# KID-REVIEW: Knowledge-GuIdeD Scientific Review Generation with Oracle Pre-training

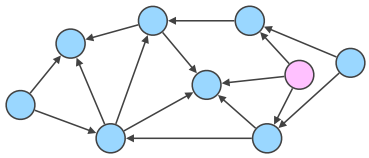

## Abstract

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 providing 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.

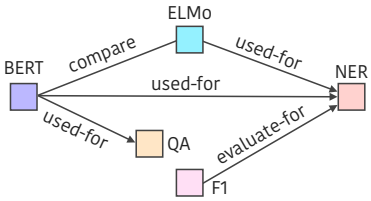
Experimentally, we perform a comprehensive evaluation (human and automatic) from different perspectives. Empirical results have shown the effectiveness of different types of knowledge as well as *oracle pre-training*. We make all code, relevant dataset available: https://github.com/Anonymous4nlp233/KIDReview as well as the KID-REVIEW system: https://nlpeer.reviews.


## 1 Introduction

The rapid growth of research publication not only requires scientists to devote more time to the literature review (Luu et al., 2020; Jha et al., 2013; Mohammad et al., 2009; Xing et al., 2020), but brings difficulties to peer review (Yuan et al., 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 et al. (2021) first



(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.

answer what the desiderata of a good automatic reviewing system are, and then design an end-to-end auto-review system using currently 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 et al., 2021). Inspired by research in the context of cognitive psychology (Kintsch and Walter Kintsch, 1998; Kamide et al., 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` or `task`) 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 Network (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, and the basic idea is instead of directly training KID-REVIEW with the whole content of a paper as input, we pre-train it by feeding oracle texts (Nallapati et al., 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 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 §4.3.1, while introducing concept graphs will lead to more detailed and critical reviews §4.3.3.

Our contributions can be summarized: (1) We make the first step towards an end-to-end knowledge-guided scientific review generation systems and present an *oracle pre-training* method to make the parameter optimization more approachable. (2) Our work not only shows the complementarity between pre-trained knowledge (e.g., BART (Lewis et al., 2020)) §4.3.1 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 §4.3.3. (3) We release a citation-aware our systems and provide a demo service.

[1] We use the greedy method to get oracle texts as described in Nallapati et al. (2017).

## 2 Preliminaries

### 2.1 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).

### 2.2 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 et al., 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*: measures whether the acceptance implied by the review is consistent with the reviewed paper.
- *Aspect Coverage*: measures how many aspects in a pre-defined typology have been covered in a review.
- *Aspect Recall*: measures how many aspects in meta-review of a paper have been covered in a review.
- *Summary Accuracy*: measures how accurate a review can summarize the core idea of a paper.
- *Constructiveness*: measures how useful 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.

## 3 Knowledge-guided Review Generation

Our proposed framework is illustrated in Fig. 2. The backbone of our model is a pre-trained

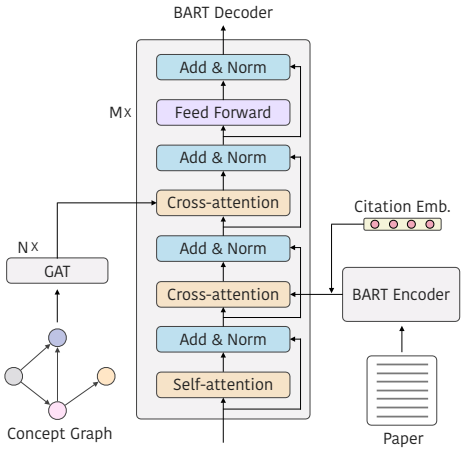
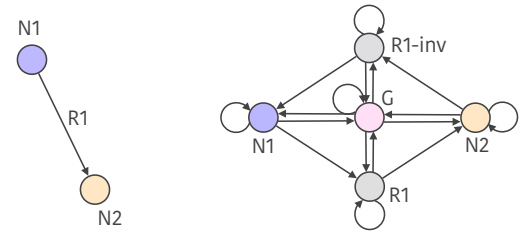



Figure 2: Architecture of our proposed model. N and M denote number of GAT layers and BART decoder layers respectively. "Emb." is the abbreviation for "Embedding".




