ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis

Qingyun Wang[1], **Qi Zeng**[1], **Lifu Huang**[1],
Kevin Knight[2], **Heng Ji**[1], **Nazneen Fatema Rajani**[3]
[1] University of Illinois at Urbana-Champaign [2] DiDi Labs [3] Salesforce Research
{qingyun4,qizeng2,lifuh2,hengji}@illinois.edu
kevinknight@didiglobal.com
nazneen.rajani@salesforce.com

Abstract

To assist human review process, we build a novel *ReviewRobot* to automatically assign a review score and write comments for multiple categories such as novelty and meaningful comparison. A good review needs to be *knowledgeable*, namely that the comments should be constructive and informative to help improve the paper; and *explainable* by providing detailed evidence. *ReviewRobot* achieves these goals via three steps: (1) We perform domain-specific Information Extraction to construct a knowledge graph (KG) from the target paper under review, a related work KG from the papers cited by the target paper, and a background KG from a large collection of previous papers in the domain. (2) By comparing these three KGs, we predict a review score and detailed structured knowledge as evidence for each review category. (3) We carefully select and generalize human review sentences into templates, and apply these templates to transform the review scores and evidence into natural language comments. Experimental results show that our review score predictor reaches 71.4%-100% accuracy. Human assessment by domain experts shows that 41.7%-70.5% of the comments generated by *ReviewRobot* are valid and constructive, and better than human-written ones for 20% of the time. Thus, ReviewRobot can serve as an assistant for paper reviewers, program chairs and authors.[1]

1 Introduction

As the number of papers in our field increases exponentially, the reviewing practices today are more challenging than ever. The quality of peer paper reviews is well-debated across the academic community (Bornmann et al., 2010; Mani, 2011; Sculley et al., 2018; Lipton and Steinhardt, 2019). How many times do we complain about a bad, random, dismissive, unfair, biased or inconsistent peer review? Authors even created various social groups at social media to release their frustrations and anger, such as the "Reviewer #2 must be stopped" group at Facebook[2]. How many times are our papers rejected by a conference and then accepted by a better venue with only few changes? As the number of paper submissions continues to double or even triple every year, so does the need for high-quality peer reviews.

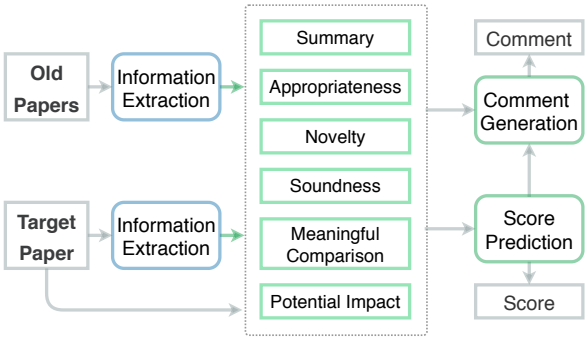


Figure 1: *ReviewRobot* Architecture Overview

The following are two different reviews for the same paper rejected by ACL2019 and accepted by EMNLP2019 without any change on content:

- *ACL 2019*: Idea is too simple and tricky.
- *EMNLP 2019*: The main strengths of the paper lie in the interesting, relatively under-researched problem it covers, the novel and valid method and the experimental results.

These reviews, including the positive ones, are too vague and generic to be helpful. We often see review comments stating a paper is missing references without pointing to any specific references, or criticizing an idea is not novel without showing

[1]The programs, data and resources are publicly available for research purpose at: https://github.com/EagleW/ReviewRobot

[2]https://www.facebook.com/groups/reviewer2/

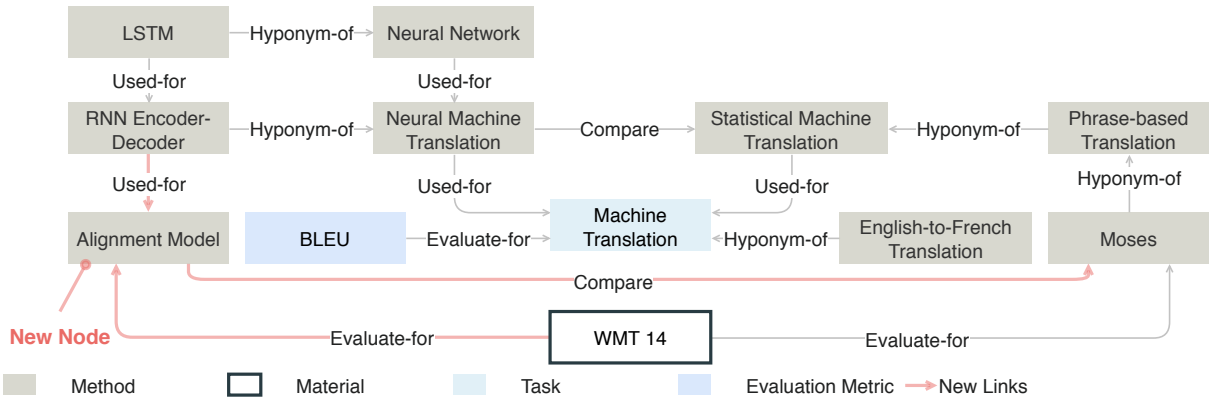


Figure 2: Knowledge Graph Construction Example for Paper (Bahdanau et al., 2015)

