Citation Text Generation

Kelvin Luu[†*], **Rik Koncel-Kedziorski**[†*], **Kyle Lo**[‡], **Isabel Cachola**[‡], and **Noah A. Smith**[†‡]
[†] Paul G. Allen School of CSE, University of Washington, Seattle, WA, USA
[‡] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{kellu,kedzior,nasmith}@cs.washington.edu
{kylel,isabelc}@allenai.org

Abstract

We introduce the task of citation text generation: given a pair of scientific documents, explain their relationship in natural language text in the manner of a citation from one text to the other. This task encourages systems to learn rich relationships between scientific texts and to express them concretely in natural language. Models for citation text generation will require robust document understanding including the capacity to quickly adapt to new vocabulary and to reason about document content. We believe this challenging direction of research will benefit high-impact applications such as automatic literature review or scientific writing assistance systems. In this paper we establish the task of citation text generation with a standard evaluation corpus and develop several strong baseline models. We provide extensive automatic and human evaluations to illustrate the successes and shortcomings of current text generation techniques for this task.

1 Introduction

The output of the world's scientists doubles roughly every nine years (Bornmann and Mutz, 2015), and their pace is quickening. As a result, scientists and other experts must devote significant time to the difficult task of literature review, or coming to understand the context in which they work. Might artificial intelligence help to reduce that time? Several lines of research seek to do so. Citation recommendations systems (Valenzuela et al., 2015; Bhagavatula et al., 2018; Cohan et al., 2019) suggest references to relevant published work for a given document such as a current draft. Summarization systems (Cohan and Goharian, 2015; Yasunaga et al., 2019) condense the information in one or more documents, allowing researchers to more quickly understand the basic ideas in a piece of research.

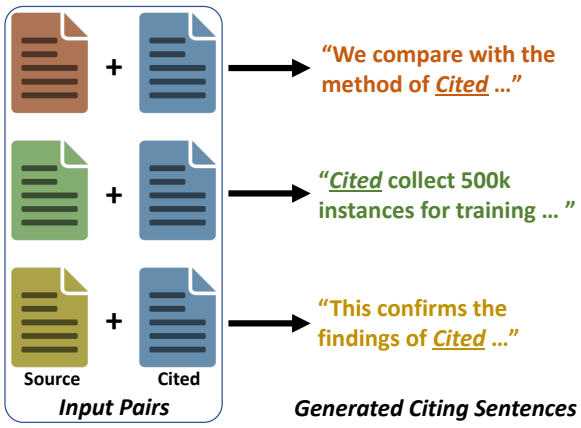

Figure 1: Overview of the citation text generation task. Given two documents, the goal is to write the sentence describing the specific relationship between them. For a given document (in blue above), the output will vary depending the content of the source document that cites it. (This image is best viewed in color.)

We introduce a complementary—but so far unaddressed—problem, citation text generation, where the relationship between a document and one or several others is expressed in natural language text. This differs from traditional summarization in that the primary focus is explaining the *relationship* between the two documents rather than their content alone. Figure 1 illustrates how the same document can be described differently by different referring texts based on the specific relationship of the two documents.

Automatically describing inter-document relationships could dramatically decrease the time researchers devote to literature review. For instance, a new paper could be explained in terms of its relationships to relevant works that a particular reader is most familiar with, rather than just those which the authors elected to cite (personalization). Fur-

[*]Denotes equal contribution.

ther, such technology could be incorporated into writing assistance systems to help less experienced or non-native writers better articulate the connection between their work and prior art. Additionally, users of citation recommendation systems can benefit from natural language explanations of recommendation system choices.

Beyond the immediate utility of citation text generation systems, the task offers interesting challenges for language understanding and generation research. A major challenge is how to represent the information in one or more scientific texts. These documents are significantly longer than those in most other domains typically studied in NLP. In our corpus, the average document length is over 5,000 words. Further, texts in the scientific domain make use of a long-tailed technical vocabulary. This requires a model that can learn phrase meanings from very few exposures, an important but unsolved problem for text generation systems. Possibly more challenging is understanding and expressing the various and nuanced relationships between related scientific papers.