(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.

sequence-to-sequence model BART (Lewis et al., 2020). We introduce two types of knowledge into BART through different ways. Citation embeddings are pre-trained 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.

## 3.1 Concept Graph

We first introduce how a concept graph for each scientific article is constructed and then detail the graph propagation process.

### 3.1.1 Graph Construction

We define concept graph as $G^{\mathrm{p}} = \{V^{\mathrm{p}}, E^{\mathrm{p}}\}$ where $V^{\mathrm{p}}$ stands for nodes one for each entity and $E^{\mathrm{p}}$ 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 dataset (Luan et al., 2018) for our concept graph construction.

**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.

### 3.1.2 Graph Initialization

The initial representation for an entity node is obtained using the $l$-th lower layers ($l$ is a hyperparameters) of BART encoder as shown in Fig. 4. 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 $l$-th 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 red circle), which is the representation learned for `[EOS]` token as the initial entity embedding.

The initialization 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.

### 3.1.3 Graph Propagation Layer

We learn the concept graph representations using Graph Attention Network (GAT) (Veličković et al., 2017). 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 multihead self-attention setup with $N$ attention heads. The updated embedding for node $i$ after going through a GAT layer can be calculated as:

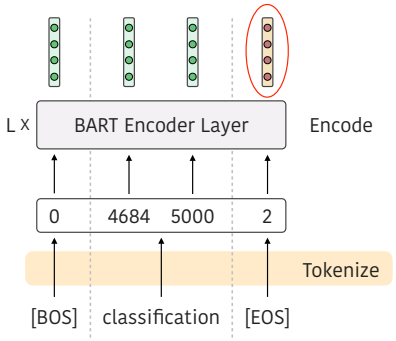



Figure 4: Illustration of an entity node embedding initialization. L denotes number of BART encoder layers we use. We take the final representation for `[EOS]` token as the entity embedding.

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

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

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

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.

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

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

#### 3.1.4 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:

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

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

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

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

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.

### 3.2 Citation Graph

#### 3.2.1 Graph Construction

To construct a citation graph, we use `S2ORC` dataset introduced by 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.

#### 3.2.2 Graph Representation Learning

We formulate the citation graph as an undirected graph, and 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.

#### 3.2.3 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 $[\mathbf{W}_c\mathbf{c}\Vert\mathbf{x}']$, where $\mathbf{W}_c$ is a trainable parameter that converts the citation embedding size to map the hidden size of BART, $\Vert$ denotes concatenation. The newly concatenated encoder output will be feed into the BART decoder to be further attended.

### 3.3 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

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

Table 1: Basic statistics of ASAP-Review dataset.

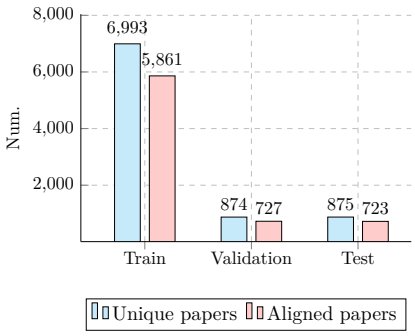



Figure 5: Statistics for paper alignment from ASAP-Review dataset to S2ORC dataset.

may be caused by the complicated mapping between lengthy input texts (>5,000 words) 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 et al., 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).

## 4 Experiment

### 4.1 Dataset

**Peer Review Dataset** We use ASAP-Review dataset introduced by (Yuan et al., 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.

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

### 4.2 Setup

**Information Extraction over scientific papers** To get desired types of entities and relations as mentioned in §3.1.1, we apply the method introduced in Wadden et al. (2019) to extract that information in the abstract section, and 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 Setting** We initialize BART's parameters using the checkpoint "bart-large-cnn" which is pre-trained on "CNN/DM" dataset (Hermann et al., 2015). We set the embedding size to be 128 when learning citation embeddings. For concept graph, we use two GAT layers, 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 $N_{\text{enc}}$ 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** We investigate two extraction strategies introduced by Yuan et al. (2021), 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 highest average ROUGE scores (Lin and Hovy, 2003) with respect to reference reviews, specifically using the greedy method described in Nallapati et al. (2017). The training for systems using oracle extraction is from scratch while others are fine-tuned based on the pre-trained models using oracle extraction. More training details can be found in Appendix.

