews. To make a fair comparison, we use the same training, validation, and test split as them. The basic statistics of this dataset are shown in Tab. 1.

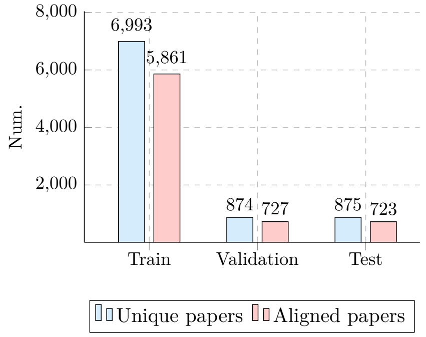


Figure 5: Statistics for paper alignment from `ASAP-Review` dataset to `S2ORC` dataset.

Citation-enriched Peer Review Dataset We align papers to be reviewed in `ASAP-Review` to the `S2ORC` dataset using title matching. The statistics for alignment are shown in Fig. 5. We assign a fixed random vector to a paper that cannot be aligned to the `S2ORC` dataset. During inference time, given a new paper, we take the average of its reference papers' embeddings as its citation embedding.

Setup

Information Extraction over scientific papers To get desired types of entities and relations, we apply the method introduced in Wadden et al. (2019) to extract that information in the abstract section. The reason is that we aim to build a *salient* concept graph, where entities serve for the main idea of the paper to be reviewed (Jain et al. 2020). We collapse co-referential entities into a single entity associated with the longest mention since we assume it to be more informative than others.

Model Settings We initialize BART's parameters using the checkpoint "`bart-large-cnn`" which is pre-trained on "`CNNDM`" dataset (Hermann et al. 2015). We set the embedding size to be 128 when learning citation embeddings. We use two GAT layers for the concept graph, each with 4 attention heads, and we set the hidden size to be 200. To get the initial concept graph embeddings, we set $l = N_{\text{enc}}/2$, where N_{enc} denotes the total number of layers in BART encoder. For each BART decoder layer, we add another cross-attention module to attend to entity node representations on top of the regular cross attention module.

Training Settings Following Yuan, Liu, and Neubig (2021), we adopt the *extract-then-generate* paradigm to deal with lengthy input texts and investigate two extraction strategies, which are (i) extracting sentences to maximize unigram entropy using cross-entropy method (Feigenblat et al. 2017); (ii) combining the abstract part of a paper as well as the extraction in (i). Besides, we also consider oracle extraction for comparison reason, which is the extraction that achieves the highest average ROUGE scores (Lin and Hovy 2003) with respect to reference reviews, specifically using the greedy method described in Nallapati, Zhai, and Zhou (2017). The

	Pre.	Knowledge	Racc	Acov	Arec
Human	–	–	49.25	50.83	58.35
Oracle	–	vanilla	2.40	67.51	65.28
	–	+ citation	**10.06**	68.66	**67.48**
	–	+ concept	6.86	**71.77**	65.74
	–	+ cit.& con.	5.03	67.67	64.09
CE	✗	vanilla	13.94	62.64	60.73
	✓	vanilla	11.43	**67.39**	**62.56**
	✓	+ citation	12.80	66.90	62.49
	✓	+ concept	12.11	62.01	60.85
	✓	+ cit. & con.	**23.31**	61.00	61.99
Abs.+CE	✗	vanilla	15.54	55.37	58.31
	✓	vanilla	17.03	63.47	63.00
	✓	+ citation	21.14	**64.69**	**63.53**
	✓	+ concept	18.06	60.64	59.80
	✓	+ cit. & con.	**25.03**	58.46	60.90

Table 2: Results on automatic evaluation metrics. **Racc**: *Recommendation Accuracy*, **Acov**: *Aspect Coverage*, **Arec**: *Aspect Recall*. "Oracle" represents *oracle pre-training*. "CE" denotes content selection of input papers with cross-entropy method. "Abs." stands for the abbreviation for abstract. The results for vanilla systems without pre-training are taken from Yuan, Liu, and Neubig (2021).

training for systems using oracle extraction is from scratch, while others are fine-tuned based on the pre-trained models using oracle extraction. See Appendix for more details.