In this work, we introduce the task of citation text generation. Leveraging the full texts of English-language computer science research papers, we construct a dataset of citation sentences for training and evaluating citation text generation models. We investigate strong retrieval and neural baseline models against which future work can compare. Our neural generation models extend the successful GPT2 architecture (Radford et al., 2019) to the scientific domain with additional pre-training and subsequent fine-tuning on the citation generation task. We experiment with different kinds of document context in the fine-tuning and inference stages. We also explore retrieval-based techniques which may more easily generalize to lower-resource settings. These models retrieve citation sentences from training documents which are most similar to test inputs. Our human and automatic evaluations show that these techniques often produce plausible citation sentences, but indicate clear directions for improvement.

2 Task

Citation text generation is the task of generating a natural language *citing sentence* which explains the relationship between two documents. Examples of such citing sentences can be found in scientific documents as in-text citations to a previous work. Thus, we will formally distinguish one document as the *source* document, from which we will draw citing sentences which reference the *cited* document.

This framing suggests a supervised learning setup. Let t denote a citing sentence drawn from source document S, and S' denote S without t. Then let

$$P(t \mid S', C) \tag{1}$$

be the probability of t given S' and the cited document C. A good citation text generation model would maximize this probability across a large number of $\langle t, S, C \rangle$ triples so that at inference time the model is able to generate a sentence t^* which accurately describes the relationship between new documents $\hat{S}$ and $\hat{C}$.

Optimizing Equation 1 is made easier by modern representation learning, including neural text generation systems. However, if we want to leverage these powerful techniques, we are faced with the problem of how to represent the input documents in a way that such models can consume. In particular, language models like GPT2 are trained to predict next token probabilities given long stretches of contiguous text from a single document. It is not clear how to mix information from more than one document (here, S and C) when providing context to these models.

An additional difficulty of the citation text generation task is the vocabulary. Low-frequency, highly meaningful terms regularly appear in human-authored citing sentences. These terms may be completely novel to a single or small collection of papers (consider the phrase "citation text generation", for instance), yet they are necessary for explaining the paper.

A final consideration is evaluation. The most appropriate evaluation metric for most text generation tasks is human judgment by potential users of the system. Evaluating citation text requires human judges with scientific expertise, whose time and effort can be costly. However, as increasingly powerful text generation systems come into existence, it is important that we begin to broach more sophisticated textual domains of consequence such as scientific writing, since automating this complex process could have a transformational impact on researcher productivity. For exploratory purposes, we use the standard automatic metrics for text generation tasks described in Section 4. We also conduct a thorough human evaluation with

	total	average/doc.
documents	154K	–
tokens	813M	5.3K
unique tokens	7.1M	1.3K
citing sentences	622K	4.0
citing sentence length	–	30.3

Table 1: Dataset statistics.

expert judges, and we analyze the relationship of these judgments to the more affordable automatic metrics.

For source and cited documents, we use English-language computer science articles and annotation from the S2-GORC dataset (Lo et al., 2019). S2-GORC is a large citation graph dataset which includes full texts of 8.1 million scientific documents. We select a subset of 154K computer science articles as our corpus. From these, we extract 622K citing sentences that link back to other documents in our corpus. In this work, we focus on citing sentences which contain a single reference. We hold 5000 examples for each of the validation and test sets. Detailed statistics can be found in Table 1.

3 Models

We explore two basic models for citation text generation. Following current work in neural text generation, we fine-tune the predictions of a large pre-trained language model to the citation text generation task (Section 3.1). To help bring the language model into the scientific text domain, we do additional pre-training with a language modeling objective over full scientific texts (Section 3.2). We also investigate approximate nearest neighbor methods to retrieve plausible human-authored citation sentences from the training data (Section 3.3).