similar ideas in previous work. Some bad reviewers often ask to add citations to their own papers to inflate their citation record and h-index, and these papers are often irrelevant or published after the submission deadline of the target paper under review. Early study (Anderson, 2009) shows that the acceptance of a computer systems paper is often random and the dominant factor is the variability between reviewers. The inter-annotator agreement between two review scores for the ACL2017 accepted papers (Kang et al., 2018) are only 71.5%, 68.4%, and 73.1% for substance, clarity and overall recommendation respectively. (Pier et al., 2018) found no agreement among reviewers in evaluating the same NIH grant application. The organizers of NIPS2014 assigned 10% submissions to two different sets of reviewers and observed that these two committees disagreed for 25.9% of the papers (Bornmann et al., 2010), and half of NIPS2016 papers would have been rejected if reviews are done by a different group (Shah et al., 2017).

These findings highlight the subjectivity in human reviews and call for *ReviewRobot*, an automatic review assistant to help human reviewers generate knowledgeable and explainable review scores and comments, along with detailed evidence. We start by installing a brain for *ReviewRobot* with a large-scale background knowledge graph (KG) constructed from previous papers in the target domain using domain-specific Information Extraction (IE) techniques. For each current paper under review, we apply the same IE method to construct two KGs, from its related work section and its other sections. By comparing the differences among these KGs, we extract pieces of evidence (e.g., novel knowledge subgraphs which are in the current paper but not in background KGs) for each review category and use them to predict review scores. We manually select constructive human review sentences and generalize them into templates for each category. Then we apply these templates to convert structured evidence to natural language comments for each category, using the predicted scores as a controlling factor.

Experimental results show that our review score predictor reaches 71.4% overall accuracy on overall recommendation, which is very close to inter-human agreement (72.2%). The score predictor achieves 100% accuracy for both appropriateness and impact categories. Human assessment by domain experts shows that up to 70.5% of the comments generated by ReviewRobot are valid, and better than human-written ones 20% of the time.

In summary, the major contributions of this paper are as follows:

- We propose a new research problem of generating paper reviews and present the first complete end-to-end framework to generate scores and comments for each review category.
- Our framework is knowledge-driven, based on fine-grained knowledge element comparison among papers, and thus the comments are highly explainable and constructive, supported by detailed evidence.
- We create a new benchmark that includes 8K paper and review pairs, 473 manually selected pairs of paper sentences and constructive human review sentences, and a background KG constructed from 174K papers.

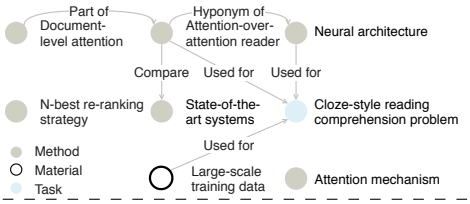
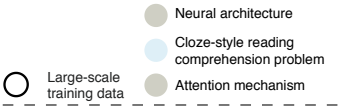
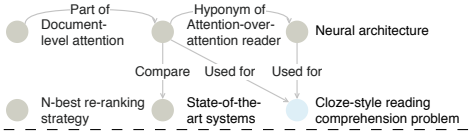
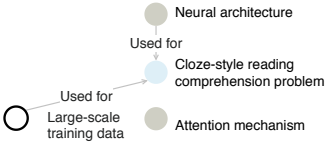
Category	Evidence	Example
Summary	• G_{P_τ}	
Appropriateness	• The number of entities overlapped between the target paper and the domain's background KG: $\lvert\{v \lvert v \in G_{P_\tau} \cap G_B\}\rvert$ • Abstract	
Novelty	• New knowledge elements that appear in the target paper but not in the background KG: $\lvert G_{P_\tau} - G_B \rvert$ • Paper sentences that contain new knowledge elements	
Soundness	• The number of knowledge elements that appear in the contribution claims in the introduction section and that are verified in the experiment section • Abstract	• attention-over-attention reader, n-best re-ranking strategy is verified in the related work section
Meaningful Comparison	• The number of papers about relevant knowledge elements which are missed in the related work section: $(G_B \cap G_{P_\tau}) - \bar{G}_{P_\tau}$ • The number of papers about relevant knowledge elements which are claimed new in the related work section: $G_B \cap G_{P_\tau} \cap \bar{G}_{P_\tau}$ • The description sentences about comparison with related work • If the related work section is not available, we use the difference between G_{P_τ} and G_B instead	 (Bahdanau et al., 2015; Hermann et al., 2015)
Potential Impact	• The number of new knowledge elements in the future work section • The number of new software, systems, data sets, and other resources	• 5 new knowledge elements • 1 new architecture
Overall Recommendations	• All features mentioned in the above categories • Abstract	

Table 1: Evidence Extraction for the example paper *Attention-over-Attention Neural Networks for Reading Comprehension* (Cui et al., 2017)

2 Approach

2.1 Overview

Figure 1 illustrates the overall architecture of *ReviewRobot*. ReviewRobot first constructs knowledge graphs (KGs) for each target paper and a large collection of background papers, then it extracts evidence by comparing knowledge elements across multiple sections and papers, and uses the evidence to predict scores and generate comments for each review category.