Generation Settings We use beam search decoding during generation and adopt the same parameters following Yuan, Liu, and Neubig (2021) for all systems.

Results and Analysis

In all our experiment results, we use the following notations. "cit." and "con." denote citation and concept knowledge. "Pre." stands for *oracle pre-training*.

Automatic & Human Evaluation As mentioned before, we use the following metrics to characterize human-written reviews and system-generated reviews: *Recommendation Accuracy*, *Aspect Coverage*, *Aspect Recall*, *Summary Accuracy* and *Constructiveness*. The former three can be automated using fine-grained aspect information within a review while the latter two require human annotations. We follow the aspect typology introduced by Yuan, Liu, and Neubig (2021) and use their provided aspect tagger to get aspect information within each review. More details can be found in Appendix. Automatic evaluation metrics are performed on `ASAP-Review` test set, the results [4] are shown in Tab. 2.

Overall, we make the following observations: (i) pre-training on oracle texts and then fine-tuning on other input texts can significantly improve *Aspect Coverage* and *Aspect Recall* compared to directly training with other input texts, with the largest improvement 8.1 for *Aspect Coverage* and 4.69 for *Aspect Recall* respectively. (ii) For systems that have been equipped with *oracle pre-training*, using citation graph and concept graph can both achieve consistently higher *Recommendation Accuracy* than vanilla system without knowledge enhancement. The observed largest improvements are 7.66 and 4.46 for adding citation knowledge and concept knowledge, respectively. Besides, the combination of both knowledge can get an even higher *Recommendation Accuracy* boost, at most 11.88. (iii) Training directly based on oracle texts of a paper can reach the highest *Aspect Coverage* and *Aspect Recall* scores, which suggests that it is still valuable to explore more effective content selection strategies when dealing with lengthy source input.

	Vanilla	Vanilla (Pre.)	+ cit.&con. (Pre.)
Sacc	39/40	40/40	39.5/40

Table 3: *Summary Accuracy* for three systems.

However, to better assess the helpfulness of peer reviews, human judgements are necessary. Therefore, we also conduct human evaluation to measure *Summary Accuracy* and *Constructiveness*. We take three systems into comparison: (i) vanilla system without *oracle pre-training* (Yuan, Liu, and Neubig 2021), (ii) vanilla system with *oracle pre-training*, (iii) system equipped with both citation knowledge and concept knowledge, as well as *oracle pre-training*. We select 40 papers from CV/NLP domain that have not been included in the training set and use abstract plus cross-entropy extraction to get system-generated reviews. For each paper, we ask one of the co-authors to annotate the generated reviews.[5] More specifically:

- For *Summary Accuracy*, we ask them to rate the summary part in a review, with a score of 1 denoting agree, 0.5 denoting partially agree, and 0 denoting absent or disagree.
- For *Constructiveness*, we pair the system-generated reviews for each paper and asked the author to give a pair-wise ranking based on how constructive he or she thinks each review is.

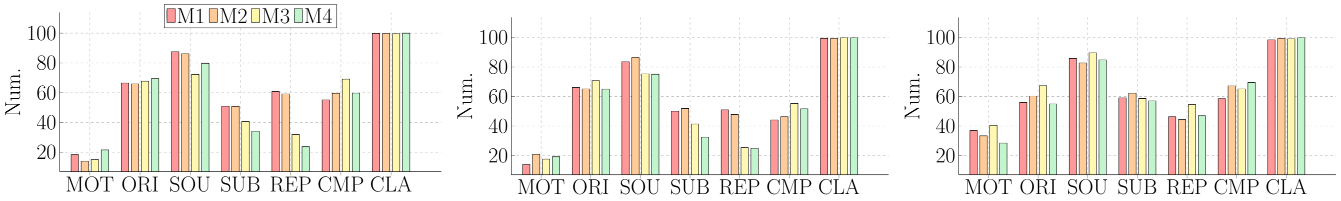
The *Summary Accuracy* for three systems are shown in Tab. 3. All systems can correctly summarize the core idea of given papers almost always. This may be because we have explicitly fed abstract as input text at our extraction stage, which will better guide the summary generation.