3.1 Neural Text Generation

Recent work has shown that adapting large pre-trained language models to text generation tasks yields strong results (Zellers et al., 2019). Therefore, we introduce a model SciGen for citation text generation. SciGen extends the GPT2 model of Radford et al. (2019), a transformer model trained on 40 gigabytes of internet text with a traditional language modeling objective (Vaswani et al., 2017); given a prefix, the model predicts the next token in the sequence. The adaptation process, called *fine-tuning*, involves continued training of the model on the target objective, in our case citation text generation.

To fine-tune GPT2 for text generation, it is typical to concatenate the conditioning context $X = x_1 \dots x_n$ and target sentence $Y = y_1 \dots y_m$ with a special separator token $\mho$. The model learns to approximate next token probabilities for each index after $\mho$:

$$P(y_{i+1} \mid X, \mho, y_1, \dots, y_i) = \text{GPT2}(X, \mho, y_1, \dots, y_i \mid \theta) \quad (2)$$

for $0 < i < m$ and model parameters θ. Cross-entropy loss is calculated for each y_i and back-propagation is used to find parameters θ which maximize $p(y_{i+1} \mid X, \mho, y_1, \dots, y_i)$.

To adapt Equation 2 to the citation text generation task, we construct the conditioning context X from the source and cited documents and use the citing sentence as Y. We take j tokens from source document $s_1, \dots, s_j$ along with k tokens from the cited document $c_1, \dots, c_k$ (which tokens to draw from the two documents is an independent variable that we explore experimentally). We then condition the generation of citing sentence Y on $X = s_1, \dots, s_j, \mho, c_1, \dots, c_k$. This model is trained to predict the citing sentence one token at a time as described above.

At inference the model is provided with an unseen source document and a document cited in the source. The citing sentence is generated one token at a time using greedy decoding. At timestep t, output token $\hat{y}_t$ is the token which maximizes $P(\hat{y}_t \mid X, \mho, \hat{y}_1, \dots, \hat{y}_{t-1})$. The selected $\hat{y}_t$ is used to condition the prediction of subsequent tokens.

Context The primary question we investigate with this model is what kind of input is best for generating accurate and informative citation sentences. Prior works studying the citation recommendation task have made use of abstracts, which perhaps act as sufficient summaries of document content. We also investigate this setting. Additionally, we explore the use of extended contexts such as the introduction or first section after the abstract. Since full scientific texts are too long to fit into the context window of our generation model, we also investigate a "sampling" approach which samples sentences from throughout the document until the context window is full. In this work, we combine either the abstract or introduction of the source document with each of the abstract, introduction, or sampled sentences from the cited document. For all variants, we finetune the underlying language model for an additional 10 epochs,

		BLEU	Rouge-1	Rouge-2	Rouge-L
generation	source abs × cited abs	9.82	0.107	0.006	0.084
	source abs × cited intro	9.39	0.107	0.006	0.084
	source abs × cited sample	9.60	0.107	0.007	0.085
	source intro × cited abs	9.92	0.111	0.010	0.087
	source intro × cited intro	9.80	0.011	0.011	0.088
	source intro × cited sample	9.81	0.109	0.009	0.087
retrieval	source abs × cited abs	9.93	0.142	0.007	0.097
	+ MERT (BLEU)	10.23	0.143	0.007	0.098
	no source × cited abs	9.79	0.141	0.006	0.096

Table 2: Automatic evaluation of generated texts. Statistical significance is discussed in Section 4.

or approximately 100k gradient updates with batch size of 64.[1] We save checkpoints every 10k gradient updates and select the best performing model based on validation perplexity.

3.2 Language Model Pretraining

GPT2-based models have demonstrated an ability to capture long distance dependencies over hundreds of tokens, which we hypothesize will allow them to synthesize information in both the source and cited documents. But citation text generation models must also handle the challenging technical syntax and vocabulary of the scientific domain.