We adopt the following most common categories from NeurIPS2019[3] and PeerRead (Kang et al., 2018):

- **Summary**: What is this paper about?
- **Appropriateness**: Does the paper fit in the venue?
- **Clarity**: Is it clear what was done and why? Is the paper well-written and well-structured?
- **Novelty**: Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
- **Soundness**: Can one trust the empirical claims of the paper – are they supported by proper experiments and are the results of the experiments correctly interpreted?
- **Meaningful Comparison**: Do the authors make clear where the problems and methods sit with respect to existing literature? Are the references adequate?
- **Potential Impact**: How significant is the work described? If the ideas are novel, will

[3] https://nips.cc/Conferences/2019/PaperInformation/ReviewerGuidelines

Category	# of Pairs	Evidence Sentence in Paper	Corresponding Review Sentence
Summary	236	In this paper, we present a simple but novel model called **attention-over-attention** reader for better solving cloze-style **reading comprehension** task. (Cui et al., 2017)	The paper describes a new method called **attention-over-attention** for **reading comprehension**.
Novelty	33	The paper presents a new framework to solve the **SR** problem - amortized MAP inference and adopts a pre-learned affine projection layer to ensure the output is consistent with LR. (Sønderby et al., 2017)	It introduces a novel neural network architecture that performs a projection to the affine subspace of valid **SR** solutions ensuring that the high resolution output of the network is always consistent with the low resolution input.
Soundness	174	In high dimensions we empirically found that the **GAN** based approach, AffGAN produced the most visually appealing results. (Sønderby et al., 2017)	Combined with **GAN**, this framework can obtain plausible and good results.
Meaningful Comparison	16	As a concrete instantiation, we show in this paper that we can enable recursive neural programs in the **NPI** model, and thus enable perfectly generalizable neural programs for tasks such as sorting where the original, **non-recursive NPI** program fails. (Cai et al., 2017)	This paper improves significantly upon the original **NPI** work, showing that the model generalizes far better when trained on traces in recursive form.
Potential Impact	14	Since there may be several rounds of questioning and reasoning, these requirements bring the problem closer to task-oriented dialog and represent a significant increase in the difficulty of the challenge over the original **bAbI** (supporting fact) problems. (Guo et al., 2017)	I am a bit worried that the tasks may be too easy (as the **bAbI** tasks have been), but still, I think locally these will be useful.

Table 2: Annotation Statistics and Examples for Template Generalization

they also be useful or inspirational? Does the paper bring any new insights into the nature of the problem?

2.2 Knowledge Graph Construction

Generating meaningful and explainable reviews requires ReviewRobot to understand the knowledge elements of each paper. We apply a state-of-the-art Information Extraction (IE) system for Natural Language Processing (NLP) and Machine Learning (ML) domains (Luan et al., 2018) to construct the following knowledge graphs (KGs):

- G_{P_τ}: A KG constructed from the abstract and conclusion sections of a target paper under review P_τ, which describes the main techniques.
- $\bar{G}_{P_\tau}$: A KG constructed from the related work section of P_τ, which describes related techniques.
- G_B: A background KG constructed from all of the old NLP/ML papers published before the publication year of P_τ, in order to teach *ReviewRobot* what's happening in the field.

Each node $v \in V$ in a KG represents an entity, namely a cluster of co-referential entity mentions, assigned one of six types: *Task*, *Method*, *Evaluation Metric*, *Material*, *Other Scientific Terms*, and *Generic Terms*. Following the previous work on entity coreference for scientific domains (Koncel-Kedziorski et al., 2019), we choose the longest informative entity mention in each cluster to represent the entity. We consider two entity clusters from different papers as coreferential if one's representative mention appears in the other. Each edge represents a relation between two entities. There are seven relation types: *Used-for*, *Feature-of*, *Evaluate-for*, *Hyponym-of*, *Part-of*, *Compare*, and *Conjunction*. Figure 2 shows an example KG constructed from (Bahdanau et al., 2015).

2.3 Evidence Extraction

We compare the differences among the constructed KGs to extract evidence for each review category. Table 1 shows the methods to extract evidence and some examples for each category.

2.4 Score Prediction

Following (Kang et al., 2018), we consider review score prediction as a multi-label classification task. For a target paper, we first encode its category related sentences with an attentional Gated Recurrent Unit (GRU) (Cho et al., 2014; Bahdanau et al., 2015) to obtain attentional contextual sentence embedding. We also encode the extracted evidence for each review category with an embedding layer. Then we concatenate the context embedding and

evidence embedding to predict the quality score r in the range of 1 to 5 with a linear output layer. We use the prediction probability as the confidence score.