The pair-wise comparison results for *Constructiveness* are shown in Tab. 4. By pairwise comparison, the vanilla system without *oracle pre-training* performs worse than its counterpart with *oracle pre-training*, while the system enhanced with knowledge can outperform the vanilla system with *oracle pre-training*. This suggests that adding knowledge can generate more informative and constructive texts.

Fine-grained Analysis Results from the above section present a holistic view of how different knowledge (e.g.,

[4] Samples of generated reviews can be found in Appendix.

[5] There are eight annotators in total, and all of them are Ph.D. students in CV/NLP domain.



(a) CE extraction (Pre.).

(b) Abstract + CE extraction (Pre.).

(c) Oracle extraction.

Figure 6: Fine-grained *Aspect Coverage* for different extraction strategies equipped with different knowledge. M1: vanilla; M2: citation graph; M3: concept graph; M4: citation + concept graph. MOT: Motivation, ORI: Originality, SOU: Soundness, SUB: Substance, REP: Replicability, CMP: Meaningful Comparison, CLA: Clarity. "Pre." denotes *oracle pre-training*.

	Sys.1	**Sys.2**	**Sys.3**
Sys.1	×	47.73	45.45
Sys.2	**52.27**	×	42.86
Sys.3	**54.55**	**57.14**	×

Table 4: Pair-wise comparisons for three systems. Sys.1 represents vanilla system without *oracle pre-training*. Sys.2 represents vanilla system with *oracle pre-training*. Sys.3 represents citation graph knowledge and concept graph knowledge enhanced system with *oracle pre-training*. Each (i, j) entry in the table means the percentage of times system i is preferred than system j.

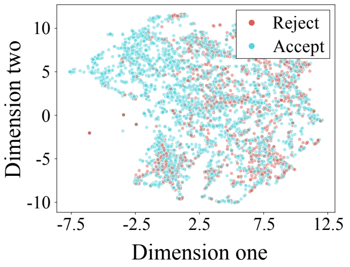


Figure 7: Citation embeddings for accepted/rejected papers using T-SNE visualization.

citation graph) and extraction strategies (e.g., CE) influence KID-REVIEW's performance w.r.t different evaluation metrics (e.g., *Aspect Coverage*). To better understand their interplay, we propose to conduct a fine-grained analysis. Specifically, we adopt the metric "*Aspect Coverage*" as a case study and break down the holistic result into different groups based on aspects. As shown in Fig.6, we find: (1) No matter which extraction strategy has been used, introducing knowledge such as citation graph, concept graph, or both of them can consistently improve the *Aspect Coverage* of Meaningful Comparison. (2) However, the influence of external knowledge on other aspects is variable, depending on which extraction strategy has been adopted. These observations suggest a potential future direction on making a better combination of different types of knowledge and extraction strategies.

Knowledge Understanding Besides holistic and fine-grained evaluation, in this section, we aim to understand how different types of knowledge work in KID-REVIEW.

Citation graph From Tab. 2, the improvements on *Recommendation Accuracy* are consistent by adding citation graph. To explore the potential reasons, we use T-SNE visualization (Van der Maaten and Hinton 2008) to understand the underlying citation embedding space. Specifically, The plot is shown in Fig. 7, red dots represent rejected papers while blue dots denote accepted papers. It is clear that certain region contains more accepted (rejected) papers (e.g., the upper left region contains almost exclusively accepted papers.). Therefore, providing citation embeddings would suggest information about the quality of a paper, thus helping the system distinguish papers of different quality. Intuitively, the citation graph will place a paper within context. If a paper has not cited its most relevant papers, it probably lacks many necessary comparisons with prior works.

Concept graph Based on human judgments for *Constructiveness*, reviews with more specific details are considered to be more constructive. We speculate that with the addition of the concept graph, a model can generate more detailed and specific reviews due to its awareness of salient entities and their relations. To understand how a concept graph would generate more informative reviews, we characterize the generated reviews by looking at how frequently certain words or phrases appear. This is performed on `ASAP-Review` test set using oracle extraction and the results are shown in Tab. 5.