Prior work has shown that pretraining on in-domain data improves the performance of large language models on domain-specific tasks (Beltagy et al., 2019). Inspired by this, we do continued pretraining of the GPT2 model in the science domain to produce SciGPT2, which we use as the underlying language model in SciGen. SciGPT2 starts from the standard pretrained GPT2-base model and is trained for an additional 75k gradient updates at batch size of 64 (effectively a single epoch over 4.8 million abstracts and body paragraphs) with a language modeling objective.[2] We observed significant improvements in the quality of SciGen outputs after replacing the underlying GPT2 language model to the domain-specific SciGPT2 model.

When using pretrained language models in downstream applications, text from task-specific test data cannot be guaranteed to be absent from the large task-independent corpora upon which these models are trained, which may improve model performance compared to models without this exposure. For the experiments described in this work, we train a version of SciGPT2 only on documents appearing in the citation text generation task training data, so that the source documents and citing sentences in the test data are unseen by the language model. We provide both this and full-corpus versions of SciGPT2 as resources for future research.

3.3 Retrieval with Approximate Nearest Neighbors

While neural text generation techniques have advanced significantly in recent years, their outputs are still inferior to human authored texts. For some tasks, it is better to retrieve a relevant human-authored text rather than generating novel text automatically (Fan et al., 2018). Is this also the case for citation text generation?

To answer this question, we adapt an approximate nearest neighbor search algorithm to find similar pairs of documents. The basic search procedure is as follows: Given a test instance input (S, C) for source S and cited document C, we find the set $\mathbf{N}_C$, the nearest neighbors to C in the training data. For each document N_C from $\mathbf{N}_C$, let $\mathbf{N}_S$ be the set of documents that cite N_C. This means that each $N_S \in \mathbf{N}_S$ contains at least one citing sentence t' which cites N_C. We return the t' associated with the (N_S, N_C) pair from the training which is closest to (S, C).

We measure the closeness of two pairs of documents by measuring cosine distances between vector representations of their abstracts. We choose to consider the abstracts of the documents under the assumption that they will summarize document content, and because they generally fit in the contextual window of the pretrained language model

[1] We use a triangular learning rate schedule with 10% warmup and a maximum learning rate of 1e-4.

[2] Learning rate and warmup as above.

we will use to encode them. The abstract of each document is encoded as a single dense vector by averaging the contextualized embeddings provided by the SciBERT model of Beltagy et al. (2019) and normalizing. The distance between (S, C) and candidate (N_S, N_C) is computed as:

$$\alpha \cos(S, N_S) + \beta \cos(C, N_C) \quad (3)$$

where α and β control the relative contribution of the two document similarities. We explore setting both α and β to 1, or tuning them to optimize BLEU on the validation data.

4 Evaluation

4.1 Automatic Evaluation

We compare the different baseline systems using BLEU (Papineni et al., 2002) and ROUGE (specifically ROUGE 1, 2, and L; (Lin, 2004)). Table 2 (above the double line) shows the performance of the SciGen model on the test set when provided with the different input context combinations outlined in Section 3.1. We find that context does make a difference for this category of model, and that a slight performance improvement comes from using the intro of the source document. Automatic evaluation of the retrieval-based methods on the test data is shown below the double line in Table 2. We see that tuning the α and β parameters to optimize BLEU on the validation set does lead to improved performance at test time, but the effect is less pronounced under other metrics. We also evaluate a model which uses only the cited document to retrieve citing sentences, ignoring the source. This model does surprisingly well, indicating the importance of modeling the cited document in this task. We discuss phenomenon in more detail in Section 5.

Statistical significance is assessed for select results using bootstrapping with 1000 samples in each of 100 iterations. This test shows that conditioning on the introduction of the source document improves performance compared to conditioning on the abstract when using the SciGen model. However, we see that IR methods perform better than the best neural models under these metrics.[3] We do not find enough evidence to reject the null hypothesis that any particular representation of the cited document's content (abstract, intro, or random sample) is sufficient.

[3] $p < 0.01$ after Bonferonni correction for both cases.

