Contextualizing Generated Citation Texts

Biswadip Mandal, Xiangci Li, Jessica Ouyang
The University of Texas at Dallas
Richardson, TX, USA
{biswadip.mandal, xiangci.li, jessica.ouyang}@utdallas.edu

Abstract

Abstractive citation text generation is usually framed as an infilling task, where a sequence-to-sequence model is trained to generate a citation given a reference paper and the context window around the target; the generated citation should be a brief discussion of the reference paper as it relates to the citing context. However, examining a recent LED-based citation generation system, we find that many of the generated citations are generic summaries of the reference paper's main contribution, ignoring the citation context's focus on a different topic. To address this problem, we propose a simple modification to the citation text generation task: the generation target is not only the citation itself, but the entire context window, including the target citation. This approach can be easily applied to any abstractive citation generation system, and our experimental results show that training in this way is preferred by human readers and allows the generation model to make use of contextual clues about what topic to discuss and what stance to take.

Keywords: scientific summarization, conditional generation, citation text generation

1. Introduction

Citation text generation is the task of generating a short summary of a reference paper, focused on how it relates to a citing paper. Modern approaches (Xing et al., 2020; Ge et al., 2021; Luu et al., 2021; Chen et al., 2021; Li et al., 2022) frame this problem as an infilling task: a context window of up to three sentences before and after the target citation is extracted, and the target is masked; given this context window and the reference paper abstract, a sequence-to-sequence (seq2seq) model is trained to generate the masked target citation.

If we examine the context window of a target citation, we find citations of similar related works, and the author may criticize them in order to draw attention to a gap that the citing paper intends to fill; to a human reader, the context window provides clues indicating the topic the citation will discuss and the stance the author will take towards the reference paper. Thus, a citation should not be a generic summary of the reference paper, but rather a query-focused summary, based on the context. After all, different papers can cite the same reference paper in completely different ways, depending on whether they share a task, an approach, or a dataset; even within the same citing paper, a reference paper can be cited differently in the introduction versus in the methodology or comparison of results.

It has been observed that neural language models often generate text that is generic and vague (Holtzman et al., 2019). Examining the recent LED-based citation generation system of Li et al. (2022), we see a similar problem. Figure 1 shows an example of their LED model's output. Although the generated citation is a valid summary of the reference paper, it is too generic, mentioning only "syntactic information." The context sentence containing the masked target already describes a constituent-structure approach, making it clear to a human reader that the target should discuss a corresponding dependency-structure based approach. However, the LED model produces a generic summary of the reference paper, which while factually correct, does not fit the context well and is actually redundant with information given earlier in the context.

Context: Broadly speaking, prior work on SRL makes use of syntactic information in two different ways. Carreras and Marquez (2005); Pradhan et al. (2013) incorporate constituent-structure span-based information, while **[MASK]**...

Ground truth: Haji et al. (2009) incorporate dependency-structure information.

LED-baseline: Haji et al. (2009) integrate syntactic information into a neural SRL system.

Contextualized: Haji et al. (2009) use dependency-level information.

Figure 1: An infilling-style baseline approach produces a generic summary of the reference paper that does not fit well in the citation context.

To address this issue, we propose *contextualized citation generation*, a context-focused modification to the citation text generation task. In this version of the task, the model must generate not only the target citation, but the surrounding context as well. In this way, the context window serves as a sort of prompt for the decoder. Figure 1 shows how

LREC-COLING 2024, pages 3849–3854
20-25 May, 2024.

the output of our contextualized approach correctly identifies dependency parsing as the topic of the citation, and our experimental results show that human readers prefer citations generated using our approach over the baseline infilling approach. Our proposed training method can be easily applied to any seq2seq citation generation model to produce coherent citations that fit their contexts.

2. Related Work

The citation text generation task was proposed by Hoang and Kan (2010). They and other extractive approaches (Hu and Wan, 2014; Chen and Zhuge, 2019) selected the most salient sentences from the reference papers to serve as their citations.

More recently, neural abstractive approaches (Xing et al., 2020; Ge et al., 2021; Luu et al., 2021; Chen et al., 2021; Li et al., 2022) have trained seq2seq models for the citation infilling task, where the input is the citation context (with target citation masked) and the reference paper abstract. Our work aims to condition the generated citation even more strongly on the context by training the model to generate the entire context window, not just the target citation. While the above works differ in features and datasets used, as well as the size and shape of the context window, it is important to note that our approach changes only the generation target, and so is compatible with all of them. In our experiments, we use the most recent approach, Li et al. (2022), as the baseline model.

3. Methodology

We propose a simple modification of the citation text generation task in order to produce citations that are more coherent with their surrounding contexts. The task is to generate the entire context window, with the masked target citation filled in. For ease of evaluation, we add a meta-token `[SEP]` to distinguish the context from the target citation. Figure 2 illustrates the difference between our approach and the standard infilling approach.