It is evident that by adding a concept graph, the generated

	Vanilla	**+ cit.**	**+ con.**
`for example`	615	616	**680**
`e.g.`	740	757	741
`such as`	255	261	**282**
`for instance`	294	294	**394**
`should compare`	90	115	**170**
`questions`	22	25	**38**
`?`	378	347	**411**

Table 5: The frequency of certain words/phrases in reviews from different systems.

	Vanilla	+ cit.	+ con.	+ cit.&con.
CE	51.37	85.79	76.53	**50.82**
Abs.+CE	61.85	66.17	56.70	**54.01**

Table 6: Total disparity difference between generated reviews and reference reviews in terms of native bias. All systems have been oracle pre-trained. "CE" denotes content selection of input papers with cross-entropy method. "Abs." is the abbreviation for abstract.

reviews are more likely to give specific examples and are more prone to ask questions. These may account for the better review-level constructiveness observed in Tab. 4.

Bias Analysis Here we also conduct bias analysis to see if adding different knowledge will result in a bias for certain groups. We consider bias analysis regarding nativeness, which measures whether there is at least one native English speaker in the author list. We split the papers in the test set into "native" and "non-native". We follow the *aspect score* and *disparity difference* defined by Yuan, Liu, and Neubig (2021) to characterize the bias of different systems. *Aspect score* is the percentage of positive occurrences for a specific aspect, and *disparity difference* measures the system bias compared to human reviewers.

In particular, we look at disparity difference which measures the distance between system preferences and human preferences for certain groups. Here we consider two extraction strategies which are (i) cross-entropy extraction and (ii) abstract part of a paper plus cross-entropy extraction. The results are shown in Tab. 6. Although adding citation graph knowledge and concept graph knowledge individually may not result in smaller disparity difference to human reviews, adding both will consistently result in smaller disparity difference to human reviews. This also demonstrates the complementarity between different knowledge.

Related Work

Knowledge-guided Text Generation For text generation tasks, knowledge beyond the input sequence is often required to produce informative output text. Researchers have tried to incorporate different flavours of knowledge to guide text generation, including topic information (Wei et al. 2019b; Xu et al. 2020), keywords (Wei et al. 2019a; Li et al. 2020), linguistic features (Zhou et al. 2017; Dong et al. 2020), knowledge base (Yang et al. 2019; Feng et al. 2020), knowledge graph (Guan, Wang, and Huang 2019; Huang, Wu, and Wang 2020), etc. Benefits of incorporating knowledge into text generation have been observed in different tasks. For example, it can greatly alleviate hallucination problem in abstractive summarization (Zhu et al. 2020), generating more appropriate and informative responses in conversation generation (Zhou et al. 2018), etc. In our work, we consider two types of knowledge for scientific review generation: citation graph and concept graph.

Peer Review Peer review is an essential component in research community and has been studied from multiple perspectives including bias analysis (Tomkins, Zhang, and Heavlin 2017; Stelmakh, Shah, and Singh 2019), aspect-based sentiment analysis (Chakraborty, Goyal, and Mukherjee 2020), decision classification (Kang et al. 2018; Qiao, Xu, and Han 2018), automatic review generation (Wang et al. 2020; Yuan, Liu, and Neubig 2021). Relevant dataset includes `PeerRead` (Kang et al. 2018) and `ASAP-Review` (Yuan, Liu, and Neubig 2021). Our work extends Yuan, Liu, and Neubig (2021) and provide a novel framework for incorporating external knowledge into pre-trained models. As far as we know, this is the first work that proposes an end-to-end knowledge-fused system for scientific review generation.

Implications and Future Directions

More Nuanced General World Knowledge In this work, we only use a single citation embedding for each paper to incorporate domain background knowledge. It has been proven to work in terms of distinguishing papers of different quality as well as detecting more missing comparisons. However, our systems still suffer from constructiveness due to factuality errors. If a system can understand the more fine-grained relationships between papers (e.g., paper A is a combination of existing work B and C), then it can better judge the novelty of submission and give more constructive comments.