4.2 Human Evaluation

We conduct a human evaluation of the generated text to determine how *correct*, *specific*, and *plausible* these outputs are. By correct we mean: does the citation sentence correctly express the factual relationship between the source and cited documents? We are also interested in specificity, as generic statements such as "We extend the ideas of Chomsky and Halle (1968)", which may be factual, do not express a detailed understanding of the documents' relationship. We ask judges whether the citing sentence describes a specific relationship between the two works or is vague enough to be used to cite many different papers. A citing sentence can be specific even it is incorrect. Lastly, we define a *plausible* citing sentence as one which could believably fit into the source document. Plausibility judgments give us a measure of how well source topicality and tone is captured without penalizing factual errors about the cited document. To measure plausibility, we show judges a source abstract and citing sentence; for specificity and correctness, judges are shown source and cited documents' abstracts along with a citing sentence.

We compare the *source intro* $\times$ *cited abs* SciGen setting against the untuned IR system. For calibration, we also elicit judgments for the gold citing sentences extracted from source documents along with the correct source and cited abstracts. For each system, we randomly select 50 datapoints from the test set. We collect judgments from 37 NLP researchers with varying levels of expertise, the majority of whom are graduate students. Each judge is given 15 datapoints for each of the plausibility, specificity, and correctness qualities. In order to facilitate expert judgment, we ensure that the source papers of these datapoints appear in the ACL anthology. We ask judges to indicate whether each datapoint does or does not meet the condition, allowing them to skip examples they feel unsure of. In total we collect over 1200 judgments, with over 100 for each system/quality combination.

Table 3 shows the percentage of "yes" judgments for each system/quality combination, along with pairwise agreement rates.That gold texts do not achieve perfect scores demonstrates the limitation of our evaluation setup, due in part to the fact that judgments are based on document abstracts rather than their full texts. Still, we observe the highest scores for all text qualities over the gold text, as well as the highest agreement rate. We

	Plausible	Specific	Correct	*agreement*
IR	68.2	74.8	46.3	*77.5*
SciGen	87.5	72.3	64.0	*70.5*
Gold	89.0	81.4	72.1	*83.8*
agreement	*87.4*	*69.8*	*71.4*	

Table 3: Human evaluation of SciGen and IR systems compared with gold citing sentences (percentages).

can also see that, despite the IR system's strong performance in automatic metrics, it produces implausible and incorrect citing sentences more often than not. A more sophisticated IR system could perhaps achieve better results. The SciGen system performs quite well in this analysis, with a majority of outputs deemed correct and plausibility ratings approaching those of the gold texts. We observe a larger difference in terms of specificity between SciGen and gold texts, indicating that SciGen, like many neural text generation systems, often generates vague and generic sentences.

Correlation with Automatic Metrics To determine if any of the studied automatic metrics correlate with any of the human evaluated text qualities, we conduct a correlation analysis. Specifically, for each of Rouge-1, Rouge-2, Rouge-L, and BLEU we compare the distribution of scores between the outputs of automatic systems and their human judgments of plausibility, specificity, and correctness. For outputs which have been assessed by multiple judges, we include the metric score for that output in the corresponding distribution once per judgment.

To compare the continuous scores of the automatic metrics with categorical human judgements (which are dichotomous nominal variables), we compute a variant of Pearson correlation called point biserial correlation (Lev et al., 1949). We find no significant correlation between any judged text quality and automatic metrics for the IR system. However, we do find a weak correlation between the BLEU scores of SciGen and their human evaluation scores for correctness ($r_{pb} = 0.29, p < 0.01$ after Bonferroni correction). A box plot of the SciGen BLEU scores by correctness category is presented in Figure 2. We can see a difference in the means of these distributions, but the variance of incorrect responses is too great to derive a stronger correlation. Extended human evaluations on correctness may improve the sharpness of this contrast. Further, introducing a domain-specific term weighting to the BLEU metric may increase its ability to act as an approximation to correctness in future model development. These are promising directions for future work.