3.1. Experimental Setup

Data. We use the NLP-domain CORWA citation text generation dataset (Li et al., 2022). Table 1 shows the size of the dataset and its partitions. We merge the human-annotated *train* and automatically-labeled *distant* sets to create a single large, unified training set.

Partition	Papers	Citations
train	565	2,243
distant	11,564	32,512
test	362	1,322

Table 1: CORWA dataset partitions and statistics.

Model. We use the Longformer Encoder-Decoder (LED; Beltagy et al., 2020) citation generation model of Li et al. The input to the model is the concatenation of the citing paper's introduction section; the paragraph containing the (masked) target citation; and the citation mark (eg. "Smith et al. (2023)"), title, and abstract of the reference paper[1].

4. Results and Analysis

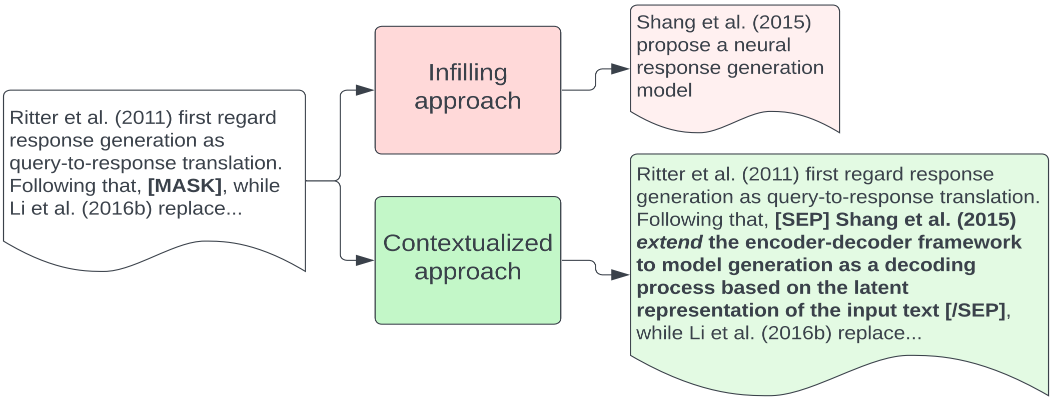
4.1. Human Evaluation

We recruit six graduate students from our university's Computer Science Department to serve as judges. We split the them into two groups of three, assigning each group 30 samples. The judges are shown the input to the generation model and the ground truth, baseline, and contextualized citations; the order of citations is randomly shuffled to anonymize the models as shown in Table 3. The judges are asked to indicate which citation they prefer based on *Fluency*, *Relevance* to the reference paper, *Coherence* in the citation context, and *Overall* quality.

Table 2 shows our human evaluation results. We see that the proposed contextualized generation model is slightly preferred with respect to Relevance, Coherence, and Overall, while the baseline model is more fluent. Comparing the contextualized model with the ground truth, we find that many samples have indistinguishable (ie. comparable) performance. Interestingly, we observe that the Relevance of ground truth is judged lower than that of the contextualized model. This may be because the ground truth citations are written by human authors who have access to the entire reference paper, while the generation models and human judges only see the reference paper abstracts.

We achieve moderate inter-annotator agreement for the first group of three judges, with pairwise Kendall's τ of 0.31, 0.17, and 0.35. We had lower agreement for the second group, with Kendall's τ of 0.15, 0.08, and 0.04. Examining the judges' scores, we find that one judge from the second group consistently disagreed with the other two; we also tallied the results excluding this third judge, but we did not find any significant difference from the results shown in Table 2.

[1] We also experimented with prompting GPT-3.5 Turbo model with the same input, but the performance was much lower than our baseline Li et al. model



Ritter et al. (2011) first regard response generation as query-to-response translation. Following that, **[MASK]**, while Li et al. (2016b) replace...

Shang et al. (2015) propose a neural response generation model

Ritter et al. (2011) first regard response generation as query-to-response translation. Following that, **[SEP] Shang et al. (2015) *extend* the encoder-decoder framework to model generation as a decoding process based on the latent representation of the input text [/SEP]**, while Li et al. (2016b) replace...

Figure 2: Comparison of the standard infilling-based citation text generation approach, where the generation target is just the target citation itself, and our proposed contextualized approach, where the generation target is the entire context window with the target citation filled in.

Cited Paper Title: Latent Predictor Networks for Code Generation
Citation Mark: Ling et al. (2016)
Abstract: Many language generation tasks require the production of text conditioned on both structured and unstructured inputs. We present a novel neural network architecture which generates an output sequence conditioned on an arbitrary number of input functions . . .