Connecting Text Editing Research with Scientific Review Generation Text editing (Iso, Qiao, and Li 2020), as exemplified as grammar error correction (Ng et al. 2014; Dong et al. 2019), has been studied in different settings. We claim that editing text towards grammatically correct descriptions is crucial for a high-quality scientific review generation system. For example, although our current systems can generate descriptions like "There is a typo in the abstract.", these claims are usually not factual since current systems do not have the sufficient ability to judge the quality of the text, which, however, matters for the evaluation of "Clarity" aspect.

Ethics Statement

We discuss ethical issues from the following aspects:

Intended Use If the system is functioning as intended, both reviewers and paper authors could benefit since our model aims to make research papers better by generating informative comments.

Failure Modes While our system may be helpful in some cases, it is not a replacement for a skilled human reviewer. Completely relying on it will result in unfair reviews since, based on our observations, there are still many factually incorrect comments being generated.

Biases Biases commonly exist in peer reviews (Manzoor and Shah 2020). In this work, we have quantified biases in generated reviews and found that adding knowledge graphs will lead to lower total disparity difference. Moreover, based on some evidence that bias even exists in human reviews (Manzoor and Shah 2020), we believe the advantage of a review generation system is: reviews can be given in a more controllable way. For example, quantify biases of generated reviews first and then (i) either filter biased systems (ii) or biased aspects (e.g., originality).

References

Chai, D.; Wu, W.; Han, Q.; Wu, F.; and Li, J. 2020. Description based text classification with reinforcement learning. In *International Conference on Machine Learning*, 1371–1382. PMLR.

Chakraborty, S.; Goyal, P.; and Mukherjee, A. 2020. Aspect-based Sentiment Analysis of Scientific Reviews. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*.

Chen, X.; Yang, W.; Ma, L.; and Li, J. 2018. Integration of World Knowledge and Temporary Information about Changes in an Object's Environmental Location during Different Stages of Sentence Comprehension. *Frontiers in psychology*, 9: 211.

Dong, X.; Yu, W.; Zhu, C.; and Jiang, M. 2020. Injecting Entity Types into Entity-Guided Text Generation. *arXiv preprint arXiv:2009.13401*.

Dong, Y.; Li, Z.; Rezagholizadeh, M.; and Cheung, J. C. K. 2019. EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3393–3402. Florence, Italy: Association for Computational Linguistics.

Dou, Z.-Y.; Liu, P.; Hayashi, H.; Jiang, Z.; and Neubig, G. 2020. GSum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Feigenblat, G.; Roitman, H.; Boni, O.; and Konopnicki, D. 2017. Unsupervised Query-Focused Multi-Document Summarization Using the Cross Entropy Method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, 961–964. New York, NY, USA: Association for Computing Machinery. ISBN 9781450350228.

Feng, X.; Sun, Y.; Qin, B.; Gong, H.; Sun, Y.; Bi, W.; Liu, X.; and Liu, T. 2020. Learning to Select Bi-Aspect Information for Document-Scale Text Content Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7716–7723.

Guan, J.; Wang, Y.; and Huang, M. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6473–6480.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, 1684–1692.

Huang, L.; Wu, L.; and Wang, L. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. *arXiv preprint arXiv:2005.01159*.

Iso, H.; Qiao, C.; and Li, H. 2020. Fact-based Text Editing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 171–182. Online: Association for Computational Linguistics.

Jain, S.; van Zuylen, M.; Hajishirzi, H.; and Beltagy, I. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7506–7516. Online: Association for Computational Linguistics.

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. Florence, Italy: Association for Computational Linguistics.

Jefferson, T.; Alderson, P.; Wager, E.; and Davidoff, F. 2002. Effects of editorial peer review: a systematic review. *Jama*, 287(21): 2784–2786.

Jha, R.; Abu-Jbara, A.; and Radev, D. 2013. A System for Summarizing Scientific Topics Starting from Keywords. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 572–577. Sofia, Bulgaria: Association for Computational Linguistics.