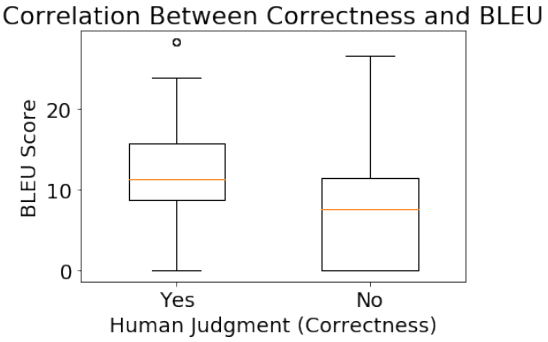

Figure 2: Box plot of BLEU scores by correctness judgment category for SciGen system.

5 Analysis

To test the validity of the human judgements, we conduct an additional evaluation of gold citing sentences paired with different kinds of mismatched inputs: (1) the correct source document and a random cited document, (2) the correct cited document but a random source document (3) random source and cited documents.[4] Conditions 1 and 2 allow us to see whether human judges accept sentences which align with only one or the other of the input documents; condition 3 provides a lower bound. We collect over 107 human evaluations of correctness across these conditions, again allowing annotators to skip datapoints they are unsure of. The results, shown in Table 4, indicate that human judges often will often accept a citing sentence as long as one of the source or cited documents is correct, but not at the rate seen in Table 3 when both documents are correct. There is no indication from this experiment that either the source or cited document is a stronger influence on a judge's correctness decision, although a larger sample size is needed to make a clear determination.

[4]Random documents selected from ACL anthology

	Correct
random cited	45.8
random source	46.9
both random	17.6

Table 4: Correctness judgements of incorrect citing sentences (percentages).

In Table 5, we take a detailed look at some selected plausible example outputs of the SciGen system with different correctness judgments. The first instance is an example showcasing the power of model to accurately depict the different focus of the source and cited works. In the other examples, despite the fact that the generated outputs are incorrect, they are still topical and specific. This phenomenon often occurs when the model output references a dataset. While the dataset would be potentially relevant to both papers, the cited papers focus on modeling contributions and do not introduce a novel corpus.

5.1 Examples

Example system outputs for randomly selected validation instances are shown in Table 6. We see that both the SciGen and IR model outputs regularly hit on the correct broad topic of the cited text (such "literary analysis" or "image captioning evaluation metrics"). It is notable that the SciGen model outputs syntactically correct and coherent citation sentences, even given the difficulty of the vocabulary in this domain. This is a testament to the power of the domain-specific language model training.

We also observe that the outputs of the SciGen model are often shorter than the gold sentences. Brevity is a known issue for neural text generation and may be alleviated by penalizing brevity in the inference procedure. More problematic are the factual errors in the generated text. In the last example, for instance, we see that SciGen fails to cite the specific image captioning dataset described in the cited paper (Pascal1K) and instead focuses on the more general evaluation metric for the image captioning task (CIDEr). This is typical of neural text generation systems, which often assign high probability to generic or frequent phrases and revert to these in the face of uncertainty.

5.2 Future Work

The fluency and appropriateness of the examples in Tables 5 and 6, along with the strong performance of the SciGen baseline in human evaluations, show that generating citing sentences which accurately capture the relationship between two documents should be increasingly possible in the near future. Based on the results obtained in this work, the most promising path forward is the exploration of more sophisticated models based on pretrained scientific language models.

Future work should focus on two complementary goals: ensuring the factual accuracy of the generated text and improved modeling of the cited document. As suggested by the correlation analysis in Section 4, it may be possible to use BLEU during development as a proxy for correctness, especially with some improved weighting scheme. Factual accuracy is difficult to enforce in statistical text generation systems, especially where inference includes sampling procedures. Grounding to knowledge bases could help. For this task, knowledge extracted from candidate generations could be compared with knowledge from the full source and cited documents to prune false or irrelevant statements. Further, modeling input documents as knowledge graphs of their contents may help these algorithms better understand the cited document, resulting in better outputs. However, such a model will have to address the problem of combining pretrained language models with graph encoding techniques, about which little is yet known.

6 Related Work

The current work builds on recent research in scientific document understanding, including citation recommendation and categorization, as well as scientific document summarization.

