BACO: A Background Knowledge- and Content-Based Framework for Citing Sentence Generation

Yubin Ge[1], Ly Dinh[1], Xiaofeng Liu[2], Jinsong Su[3], Ziyao Lu[3], Ante Wang[3], Jana Diesner[1]
[1]University of Illinois at Urbana-Champaign, USA
[2]Harvard University, USA
[3]Xiamen University, China
{yubinge2, dinh4, jdiesner}@illinois.edu
jssu@xmu.edu.cn

Abstract

In this paper, we focus on the problem of citing sentence generation, which entails generating a short text to capture the salient information in a cited paper and the connection between the citing and cited paper. We present **BACO**, a **BA**ckground knowledge- and **CO**ntent-based framework for citing sentence generation, which considers two types of information: (1) *background knowledge* by leveraging structural information from a citation network; and (2) *content*, which represents in-depth information about *what to cite* and *why to cite*. First, a citation network is encoded to provide *background knowledge*. Second, we apply salience estimation to identify *what to cite* by estimating the importance of sentences in the cited paper. During the decoding stage, both types of information are combined to facilitate the text generation. We then conduct joint training of the generator and citation function classification to make the model aware of *why to cite*. Our experimental results show that our framework outperforms comparative baselines.

1 Introduction

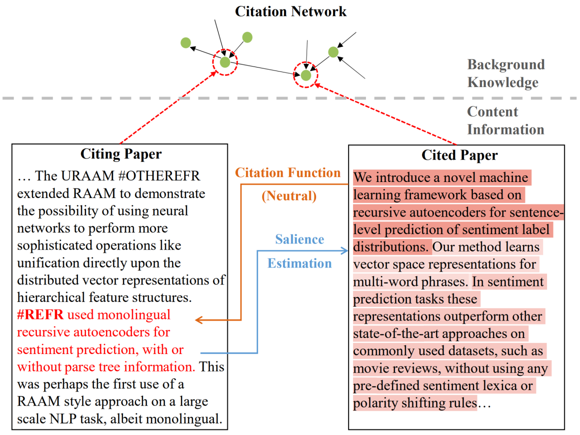
A citation systematically, strategically, and critically synthesizes content from a cited paper in the context of a citing paper (Smith, 1981). A paper's text that refers to prior work, which we herein refer to as citing sentences, forms the conceptual basis for a research question or problem; identifies issues, contradictions, or gaps with state of the art solutions; and prepares readers to understand the contributions of a citing paper, e.g., in terms of theory, methods, or findings (Elkiss et al., 2008). Writing meaningful and concise citing sentences that capture the gist of cited papers and identify connections between citing and cited papers is not trivial (White, 2004). Learning how to write up information about related work with appropriate and meaningful citations is particularly challenging for new scholars (Mansourizadeh and Ahmad, 2011).

To assist scholars with note taking on prior work when working on a new research problem, this paper focuses on the task of citing sentence generation, which entails identifying salient information from cited papers and capturing connections between cited and citing papers. With this work, we hope to reduce scientific information overload for researchers by providing examples of concise citing sentences that address information from cited papers in the context of a new research problem and related write up. While this task cannot and is not meant to replace the scholarly tasks of finding, reading, and synthesizing prior work, the proposed computational solution is intended to support especially new researchers in practicing the process of writing effective and focused reflections on prior work given a new context or problem.

A number of recent papers have focused on the task of citing sentence generation (Hu and Wan, 2014; Saggion et al., 2020; Xing et al., 2020), which is defined as generating a short text that describes a cited paper B in the context of a citing paper A, and the sentences before and after the citing sentences in paper A are considered as context. However, previous work has mainly utilized limited information from citing and cited papers to solve this task. We acknowledge that any such solution, including ours, is a simplification of the intricate process of how scholars write citing sentences.