Kamide, Y.; Altmann, G. T.; and Haywood, S. L. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1): 133–156.

Kang, D.; Ammar, W.; Dalvi, B.; van Zuylen, M.; Kohlmeier, S.; Hovy, E.; and Schwartz, R. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. *ArXiv*, abs/1804.09635.

Kintsch, W.; and Walter Kintsch, C. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.

Koncel-Kedziorski, R.; Bekal, D.; Luan, Y.; Lapata, M.; and Hajishirzi, H. 2019. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ArXiv*, abs/1910.13461.

Li, H.; Zhu, J.; Zhang, J.; Zong, C.; and He, X. 2020. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8196–8203.

Lin, C.-Y.; and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 150–157.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. Online: Association for Computational Linguistics.

Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 2018. Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Luu, K.; Koncel-Kedziorski, R.; Lo, K.; Cachola, I.; and Smith, N. A. 2020. Citation Text Generation. *arXiv preprint arXiv:2002.00317*.

Manzoor, E.; and Shah, N. B. 2020. Uncovering latent biases in text: Method and application to peer review. *arXiv preprint arXiv:2010.15300*.

Mohammad, S.; Dorr, B.; Egan, M.; Hassan, A.; Muthukrishan, P.; Qazvinian, V.; Radev, D.; and Zajic, D. 2009. Using Citations to Generate surveys of Scientific Paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 584–592. Boulder, Colorado: Association for Computational Linguistics.

Mumper, M. L. 2013. The role of world knowledge and episodic memory in scripted narratives. *Psychology Honors Projects*, 32.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. *ArXiv*, abs/1611.04230.

Ng, H. T.; Wu, S. M.; Briscoe, T.; Hadiwinoto, C.; Susanto, R. H.; and Bryant, C. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, 1–14. Baltimore, Maryland: Association for Computational Linguistics.

Qiao, F.; Xu, L.; and Han, X. 2018. Modularized and attention-based recurrent convolutional neural network for automatic academic paper aspect scoring. In *International Conference on Web Information Systems and Applications*, 68–76. Springer.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Smith, R. 2006. Peer Review: A Flawed Process at the Heart of Science and Journals. *Journal of the Royal Society of Medicine*, 99: 178 – 182.

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On testing for biases in peer review. In *Advances in Neural Information Processing Systems*, 5286–5296.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, 1067–1077.

Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48): 12708–12713.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wadden, D.; Wennberg, U.; Luan, Y.; and Hajishirzi, H. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Wang, Q.; Zeng, Q.; Huang, L.; Knight, K.; Ji, H.; and Rajani, N. F. 2020. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. In *Proceedings of INLG*.

Wei, W.; Liu, J.; Mao, X.; Guo, G.; Zhu, F.; Zhou, P.; and Hu, Y. 2019a. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1401–1410.

Wei, X.; Hu, Y.; Xing, L.; Wang, Y.; and Gao, L. 2019b. Translating with bilingual topic knowledge for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7257–7264.

Xing, X.; Fan, X.; and Wan, X. 2020. Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6181–6190. Online: Association for Computational Linguistics.

Xu, M.; Li, P.; Yang, H.; Ren, P.; Ren, Z.; Chen, Z.; and Ma, J. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. *arXiv preprint arXiv:2002.02153*.

Yang, M.; Qu, Q.; Tu, W.; Shen, Y.; Zhao, Z.; and Chen, X. 2019. Exploring human-like reading strategy for abstractive text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7362–7369.

Yuan, W.; Liu, P.; and Neubig, G. 2021. Can We Automate Scientific Reviewing? *arXiv preprint arXiv:2102.00176*.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, 4623–4629.

Zhou, Q.; Yang, N.; Wei, F.; Tan, C.; Bao, H.; and Zhou, M. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, 662–671. Springer.

Zhu, C.; Hinthorn, W.; Xu, R.; Zeng, Q.; Zeng, M.; Huang, X.; and Jiang, M. 2020. Boosting factual correctness of abstractive summarization with knowledge graph. *arXiv preprint arXiv:2003.08612*.