2.5 Comment Generation

Given the evidence graphs and predicted scores as input, we perform template-based comment generation for each category. We aim to learn good templates from human reviews. Unfortunately as we have discussed earlier, not all human written review sentences are of high quality, even for those accepted papers. Therefore in order to generalize templates, we need to carefully select those constructive and informative human review sentences that are supported by certain evidence in the papers. To avoid expensive manual selection, we design a semi-automatic bootstrapping approach. We manually annotate 200 paper-review pairs from ACL2017 and ICIR2017 datasets, and then use them as seed annotations to train an attentional GRU (Cho et al., 2014) based binary (select/not select) classifier to process the remaining human review sentences and keep high-quality reviews with high confidence. Our attentional GRU achieves binary classification accuracy 85.25%. Table 2 shows the annotation statistics and some examples.

For appropriateness, soundness, and potential impact categories, we generate generic positive or negative comments based on the predicted scores. For summary, novelty, and meaningful comparison categories, we consider review generation as a template-based graph-to-text generation task. Specifically, for summary and novelty, we generate reviews by describing the *Used-for*, *Feature-of*, *Compare* and *Evaluate-for* relations in evidence graphs. We choose positive or negative templates depending on whether the predicted scores are above 3. We use the predicted overall recommendation score to control summary generation. For related work, we keep the knowledge elements in the evidence graph with a TF-IDF score (Jones, 1972) higher than 0.5. For each knowledge element, we recommend the most recent 5 papers that are not cited as related papers.

3 Experiments

3.1 Data

We choose papers in NLP and ML domains in our experiments because it's easy for us to analyze results, and we are not the most harsh community in Computer Science: the average review score in our corpus is 3.3 out of 5 while it is 2.5/5 in the computer system area (Anderson, 2009). In addition to the review corpus constructed by (Kang et al., 2018), we have collected additional paper-review pairs from openreview[4] and NeurIPS[5]. In total, we have collected 8,110 paper and review pairs as shown in Table 3. We construct the background KG from 174,165 papers from the open research corpus (Ammar et al., 2018). Table 4 shows the data statistics of background KGs.

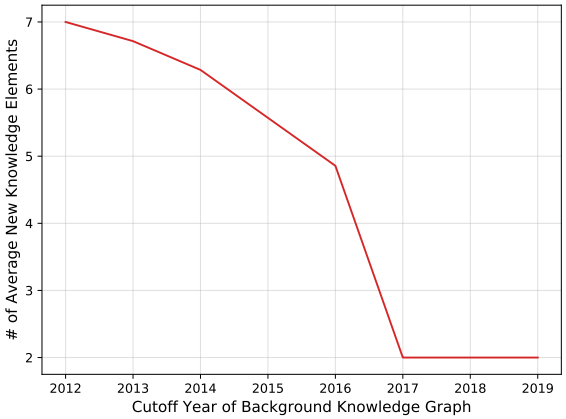


Figure 3: The average number of new knowledge elements in ACL2017 test papers given the background KG constructed from (1965∼cutoff year)

3.2 Score Prediction Performance

We use the ACL2017 dataset in the score prediction task because it has complete score annotations for each review category. We follow the data split of PeerRead (Kang et al., 2018)[6]. Unlike PeerRead which uses multiple review scores for the same input paper, we use the rounded average score of each category as the target score. Table 5 shows that our model trained from carefully selected constructed reviews has already reached a prediction accuracy of 71.43% for overall recommendation, which is very close to the human inter-annotator agreement (72.2%) and dramatically advances state-of-the-art approaches in most categories. Our model also produces the lowest mean square errors for all categories.

Our knowledge graph synthesis based approach is particularly effective at predicting Novelty score

[4]We collect ICLR paper using open review API https://openreview-py.readthedocs.io/

[5]https://papers.nips.cc/

[6]We exclude the training pairs that we fail to run IE system on. The test set remains the same as (Kang et al., 2018).

Conference	Year							
	2013	2014	2015	2016	2017	2018	2019	2020
ICLR	-	-	-	-	404	874	1,342	2,067
NeurIPS	342	399	389	545	655	963	-	-
ACL	-	-	-	-	130	-	-	-

Table 3: Data Statistics for Paper Review Corpus (# of papers)

Years (1965∼)	2011	2012	2013	2014	2015	2016	2017	2018	2019
# of Entities	535,075	585,321	628,713	683,686	737,878	801,740	870,992	950,457	1,008,955
# of Relations	160,123	175,780	188,876	205,898	222,592	242,312	263,827	288,805	307,636

Table 4: Data Statistics for Background Knowledge Graphs since 1965

Category	Human Kappa Score	Human Average Inter Annotator Agreement	CNN (Kang et al., 2018)			GRU with Abstract			GRU with Evidence		
			Score Acc.	Decision Acc.	MSE	Score Acc.	Decision Acc.	MSE	Score Acc.	Decision Acc.	MSE
Recommendation	33.63	72.2	71.43	57.14	0.714	71.43	57.14	0.714	**85.71**	**71.43**	**0.571**
Appropriateness	100	100	**85.71**	**100**	**0.143**	**85.71**	**100**	**0.143**	**85.71**	**100**	**0.143**
Meaningful Comparison	100	100	57.14	57.14	0.857	57.14	71.42	0.857	**71.43**	**71.43**	**0.714**
Soundness	100	100	42.86	42.86	1.86	14.28	**85.71**	0.857	**71.43**	**85.71**	**0.714**
Novelty	100	100	42.86	42.86	2.29	28.57	28.57	2.43	**71.43**	**71.43**	**0.714**
Clarity	70.20	86.11	**42.86**	**71.43**	**1.00**	**42.86**	**71.43**	**1.00**	**42.86**	**71.43**	**1.00**
Potential Impact	100	100	**85.71**	**100**	**0.143**	**85.71**	**100**	0.571	**85.71**	**100**	**0.143**