Given this motivation, we explore two sets of information to generate citing sentences, namely *background knowledge* in the form of citation networks, and *content* from both citing and cited papers, as shown in Figure 1. Using citation networks was inspired by the fact that scholars have analyzed such networks to identify the main themes and research developments in domain areas such as information sciences (Hou et al., 2018), busi-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1466–1478
August 1–6, 2021. ©2021 Association for Computational Linguistics



… The URAAM #OTHEREFR extended RAAM to demonstrate the possibility of using neural networks to perform more sophisticated operations like unification directly upon the distributed vector representations of hierarchical feature structures. **#REFR** used monolingual recursive autoencoders for sentiment prediction, with or without parse tree information. This was perhaps the first use of a RAAM style approach on a large scale NLP task, albeit monolingual.

We introduce a novel machine learning framework based on recursive autoencoders for sentence-level prediction of sentiment label distributions. Our method learns vector space representations for multi-word phrases. In sentiment prediction tasks these representations outperform other state-of-the-art approaches on commonly used datasets, such as movie reviews, without using any pre-defined sentiment lexica or polarity shifting rules…

Figure 1: An example from our dataset (source: ACL Anthology Network corpus (Radev et al., 2013). The red text in the citing paper is the citing sentence, and the special token #REFR indicates the citation of the cited paper. Our framework aims at capturing information from two perspectives: *background knowledge* and *content*. The *background knowledge* is learned by obtaining structural features of the citation network. The *content* information entails estimated sentence salience (higher salience is highlighted by darker color) in the cited paper and the corresponding citation function of the cited paper to the citing paper.

ness modeling (Li et al., 2017), and pharmaceutical research (Chen and Guan, 2011).

We use the *content* of citing and cited papers as a second set of features to capture two more in-depth *content* features: (1) *What to cite* - while the overall content of a cited paper needs to be understood by the authors of the citing paper, not all content is relevant for writing citing sentences. Therefore, we follow the example of estimating salient sentences (Yasunaga et al., 2019) and use the predicted salience to filter crucial information that should be integrated into the resulting citing sentence; (2) *Why to cite* - we define "citation function" as an approximation of an author's reason for citing a paper (Teufel et al., 2006). A number of previous research on citation functions has used citing sentences and their context for classification (Zhao et al., 2019; Cohan et al., 2019). Our paper involves citation functions into citing sentence generation so that the generated citing sentences can be coherent given their context, and can still contain the motivation for a specific citation.

In this paper, we propose a **BA**ckground knowledge- and **CO**ntent-based framework, named **BACO**. Specifically, we encode a citation network based on citation relations among papers to obtain *background knowledge*, and the given citing and cited papers to provide *content* information. We extend a standard pointer-generator (See et al., 2017) to copy words from cited and citing papers, and determine *what to cite* by estimating sentence salience in the cited paper. The various pieces of captured information are then combined as the context for the decoder. Furthermore, we extend our framework to include *why to cite* by jointly training the generation with citation function classification and facilitate the acquisition of the *content* information.

As for the dataset, we extended the ACL Anthology Network corpus (AAN) (Radev et al., 2013) with extracted citing sentences by using RegEx. We then hand-annotated the citation functions on a subset of the dataset, and trained a citation function labeling model based on SciBERT (Beltagy et al., 2019). The resulting labeling model was then used to automatically label the rest data to build a large-scale dataset.

We summarize our contributions as follows:

- We propose a **BA**ckground knowledge- and **CO**ntent-based framework, named **BACO**, for citing sentence generation.
- We manually annotated a subset of citing sentences with citation functions to train a SciBERT-based model to automatically label the rest data for citing sentence generation.
- Based on the results from experiments, we show that BACO outperforms comparative baselines by at least 2.57 points on ROUGE-2.