System 0: Neural encoder-decoder models have proved effective in mapping NL to logical forms (Dong and Lapata, 2016) and also for directly producing general purpose programs (Iyer et al., 2017; 2018). **Ling et al. (2016) use a sequence-to-sequence model with attention and a copy mechanism to generate source code.** Instead of directly . . .

System 1: Neural encoder-decoder models have proved effective in mapping NL to logical forms (Dong and Lapata, 2016) and also for directly producing general purpose programs (Iyer et al., 2017; 2018). **Ling et al. (2016) show that a neural encoder-decoder model can outperform strong baselines at the task of generating Java/Python code from card games.** Instead of directly . . .

Figure 3: Evaluation setup for our human judges, who are asked if they prefer System 0 or System 1 in terms of Fluency, Relevance, Coherence, and Overall. We randomly assign System 0 and System 1 as the baseline, ground truth, or our contextualized approach.

4.2. Analyzing Contextualized Citations

Examining the generated citations, we find several common scenarios where the baseline model ignores clues in the context window that the contextualized model is able use.

Model	Flu.	Rel.	Coh.	Overall
Contextualized vs Baseline				
♡Contextualized	12	**49**	**41**	**53**
♡Baseline	**33**	44	36	50
Indistinguishable	135	87	103	77
Contexualized vs Ground Truth				
♡Contextualized	**19**	**38**	**28**	**27**
♡Ground Truth	18	13	21	18
Indistinguishable	143	129	131	135

Table 2: Human evaluation comparing contextualized and infilling-based citation generation.

Context: . . . To the best of our knowledge, **[MASK]**. Most similar to our work in spirit, Ding et al. (1997) used Layer-wise Relevance Propagation (LRP; Bach et al al. 2015), *an interpretation method resembling saliency*, to interpret . . .

Ground truth: Li et al. (2016) presented *the only work that directly employs saliency methods* to interpret NLP models.

Baseline: Li et al. (2016) proposed three strategies for visualizing compositionality in NMT.

Contextualized: Li et al (2016) is *the only work that applied saliency* to NMT. They used a simple recurrent neural network (RNN) to learn a saliency function for each word in the sentence.

Figure 4: Example where the context hints the citation should discuss saliency (in *italics*).

The context discusses the historical development of a particular topic. A coherent citation should focus on the part of the reference paper that is related to that topic; even if that topic is not

Context: . . . Previous work on **[MASK]**. *In contrast with their work,* we train models to . . .

Ground truth: curriculum learning for MT (Kocmi and Bojar, 2017; Wang et al., 2018) proposed methods which feed easier samples to the model first and later show more complex sentences. *However,* their focus is on improving convergence time while providing limited success on improving translation quality.

Baseline: dynamic sentence sampling (Kocmi and Bojar, 2017; Wang et al., 2018) focuses on how to dynamically sample sentences in order to accelerate the training of NMT.

Contextualized: fine-tuning neural machine translation models (Kocmi and Bojar, 2017; Wang et al., 2018) has tried to dynamically sample sentences in order to accelerate training. *However,* these methods require training models to learn a prior distribution of similar sentences.

Figure 5: Example where the context hints the citation should criticize the reference paper (in *italics*).

especially important in the reference paper, a good citation generation model should still be able to identify it as the topic of focus. Figure 4 shows an example where the context paragraph contains two clues. First, the introductory phrase "to the best of our knowledge" hints that the citation should claim that few, if any prior works have addressed the topic, identifying a gap that the citing paper intends to fill. Second, the following sentences states that the most similar work uses a "method resembling saliency." The contextualized model is able to put these clues together, highlighting how "Li et al (2016) is the only work that applied saliency."

The context implies that the reference paper should be compared or contrasted. Teufel et al. (2006) identifies several author intents for citing a given paper, including contrasting methods or results, claiming the cited work is worse than or comparable to the current work, or comparing or contrasting with other cited works. A common pattern is that the context sentence following the masked target discusses how the current work addresses limitations in the prior work, making it evident that the target citation criticizes the reference paper. Figure 5 shows an example where the context sentence immediately following the target citation begins with "in contrast with their work."

The context implies that the reference paper should be discussed in a similar way as another work. Figure 6 shows an example where

Context: . . . Rubinstein et al. (2015) demonstrated that *state-of-the-art distributional semantic models fail to predict attributive properties of concept words* (e.g. the properties is-red and is-round for the word apple) as accurately as taxonomic properties (e.g. is-a-fruit). *Similarly*, **[MASK]** . . .

Ground truth: Sommerauer and Fokkens (2018) investigated the *types of semantic knowledge encoded within pretrained word embeddings, concluding that some properties cannot be learned* by supervised classifiers.

Baseline: Somasomauer and Fokkens (2018) compared word embeddings to a feature-identification method based on cosine similarity.