Citation recommendation, or the task of selecting works related to a source document which would be suitable for citing, is a longstanding goal of AI research (McNee et al., 2002; Bhagavatula et al., 2018; Nallapati et al., 2008). Recently, researchers have sought to categorize citations using various ontologies of citation intents. Valenzuela et al. (2015) sought to discern "highly influential" citations from others. Jurgens et al. (2016) uses six categories including "motivation", "uses", and "future work" among others. Cohan et al. (2019) condense this ontology to just three: "background","method", and "result comparison".

We view the citation text generation task as an extension of these classification approaches with distinct advantages. While classification requires an extant citation link to exist, our generation task can describe possible relationships between works which do not cite each other, such as contemporaneous works. Additionally, because gold citation texts are readily available in scientific documents, the citation text generation task requires no task-specific annotated training data. In practice, citation classification is used to assist in suggesting relevant works to researchers; citation text generation complements this goal by providing rationales for the recommendation and furthering progress toward explainable AI.

Generating a citation is also connected to summarizing scientific documents. There is a long history research on summarizing scientific documents (Luhn, 1958; Paice, 1980). More recently, researchers have included citing sentences as part of the input for summarization, hoping to capture the contribution of a work along with its content (Nakov et al., 2004; Cohan and Goharian, 2017; Yasunaga et al., 2019). Ours is the first to focus on the specific relationship between two documents when generating such sentences. Because of the emphasis on relational document understanding in our task, citation generation models can be used to assist with drafting papers as well, reducing researcher workload and providing non-native writers with a helpful first draft.

Our work builds on recent advances in transfer learning in NLP. In particular, large pretrained models such as BERT (Devlin et al., 2018) and GPT2 (Radford et al., 2019) have made strong advances on a number of tasks (Wang et al., 2019). It has also been shown that pretraining these models on domain-specific data further improves results on domain-specific tasks (Beltagy et al., 2019; Lee et al., 2019). In this work, we apply that methodology by adding an additional pretraining phase on in-domain data before finetuning a GPT2 model on the citation text generation task.

7 Conclusion

We have introduced the challenging but potentially highly impactful task of citation text generation. This task requires reasoning about the relationships between documents and expressing these relationships in natural language text. We have established a dataset for this task and studied the performance of contemporary neural text generation and information retrieval models with automatic and human evaluations. Our analysis shows that a neural text generation model based on a strong domain-pretrained language model produces fluent and topical outputs which are often correct. More research is needed to improve factual accuracy and specificity of the generated text.

Source
This paper investigates the interplay between different types of user interactions on Twitter, with respect to predicting missing or unseen interactions . . . Interestingly, the most predictive features vary with the user profiles, and are not the same across all users. For example, for a pair of users that interact with a large number of other Twitter users, we find that certain"higher-dimensional"triads, i.e., triads that involve multiple types of interactions, are very informative, whereas for less active Twitter users, certain in-degrees and out-degrees play a major role. . . .

Cited
We study online social networks in which relationships can be either positive (indicating relations such as friendship) or negative (indicating relations such as opposition or antagonism). Such a mix of positive and negative links arise in a variety of online settings . . .

SciGen
(Cite) analyzed tweets as graph streams for predicting friendship relationships, although they focused on friendship relationships, not triads.

Score
Plausible → Correct; citation distinguishes that the source document deals with triads while the cited document does not

---

Source
Recognition Recent models of emotion recognition strongly rely on supervised deep learning solutions for the distinction of general emotion expressions. However, they are not reliable when recognizing online and personalized facial expressions, e.g., for person-specific affective understanding. In this paper, we present a neural model based on a conditional adversarial autoencoder . . .

Cited
The continuous dimensional emotion modelled by arousal and valence can depict complex changes of emotions. In this paper, we present our works on arousal and valence predictions for One-Minute-Gradual (OMG) Emotion Challenge. Multimodal representations are first extracted from videos using a variety of acoustic, video and textual models . . .

SciGen
This dataset contains 85,110 image-class videos and their respective emotion labels *(Cited)*.