2 Related Work

Several studies on citing sentence generation have used keyword-based summarization methods (Hoang and Kan, 2010; Chen and Zhuge, 2016, 2019). To that end, they built keyword-based trees to extract sentences from cited papers as related work write-ups. These studies have two limitations: First, since related work sections are not simply (chronological) summaries of cited papers, synthesizing prior work in this manner is insufficient. Second, extractive summarization uses verbatim content from cited papers, which implies intellectual property issues (e.g., copyright violations) as well as ethical problems, such as a lack of intellectual engagement with prior work. Alternatively, abstractive summarization approaches, such as methods based on linear programming (Hu and Wan, 2014) and neural seq2seq methods (Wang et al., 2018), have also been explored. These approaches mainly focus on utilizing papers' content information, specifically on the text of cited papers directly. A recent paper that went beyond summarizing the content of cited papers (Xing et al., 2020) used a multi-source, pointer-generator network with a cross attention mechanism to calculate the attention distribution between the citing sentences' context and the cited paper's abstract.

Our paper is based on the premise that citation network analysis can provide background knowledge that facilitates the understanding of papers in a field. Prior analyses of citation networks have been used to reveal the cognitive structure and interconnectedness of scientific (sub-)fields (Moore et al., 2005; Bruner et al., 2010), and to understand and detect trends in academic fields (You et al., 2017; Asatani et al., 2018). Network analysis has also been applied to citation networks to identify influential papers and key concepts (Huang et al., 2018), and to scope out research areas.

While previous studies have shown that using text from citing papers is useful to generate citing sentences, the benefit of other content-based features of a citation (e.g., reasons for citing) is insufficiently understood (Xing et al., 2020). Extant literature on citation context analysis (Moravcsik and Murugesan, 1975; Lipetz, 1965), which focused on the connections between the citing and cited papers with respect to purposes and reasons for citations, has found that citation function (Ding et al., 2014; White, 2004) is an important indicator of why a paper chose to cite specific paper(s). Based on a content analysis of 750 citing sentences from 60 papers published in two prominent physics journals, Lipetz (1965) identified 11 citation functions, such as *questioned*, *affirmed*, or *refuted* cited paper's premises. Similarly, Moravcsik and Murugesan (1975) qualitatively coded the citation context of 30 articles on high energy physics, finding 10 citation functions grouped into 5 pairs: conceptual-operational, organic-perfunctory, evolutionary-juxtapositional, confirmative-negational, valuable-redundant.

Citation context analysis has also been used to study the valence of citing papers towards cited papers (Athar, 2011; Abu-Jbara et al., 2013) by classifying citation context as positive, negative, or neutral. In this paper, we adopt Abu-Jbara et al. (2013)'s definition of a positive citation as a citation that explicitly states the strength(s) of a cited paper, or a situation where the citing paper's work is guided by the cited paper. In contrast to that, a negative citation is one that explicitly states the weakness(es) of a cited paper. A neutral citation is one that objectively summarizes the cited paper without an additional evaluation. In addition to these three categories, we also consider mixed citation contexts (Cullars, 1990), which are citations that contain both positive and negative evaluations of a cited papers, or where the evaluation is unclear. Given that our paper is a first attempt to integrate citation functions into citing sentence generation, we opted to start with a straightforward valence category schema before exploring more complex schemas in future work.

3 Dataset and Annotation

We first extended the AAN[1] (Radev et al., 2013) with the extracted citing sentences using RegEx. We followed the process in (Xing et al., 2020) to label 1,200 randomly sampled citing sentences with their citation functions. The mark-up was done by 6 coders who were provided with definitions of positive, negative, neutral, and mixed citation functions, and ample examples for each valence category. Our codebook including definitions and examples of citation functions is shown in Table 1. After the annotation, we randomly split the dataset into 800 instances for training and the remaining 400 for testing. We then used the 800 human-annotated instances to train a citation function labeling model with 10-fold cross validation. The labeling task was treated as a multi-class classification problem.

Our labeling model was built upon SciBERT (Beltagy et al., 2019), a pre-trained language model based on BERT (Devlin et al., 2019) but trained on a large corpus of scientific text. We added a multilayer perceptron (MLP) to SciBERT, and fine-tuned the whole model on our dataset. As for the input, we concatenated each citing sentence with its context in the citing paper, and inserted a special tag `[CLS]` at the beginning and another special tag `[SEP]` to separate them. The final hidden state that corresponded to `[CLS]` was used as the aggregate sequence representation. This state was fed into the MLP, followed by the softmax function for predicting the citation function of the citing sentence. We report details of test results and dataset statistics in the Appendix, Section A.1.