### 4.3 Results and Analysis

#### 4.3.1 Automatic & Human Evaluation

As mentioned in §2.2, 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 et al. (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 [2] are shown in Tab. 2.

| | 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. **Pre.** denotes whether the system is fine-tuned from *oracle pre-training*. "cit." and "con." stand for abbreviations for citation and concept respectively.

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 that two 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.

However, to better judge the helpfulness of peer reviews, human evaluation is necessary. We also conduct human evaluation to measure *Summary Accuracy* and *Constructiveness*. We take three systems into comparison: (i) vanilla system without *oracle pre-training*, (ii) vanilla system with *oracle pre-training*, (iii) system with both citation knowledge and concept knowledge. We select 40 papers from CV/NLP domains 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. 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, at our extraction stage, we have explicitly feed abstract as input text, which will better guide the summary generation.

| | vanilla | vanilla (Pre.) | +cit.& con. (Pre.) |
|---|---|---|---|
| SACC | 39/40 | 40/40 | 39.5/40 |

Table 3: *Summary Accuracy* for three systems. "cit." and "con." stand for abbreviations for citation and concept respectively. "Pre" stands for *oracle pre-training*.

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

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 beat the vanilla system with *oracle pre-training*. This suggests that adding knowledge can generate more informative and constructive texts.

| | 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 prefered than system $j$.

#### 4.3.2 Fine-grained Analysis

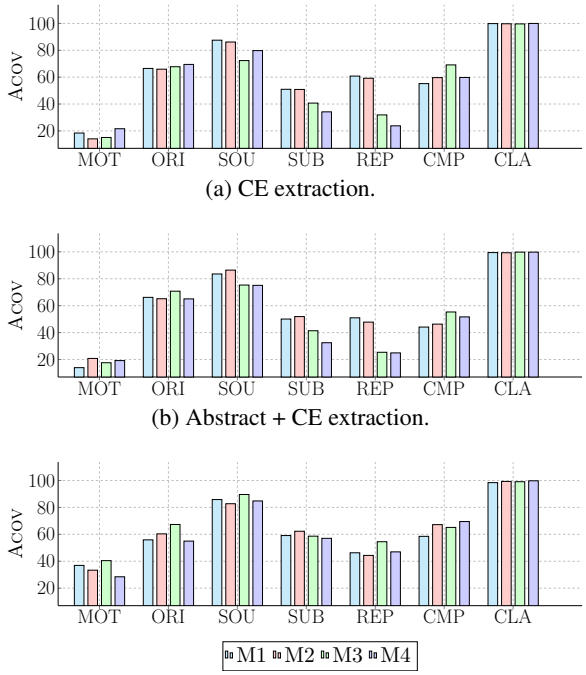
Results from the above section present a holistic view of how different knowledge (e.g., citation graph) and extraction strategies (e.g., CE) influence KID-REVIEW's performance w.r.t different evaluation metrics (e.g., *Aspect Coverage*).

To get a deeper understanding of their interplay, we propose to conduct a fine-grained analysis. Specifically, we adopt the metric "*Aspect Coverage*" as a case study and breakdown 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 (CMP). (2) However, the influence of the introduction of external knowledge on other *Aspect Coverage* 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.

#### 4.3.3 Knowledge Understanding

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



(a) CE extraction.

(b) Abstract + CE extraction.

(c) Oracle extraction.

Figure 6: Fine-grained *Aspect Coverage* (ACOV) 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.

**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.

**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

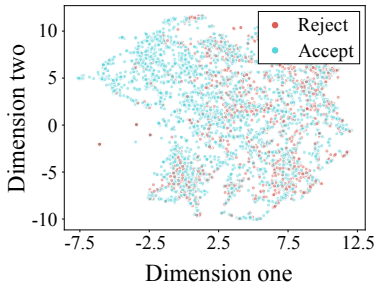



Figure 7: Citation embeddings for accepted/rejected papers using T-SNE visualization.

at how frequently certain words or phrases appear. This is performed on `ASAP-Review` test set using oracle extraction.

| | 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: Certain word/phrase frequency in reviews from different systems.

It is evident that by adding a concept graph, the generated reviews are more likely to give specific examples and are more prone to ask questions. Those may account for the better review-level constructiveness observed in Tab. 4.