Score
Plausible → Incorrect; cited paper seems to using the dataset from a challenge rather than introducing a dataset.

Source
The bag-of-words (BOW) model is the common approach for classifying documents, where words are used as feature for training a classifier, but their performance varies greatly depending on the model variant, features used and task/dataset. . . .

Cited
Variants of Naive Bayes (NB) and Support Vector Machines (SVM) are often used as baseline methods for text classification, where words are used as feature for training a classifier. This generally involves a huge number of features. This generally involves a huge number of features. Some techniques, such as Latent Semantic Analysis (LSA) . . .

SciGen
We use the subset of these datasets from *(Cited)*.

Score
Plausible → Incorrect; the cited paper seems to be an analysis paper and does not introduce any novel datasets on sentiment analysis.

Table 5: Examples of plausible SCIGEN outputs.

Source
Secure communication over a wiretap channel is considered in the disadvantaged wireless environment, where the eavesdropper channel is (possibly much) better than the main channel. ...

Cited
We consider the secure transmission of information over an ergodic fading channel in the presence of an eavesdropper. Our eavesdropper can be viewed as th e wireless counterpart of Wyner's wiretapper. ...

SciGen
In (*Cited*), an optimal SWIPT scheme was proposed with perfect CSIT.

IR
Consider the channel model shown in Figure 1 , which reflects the understanding that in an adversarial game in modern communication systems, it is the interference effects on wideband receiver front-ends rather than the baseband processing that is the significant detriment (*Cited*).

Citing sentence
However, public discussion schemes result in low secrecy rates in scenarios of interest (as discussed in detail in (*Cited*)), and the technique proposed here can be used in conjunction with public discussion approaches when two-way communication is possible.

Source
Indian epics have not been analyzed computationally to the extent that Greek epics have. In this paper, we show how interesting in sights can be derived from the ancient epic Mahabharata by applying a variety of analytical techniques based on a combination of natural language processing, sentiment/emotion analysis and social network analysis methods. ...

Cited
We present a method for extracting social networks from literature, namely, nineteenth-century British novels and serials. We derive the networks from dialogue interactions, and thus our method depends on the ability to determine when two characters are in conversation. ...

SciGen
The authors of (*Cited*) presented a method of characterizing the motivations for writing the essays by examining the topical influence of characters

IR
We present an approach to the extraction of family relations from literary narrative, which incorporates a technique for utterance attribution proposed recently by (*Cited*) .

Citing sentence
Robert (*Cited*) defined the eight basic emotion types.

Source
Automatic description generation from natural images is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities. In this survey, ...

Cited
Crowd-sourcing approaches such as Amazon's Mechanical Turk (MTurk) make it possible to annotate or collect large amounts of linguistic data at a relatively low cost and high speed. However, MTurk offers only limited control over who is allowed to particpate in a particular task. ...

SciGen
Evaluation was performed using the CIDEr metric (*Cited*).

IR
The last and the most challenging dataset, Pascal1k (*Cited*), is a collection of images with associated natural language sentences.

Citing sentence
The Pascal1K sentence dataset (*Cited*) is a dataset which is commonly used as a benchmark for evaluating the quality of description generation systems.

Table 6: Randomly selected examples of system inputs and outputs from validation set.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *NAACL-HLT*.

Lutz Bornmann and RÃijdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215âĂŞ2222.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL-HLT*.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *EMNLP*.

Arman Cohan and Nazli Goharian. 2017. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Dan Jurafsky. 2016. Citation classification for behavioral analysis of a scientific field. *ArXiv*, abs/1609.00435.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Joseph Lev et al. 1949. The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20(1):125–126.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2019. Gorc: A large contextual citation graph of academic papers.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165.

Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. 2002. On the recommending of citations for research papers. In *CSCW*.

Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, volume 4, pages 81–88.

Ramesh Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *KDD*.

Chris D. Paice. 1980. The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases. In *SIGIR*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Marco Valenzuela, Vu A. Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Richard Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *AAAI*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *ArXiv*, abs/1905.12616.