Table 5: Score Prediction Accuracy (%) and Mean Square Error (MSE) on ACL2017 Data Set

and achieves the accuracy of 71.43%, which is much higher than the accuracy (28.57%) of all other automatic prediction methods using paper abstracts only as input. In Figure 3 we show the average number of new knowledge elements of our test set consisting of ACL2017 papers, when it's reviewed during different years. When the background KG includes newer work, the novelty of these papers decreases, especially after 2017. This indicates that our approach provides a reliable measure for computing novelty.

As a fun experiment, we also run *ReviewRobot* on this paper submission itself. The predicted review scores are 5, 3, 4, 3, 4, 4, and 4 for Appropriateness, Meaningful Comparison, Soundness, Novelty, Clarity, Potential Impact and Overall Recommendation, respectively, which means this paper is likely to be accepted.

3.3 Comment Generation Performance

For the system generated review comments for 50 ACL2017 papers, we ask domain experts to check whether each comment is *constructive* and *valid*. Two researchers independently annotate the reviews and reach the inter-annotator agreement of 92%, 92%, and 82% for Novelty, Summary and Related Work, respectively. One expert annotator performs data adjudication. The percentages of constructive and valid comments are 70.5%, 44.6% and 41.7% for Summary, Novelty and Meaningful Comparison, respectively. Human assessors also find that for 20% of these papers, human reviewers do not suggest missing related work for Meaningful Comparison, while *ReviewRobot* generates constructive and informative comments. For example, the human reviewer states "The paper would be stronger with the inclusion of more baselines based on related work[7]", but fails to provide any useful references. In the following we compare the human and system generated reviews for an example paper (Niu et al., 2017):

Summary

* [SYSTEM] *The paper proposes novel skip-gram, attention scheme, sememe-encoded models and word representation learning for NLP tasks. The authors uses linguistic common-sense knowledge bases.*

* [HUMAN] *This paper proposes the use of HowNet to enrich embeddings.*

[7]Review for (Niu et al., 2017) https://github.com/allenai/PeerRead/blob/master/data/acl_2017/train/reviews/318.json

Novelty

* [SYSTEM] *The paper proposes novel attention scheme and sememe-encoded models.*

* [HUMAN] *I see the main contribution of the work to be the results which show that we can learn better word representations (unsure about wsi) by modeling sememe information than other competitive baselines.*

Meaningful Comparison

* [SYSTEM] *The following related papers are missing:*

1. *About low-dimensional semantic space:*
   (a) *Unsupervised approximate-semantic vocabulary learning for human action and video classification (Zhao and Ip, 2013) Qiong Zhao and Horace HS Ip. 2013. Unsupervised Approximate-semantic Vocabulary Learning for Human Action and Video Classification. Pattern Recognition Letters, 34(15):1870–1878.*
2. *About sememes:*
   (a) *Chinese Word Sense Disambiguation with PageRank and HowNet (Wang et al., 2008): Jinghua Wang, Jianyi Liu, and Ping Zhang. 2008. Chinese Word Sense Disambiguation with PageRank and HowNet. In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.*
   (b) *A maximum entropy approach to HowNet-based Chinese word sense disambiguation (Wong and Yang, 2002): Ping Wai Wong and Yongsheng Yang. 2002. A Maximum Entropy Approach to HowNet-based Chinese Word Sense Disambiguation. In COLING-02: SEMANET: Building and Using Semantic Networks.*
3. *About word similarity and word analogy:*
   (a) *Open IE as an Intermediate Structure for Semantic Tasks (Stanovsky et al., 2015): Gabriel Stanovsky, Ido Dagan, et al. 2015. Open IE as an Intermediate Structure for Semantic Tasks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 303–308.*

* [HUMAN] *The paper would be stronger with the inclusion of more baselines based on related work.*

3.4 Remaining Challenges and Limitations

The quality of *ReviewRobot* is mainly limited by state-of-the-art Information Extraction performance for the scientific literature domain. In the future we plan to annotate more data to cover more dimensions for paper profiling (such as goal and main contribution), and more fine-grained knowledge types to improve the extraction quality. For example, for the NLP domain we can extract finer-grained subtypes: a model can include parameters, components and features. The goal of an NLP paper could belong to: "*New methods for specific NLP problems*", "*End-user applications*", "*Corpora and evaluations*", "*New machine learning methods for NLP*", "*Linguistic theories* ", "*Cognitive modeling and psycholinguistic research*" or "*Applications to social sciences and humanities*". Our current evidence extraction framework also lacks of a salience measure to assign different weights to different types of knowledge elements.