## 5 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 (Wang et al., 2018; Narayan et al., 2018; Wei et al., 2019b; Xu et al., 2020), keywords (Mou et al., 2016; Wei et al., 2019a; Li et al., 2020), linguistic features (Zhou et al., 2017; Dong et al., 2020), knowledge base (Eric and Manning, 2017; Bi et al., 2019; Yang et al., 2019; Feng et al., 2020), knowledge graph (Bauer et al., 2018; Liu et al., 2019; Guan et al., 2019; Huang et al., 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 et al., 2017; Stelmakh et al., 2019), aspect-based sentiment analysis (Chakraborty et al., 2020), decision classification (Kang et al., 2018; Qiao et al., 2018), automatic review generation (Wang et al., 2020; Yuan et al., 2021). Relevant dataset includes `PeerRead` by Kang et al. (2018) and `ASAP-Review` by Yuan et al. (2021). Our work extends Yuan et al. (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.

## 6 Implications and Future Directions

**More Nuanced General World Knowledge** In this work, we only use a single citation embedding 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, it still suffers 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 et al., 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 some description like "There is a typo of abstract sentences", these claims commonly are not factually correct 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.

## References


Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*.

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating external knowledge into machine reading for generative question answering. *arXiv preprint arXiv:1909.02745*.

Duo Chai, Wei Wu, Qinghong Han, Fei Wu, and Jiwei Li. 2020. Description based text classification with reinforcement learning. In *International Conference on Machine Learning*, pages 1371–1382. PMLR.

Souvic Chakraborty, P. Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*.

Xuqian Chen, Wei Yang, Lijun Ma, and Jiaxin Li. 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.

Xiangyu Dong, Wenhao Yu, Chenguang Zhu, and Meng Jiang. 2020. Injecting entity types into entity-guided text generation. *arXiv preprint arXiv:2009.13401*.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 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*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Mihail Eric and Christopher D Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *arXiv preprint arXiv:1701.04024*.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 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, page 961–964, New York, NY, USA. Association for Computing Machinery.

Xiaocheng Feng, Yawei Sun, Bing Qin, Heng Gong, Yibo Sun, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. Learning to select bi-aspect information for document-scale text content manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7716–7723.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

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

Hayate Iso, Chao Qiao, and Hang Li. 2020. Fact-based Text Editing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 171–182, Online. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Tom Jefferson, Philip Alderson, Elizabeth Wager, and Frank Davidoff. 2002. Effects of editorial peer review: a systematic review. *Jama*, 287(21):2784–2786.

Rahul Jha, Amjad Abu-Jbara, and Dragomir Radev. 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)*, pages 572–577, Sofia, Bulgaria. Association for Computational Linguistics.

Yuki Kamide, Gerry TM Altmann, and Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1):133–156.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, E. Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *ArXiv*, abs/1804.09635.

Walter Kintsch and CBEMAFRS Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*.

M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, A. Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8196–8203.

Chin-Yew Lin and Eduard Hovy. 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*, pages 150–157.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. *arXiv preprint arXiv:1903.10245*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Kelvin Luu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A Smith. 2020. Citation text generation. *arXiv preprint arXiv:2002.00317*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 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*, pages 584–592, Boulder, Colorado. Association for Computational Linguistics.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*.

Micah L Mumper. 2013. The role of world knowledge and episodic memory in scripted narratives.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *ArXiv*, abs/1611.04230.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

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

R. Smith. 2006. Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99:178 – 182.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2019. On testing for biases in peer review. *arXiv preprint arXiv:1912.13188*.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *arXiv preprint arXiv:1805.03616*.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 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*, pages 1401–1410.

Xiangpeng Wei, Yue Hu, Luxi Xing, Yipeng Wang, and Li Gao. 2019b. Translating with bilingual topic knowledge for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7257–7264.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 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*, pages 6181–6190, Online. Association for Computational Linguistics.

Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. *arXiv preprint arXiv:2002.02153*.

Min Yang, Qiang Qu, Wenting Tu, Ying Shen, Zhou Zhao, and Xiaojun Chen. 2019. Exploring human-like reading strategy for abstractive text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7362–7369.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Boosting factual correctness of abstractive summarization with knowledge graph. *arXiv preprint arXiv:2003.08612*.