Contextualized: Sommerauer and Fokkens (2018) showed that the *properties of concepts are not captured by word embeddings*.

Figure 6: Example where the context hints the citation should discuss how the reference paper is similar to a previous citation (in *italics*).

the sentence containing the target citation begins with "similarly," hinting that the citation should discuss how the reference paper is similar to the citation in the previous sentence. The baseline model again produces a generic summary, while the contextualized model is able to identify the topic of focus as the failure of word embeddings to capture information about concept words.

The target citation should not be redundant with the context. We find several instances where the baseline model's generated citation repeats information already present in the context. While the citation is still a valid summary of the reference paper, the baseline model fails to recognize that the information is redundant and it should focus on something else, such as a different method or concept. The example in Figure 1 shows a generic, redundant citation generated by the baseline model that should have been more specific.

The reference paper is an extension of another work. When describing the historical development of a task or topic, it is common to present works in chronological order, where each work builds on top of the previous ones. A good citation generation model should recognize when the reference paper is an extension of another paper cited earlier in the context. Figure 7 shows an example where the context discusses a paper, and the model needs to generate a citation for a follow-up work by the same first author. The contextualized model generates the phrase "further extended" (referring

Context: . . . Rei et al. (2016) extended this model to include character embeddings in order to capture morphological similarities such as word endings. **[MASK]** . . .

Ground truth: Rei (2017) *subsequently* added a secondary LM objective to the neural sequence labeling architecture, operating on both word and character-level embeddings. This was found to be particularly useful for GED -introducing an LM objective allows the network to learn more generic features about language and composition.

Baseline: Rei (2017) proposed a semi-supervised approach to the task of GED, using a language modeling objective to predict surrounding words for every word in the sentence.

Contexualized: Rei (2017) *further extended* this model with a secondary training objective, learning to predict the surrounding words in the context.

Figure 7: Example where the contextualized model correctly describes one reference paper as an extension of another (in *italics*).

to the earlier work, which already “extended"). We observe that citations often contain an action verbs, such as *extend*, and *improve*, which are used when the reference paper is dependent on an earlier work. Compared with the baseline model, our contextualized model more frequently such verbs.

5. Conclusion

We present a simple reframing of the citation text generation task to make better use of citing context information. Our approach of generating the entire citing context window, with the target citation filled in, produces output that is more appropriate to its context than the existing method of generating the target citation alone. Our proposed approach changes only the generation target and is agnostic to any special features or input representations, so it is straightforward to apply to any existing citation generation models via retraining. We have made our code available at `https://github.com/mandalbiswadip/ContextualGeneration`. Our human evaluation reveals that readers are able to judge when a citation better fits its context, and we present a qualitative analysis of some common shortcomings of citations generated by the baseline approach that are addressed by our contextualized model.

Limitations

The standard automatic evaluation metric for text generation, ROUGE, does not capture coherence well. This is a known problem: small differences in ROUGE do not correspond to noticeable differences in generation quality (Deutsch et al., 2022), and ROUGE does not capture other aspects of the summary, such as informativeness (Schluter, 2017) and factuality (Wallace et al., 2021). We instead present a human evaluation and qualitative analysis to compare the coherence citations generated using our proposed approach with those of the baseline model. However, such evaluations are be expensive and time-consuming to conduct, requiring significant reading and cognitive effort from domain-expert judges; as a result, we are not able to evaluate a very large number of samples.

In addition, our work is focused on natural language processing papers published in the English language. Consequently, the scope of our model may not encompass the full range of diversity exhibited by papers from other fields (e.g., biology) or different sub-fields within computer science.

Ethics Statement

Our work targets the task of citation generation, where plagiarism of the cited paper is a significant concern. We not attempt to control the level of extractiveness of our generated citations; as a result, they may copy extensively from the cited paper abstract, which given as input.

Further, our approach generates citations that criticize the reference paper when it seems appropriate to do so based on the citation context. Errors or hallucinations by our model could result in false or unfair criticisms, which can mislead the reader and negatively impact their perception of the reference paper.

6. Bibliographical References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, 31(3):e4261.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th An-*

nual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6068–6077, Online. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. *arXiv preprint arXiv:2204.10216*.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A background knowledge- and content-based framework for citing sentence generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1466–1478, Online. Association for Computational Linguistics.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Coling 2010: Posters*, pages 427–435.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633, Doha, Qatar. Association for Computational Linguistics.

Xiangci Li, Biswadip Mandal, and Jessica Ouyang. 2022. CORWA: A citation-oriented related work annotation dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5426–5440, Seattle, United States. Association for Computational Linguistics.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2130–2144, Online. Association for Computational Linguistics.

Natalie Schluter. 2017. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

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.