Paper review generation requires background knowledge acquisition and comparison with the target paper content. Our novel approach on constructing background KG has helped improve the quality of review comments on novelty but the KG is still too flat to generate comments on soundness. For example, from the following two sentences in a paper: "Third, at least 93% of time expressions contain at least one time token.", and "For the relaxed match on all three datasets , SynTime-I and SynTime-E achieve recalls above 92%.", a knowledgeable human reviewer can infer 93% as the upper bound of performance and write a comment: "Section 5.2 : given this approach is close to the ceiling of performance since 93 % expressions contain time token , and the system has achieved 92 % recall , how do you plan to improve further?". Similarly, *ReviewRobot* cannot generalize knowledge elements into high-level comments such as "*deterministic*" as in "*The tasks 1-5 are also completely deterministic*".

ReviewRobot still lacks of deep knowledge reasoning ability to judge the soundness of algorithm design details, such as whether the split of data set makes sense, whether a model is able to generalize. *ReviewRobot* is not able to comment on missing hypotheses, the problems on experimental setting

and future work. *ReviewRobot* currently focuses on text only and cannot comment on mathematical formulas, tables and figures.

Good machine learning models rely on good data. We need massive amounts of good human reviews to fuel *ReviewRobot*. In our current approach, we manually select a subset of good human review sentences that are also supported by corresponding sentences in the target papers. This process is very time-consuming and expensive. We need to build a better review infrastructure in our community, e.g., asking authors to provide feedback and rating to select constructive reviews as in NAACL2018[8].

4 Related Work

Paper Acceptance Prediction. Kang et al. (2018) has constructed a paper review corpus, PeerRead, and trained paper acceptance classifiers. Huang (2018) applies an interesting visual feature to compare the pdf layouts and proves its effectiveness to make paper acceptance decision. Ghosal et al. (2019) applies sentiment analysis features to improve acceptance prediction. The KDD2014 PC chairs exploit author status and review comments for predicting paper acceptance (Leskovec and Wang, 2014). We extend these methods to score prediction and comment generation with detailed knowledge element level evidence for each specific review category.

Paper Review Generation. Bartoli et al. (2016) proposes the first deep neural network framework to generate paper review comments. The generator is trained with 48 papers from their own lab. In comparison, we perform more concrete and explainable review generation by predicting scores and generating comments for each review category following a rich set of evidence, and use a much larger data set. Nagata (2019) generates comment sentences to explain grammatical errors as feedback to improve paper writing. (Xing et al., 2020; Luu et al., 2020) extract paper-paper relations and use them to guide citation text generation.

Review Generation in other Domains. Automatic review generation techniques have been applied to many other domains including music (Tata and Di Eugenio, 2010), restaurants (Oraby et al., 2017; Juuti et al., 2018; Li et al., 2019a; Bražinskas et al., 2020), and products (Catherine and Cohen, 2018; Li et al., 2019a; Li and Tuzhilin, 2019; Dong et al., 2017; Ni and McAuley, 2018; Bražinskas et al., 2020). These methods generally apply a sequence-to-sequence model with attention to aspects and attributes (e.g. food type). Compared to these domains, paper review generation is much more challenging because it requires the model to perform deep understanding on paper content, construct knowledge graphs to compare knowledge elements across sections and papers, and synthesize information as input evidence for comment generation.

Controlled Knowledge-Driven Generation. There have been some other studies on text generation controlled by sentiment (Hu et al., 2017), topic (Krishna and Srinivasan, 2018), text style (Shen et al., 2017; Liu et al., 2019a; Tikhonov et al., 2019), and facts (Wang et al., 2020). The usage of external supportive knowledge in text generation can be roughly divided into the following three levels: (1) Knowledge Description, which transforms structured data into unstructured text, such as Table-to-Text Generation (Mei et al., 2016; Lebret et al., 2016; Chisholm et al., 2017; Sha et al., 2018; Liu et al., 2018b; Nema et al., 2018; Wang et al., 2018a; Moryossef et al., 2019; Nie et al., 2019; Castro Ferreira et al., 2019; Wang et al., 2020; Shahidi et al., 2020) and its variants in low-resource (Ma et al., 2019) and multi-lingual setting (Kaffee et al., 2018a,b), Data-to-Document (Wiseman et al., 2017; Puduppully et al., 2019; Gong et al., 2019; Iso et al., 2019), Graph-to-Text (Flanigan et al., 2016; Song et al., 2018; Zhu et al., 2019; Koncel-Kedziorski et al., 2019), and Topic-to-text (Tang et al., 2019), and Knowledge Base Description (Kiddon et al., 2016; Gardent et al., 2017; Koncel-Kedziorski et al., 2019); (2) Knowledge Synthesis, which retrieves knowledge base and organizes text answers, such as Video Caption Generation (Whitehead et al., 2018), KB-supported Dialogue Generation (Han et al., 2015; Zhou et al., 2018; Parthasarathi and Pineau, 2018; Liu et al., 2018a; Young et al., 2018; Wen et al., 2018; Chen et al., 2019; Liu et al., 2019b), Knowledge-guided comment Generation (Li et al., 2019b), paper generation (Wang et al., 2018b, 2019; Cachola et al., 2020) , and abstractive summarization (Gu et al., 2016; Sharma et al., 2019; Huang et al., 2020).

[8] https://naacl2018.wordpress.com/2018/02/26/acceptance-and-author-feedback/

5 Application Limitations and Ethical Statement

The types of evidence we have designed in this paper are limited to NLP, ML or related areas, and thus they are not applicable to other scientific domains such as biomedical science and chemistry. Whether *ReviewRobot* is essentially beneficial to the scientific community also depends on who uses it. Here are some example scenarios where *ReviewRobot* should and should not be used:

- **Should-Do**: Reviewers use *ReviewRobot* merely as an assistant to write more constructive comments and compare notes.
- **Should-Do**: Editors use *ReviewRobot* to assist filtering very bad papers during screening.
- **Should-Do**: Authors use *ReviewRobot* to get initial feedback to improve paper writing such as adding missing references and highlighting the recommended novel points.
- **Should-Do**: Researchers use *ReviewRobot* to perform literature survey, find more good papers and validate the novelty of their papers.
- **Should-Not-Do**: Reviewers submit *ReviewRobot*'s output without reading the paper carefully.
- **Should-Not-Do**: Editors send *ReviewRobot*'s output and make decisions based on it.
- **Should-Not-Do**: Authors revise their papers to fit into *ReviewRobot*'s features to boost review scores. For example, authors should not deliberately cite all related papers or add irrelevant new terms to boost their review scores.

6 Conclusions and Future Work

We build a *ReviewRobot* for predicting review scores and generating detailed comments for each review category, which can serve as an effective assistant for human reviewers and authors who want to polish their papers. The key innovation of our approach is to construct knowledge graphs from the target paper and a large collection of in-domain background papers, and summarize the pros and cons of each paper on knowledge element level with detailed evidence. We plan to enhance *ReviewRobot*'s knowledge reasoning capability by building a taxonomy on top of the background KG, and incorporating multi-modal analysis of formulas, tables, figures, and citation networks.

Acknowledgments

The knowledge extraction and prediction components were supported by the U.S. NSF No. 1741634 and Air Force No. FA8650-17-C-7715. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Thomas Anderson. 2009. Conference reviewing considered harmful. *ACM SIGOPS Operating Systems Review*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 5th International Conference on Learning Representations*.

Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. 2016. Your paper has been accepted, rejected, or whatever: Automatic generation of scientific paper reviews. In *Proceedings of Availability, Reliability, and Security in Information Systems*, pages 19–28. Springer International Publishing.

Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. 2010. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one*, 5(12).

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. Tldr: Extreme summarization of scientific documents. *Computation and Language*, arXiv:2004.15011.

Jonathon Cai, Richard Shin, and Dawn Song. 2017. Making neural programming architectures generalize via recursion. In *Proceedings of the 7th International Conference on Learning Representations*.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.

Rose Catherine and William Cohen. 2018. Transnets for review generation. In *Proceedings of 6th International Conference on Learning Representations Workshop*.

Shuang Chen, Jinpeng Wang, Xiaocheng Feng, Feng Jiang, Bing Qin, and Chin-Yew Lin. 2019. Enhancing neural data-to-text generation models with external background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3022–3032, Hong Kong, China. Association for Computational Linguistics.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 633–642, Valencia, Spain. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602, Vancouver, Canada. Association for Computational Linguistics.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. 2016. Generation from Abstract Meaning Representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego, California. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130, Florence, Italy. Association for Computational Linguistics.

Li Gong, Josep Crego, and Jean Senellart. 2019. Enhanced transformer model for data-to-text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156, Hong Kong. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Xiaoxiao Guo, Tim Klinger, Clemens Rosenbaum, Joseph P Bigus, Murray Campbell, Ban Kawas, Kartik Talamadupula, Gerry Tesauro, and Satinder Singh. 2017. Learning to query, reason, and answer questions on ambiguous texts. In *Proceedings of the 7th International Conference on Learning Representations*.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133, Prague, Czech Republic. Association for Computational Linguistics.

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 systems28*, pages 1693–1701.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1587–1596.

Jia-Bin Huang. 2018. Deep paper gestalt. *Computer Vision and Pattern Recognition Repository*, arXiv:1812.08775.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

Hayate Iso, Yui Uehara, Tatsuya Ishigaki, Hiroshi Noji, Eiji Aramaki, Ichiro Kobayashi, Yusuke Miyao, Naoaki Okazaki, and Hiroya Takamura. 2019. Learning to select, track, and generate for data-to-text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2102–2113, Florence, Italy. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Mika Juuti, Bo Sun, Tatsuya Mori, and N Asokan. 2018. Stay on-topic: Generating context-specific fake restaurant reviews. In *Proceedings of European Symposium on Research in Computer Security*, pages 132–151.

Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018a. Learning to generate Wikipedia summaries for underserved languages from Wikidata. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 640–645, New Orleans, Louisiana. Association for Computational Linguistics.

Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018b. Mind the (language) gap: Generation of multilingual wikipedia summaries from wikidata for articleplaceholders. In *Proceedings of the 15th European Semantic Web Conference*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, Texas. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1697–1705, New Orleans, Louisiana. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Jure Leskovec and Wei Wang. 2014. Data science view of the kdd 2014. In *KDD2014 PC Chair Report*.

Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. 2019a. Generating long and informative reviews with aspect-aware coarse-to-fine decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1969–1979, Florence, Italy. Association for Computational Linguistics.

Pan Li and Alexander Tuzhilin. 2019. Towards controllable and personalized review generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3237–3245, Hong Kong, China. Association for Computational Linguistics.

Wei Li, Jingjing Xu, Yancheng He, ShengLi Yan, Yunfang Wu, and Xu Sun. 2019b. Coherent comments generation for Chinese articles with a graph-to-sequence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4843–4852, Florence, Italy. Association for Computational Linguistics.

Zachary C Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship. *Queue*, 17(1):45–77.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018a. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Tianyu Liu, Fuli Luo, Pengcheng Yang, Wei Wu, Baobao Chang, and Zhifang Sui. 2019a. Towards comprehensive description generation from factual attribute-value tables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5985–5996, Florence, Italy. Association for Computational Linguistics.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018b. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019b. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792, Hong Kong, China. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

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

Shuming Ma, Pengcheng Yang, Tianyu Liu, Peng Li, Jie Zhou, and Xu Sun. 2019. Key fact as pivot: A two-stage model for low resource table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2047–2057, Florence, Italy. Association for Computational Linguistics.

Inderjeet Mani. 2011. Improving our reviewing processes. *Computational Linguistics*, 37(1):261–265.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730, San Diego, California. Association for Computational Linguistics.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Preksha Nema, Shreyas Shetty, Parag Jain, Anirban Laha, Karthik Sankaranarayanan, and Mitesh M. Khapra. 2018. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1539–1550, New Orleans, Louisiana. Association for Computational Linguistics.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711, Melbourne, Australia. Association for Computational Linguistics.

Feng Nie, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. An encoder with non-sequential dependency for neural data-to-text generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 141–146, Tokyo, Japan. Association for Computational Linguistics.

Yilin Niu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Improved word representation learning with sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2049–2058, Vancouver, Canada. Association for Computational Linguistics.

Shereen Oraby, Sheideh Homayon, and Marilyn Walker. 2017. Harvesting creative templates for generating stylistically varied restaurant reviews. In *Proceedings of the Workshop on Stylistic Variation*, pages 28–36, Copenhagen, Denmark. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Elizabeth L Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J Nathan, Cecilia E Ford, and Molly Carnes. 2018. Low agreement among reviewers evaluating the same nih grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

D Sculley, Jasper Snoek, and Alex Wiltschko. 2018. Avoiding a tragedy of the commons in the peer review process. *Computers and Society Repository*, arXiv:1901.06246.

Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. 2018. Order-planning neural text generation from structured data. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike von Luxburg. 2017. Design and analysis of the nips 2016 review process. *Computer Science Repository*, arXiv:1708.09794.

Hamidreza Shahidi, Ming Li, and Jimmy Lin. 2020. Two birds, one stone: A simple, unified model for text generation from structured and unstructured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3864–3870, Online. Association for Computational Linguistics.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems 30*, pages 6830–6841.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. 2017. Amortised map inference for image super-resolution. In *Proceedings of the 7th International Conference on Learning Representations*.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

Gabriel Stanovsky, Ido Dagan, and Mausam. 2015. Open IE as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 303–308, Beijing, China. Association for Computational Linguistics.

Hongyin Tang, Miao Li, and Beihong Jin. 2019. A topic augmented text generation model: Joint learning of semantics and structural features. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5090–5099, Hong Kong, China. Association for Computational Linguistics.

Swati Tata and Barbara Di Eugenio. 2010. Generating fine-grained reviews of songs from album reviews. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1376–1385, Uppsala, Sweden. Association for Computational Linguistics.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. 2019. Style transfer for texts: Retrain, report errors, compare with rewrites. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3936–3945, Hong Kong, China. Association for Computational Linguistics.

Jinghua Wang, Jianyi Liu, and Ping Zhang. 2008. Chinese word sense disambiguation with PageRank and HowNet. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. 2018a. Describing a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 10–21, Tilburg University, The Netherlands. Association for Computational Linguistics.

Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. 2018b. Paper abstract writing through editing mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 260–265, Melbourne, Australia. Association for Computational Linguistics.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Haoyang Wen, Yijia Liu, Wanxiang Che, Libo Qin, and Ting Liu. 2018. Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3781–3792, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. 2018. Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3992–4001, Brussels, Belgium. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Ping Wai Wong and Yongsheng Yang. 2002. A maximum entropy approach to HowNet-based Chinese word sense disambiguation. In *COLING-02: SEMANET: Building and Using Semantic Networks*.

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.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Qiong Zhao and Horace HS Ip. 2013. Unsupervised approximate-semantic vocabulary learning for human action and video classification. *Pattern Recognition Letters*, 34(15):1870–1878.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4623–4629.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better AMR-to-text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5459–5468, Hong Kong, China. Association for Computational Linguistics.