[1] http://aan.how/download/

Citation function	Definition	Example
Positive	When a citing paper explicitly supports a cited paper's premises and findings, and/or that the cited paper's finding(s) is used to corroborate with the citing paper's study.	”Our architecture is inspired by state-of-the-art model #REFR”
Negative	When a citing paper points out weaknesses in the cited paper's premises and findings, as well as explicitly rejecting the finding(s) from the cited paper.	”Unbounded Content Early versions of the customization system #REFR only allowed a small number of entries for things like lexical types.”
Neutral	When a citing paper objectively summarizes the cited paper's premises and findings, without explicitly offering any evaluations of the cited paper's finding(s).	”#REFR translates the extracted Vietnamese phrases into corresponding English ones”
Mixed	When a citing paper does not clearly express their evaluations towards the cited paper, or that the citing paper contains multiple citation functions in one sentence.	”In previous work, this has been done successfully for question answering tasks #REFR, but not for web search in general.”

Table 1: Definitions and examples of our proposed citation functions. The special token #REFR indicates the citation of a paper.

4 Methodology

Our proposed framework includes an encoder and a generator, as shown in Figure 2. The encoder takes the citation network and the citing and cited papers as input, and encodes them to provide *background knowledge* and *content* information, respectively. The generator contains a decoder that can copy words from citing and cited paper while retaining the ability to produce novel words, and a salience estimator that identifies key information from the cited paper. We then trained the framework with citation function classification to enable the recognition of why a paper was cited.

4.1 Encoder

Our encoder (the yellow shaded area in Figure 2) consists of two parts, a graph encoder that was trained to provide *background knowledge* based on the citation network, and a hierarchical RNN-based encoder that encodes the *content* information of the citing and cited papers.

4.1.1 Graph Encoder

We designed a citation network pre-training method for providing the *background knowledge*. In detail, we first constructed a citation network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is a set of nodes/papers[2] and $\mathcal{E}$ is a set of directed edges. Each edge links a citing paper (source) to a cited paper (target). To utilize $\mathcal{G}$ in our task, we employed a graph attention network (GAT) (Veličković et al., 2018) as our graph encoder, which leverages masked self-attentional layers to compute the hidden representation of each node. This GAT has been shown to be effective on multiple citation network benchmarks. We input a set of node pairs $\{(v_p, v_q)\}$ into it for training of the link prediction task. We pre-trained our graph encoder network using negative sampling to learn the node representations h_p^n for each paper p, which contains structural information of the citation network and can provide *background knowledge* for the downstream task.

4.1.2 Hierarchical RNN-based Encoder

Given the word sequence $\{cw_i\}$ of the citing sentence's context and the word sequence $\{aw_j\}$ of the cited paper's abstract, we input the embedding of word tokens (e.g., $e(w_t)$) into a hierarchical RNN-based encoder that includes a word-level Bi-LSTM and a sentence-level Bi-LSTM. The output word-level representation of the citing sentence's context is denoted as $\{\mathbf{h}_i^{cw}\}$, and the cited paper's abstract is encoded similarly as its word-level representation $\{\mathbf{h}_j^{aw}\}$. Meanwhile, their sentence-level representations are represented as $\{\mathbf{h}_m^{cs}\}$ and $\{\mathbf{h}_n^{as}\}$.

4.2 Generator

Our generator (the green shaded area in Figure 2) is an extension of the standard pointer generator (See et al., 2017). It integrates both *background knowledge* and *content* information as context for text generation. The generator contains a decoder and an additional salience estimator that predicts the salience of sentences in the cited paper's abstract for refining the corresponding attention.

[2] We use *node* and *paper* interchangeably

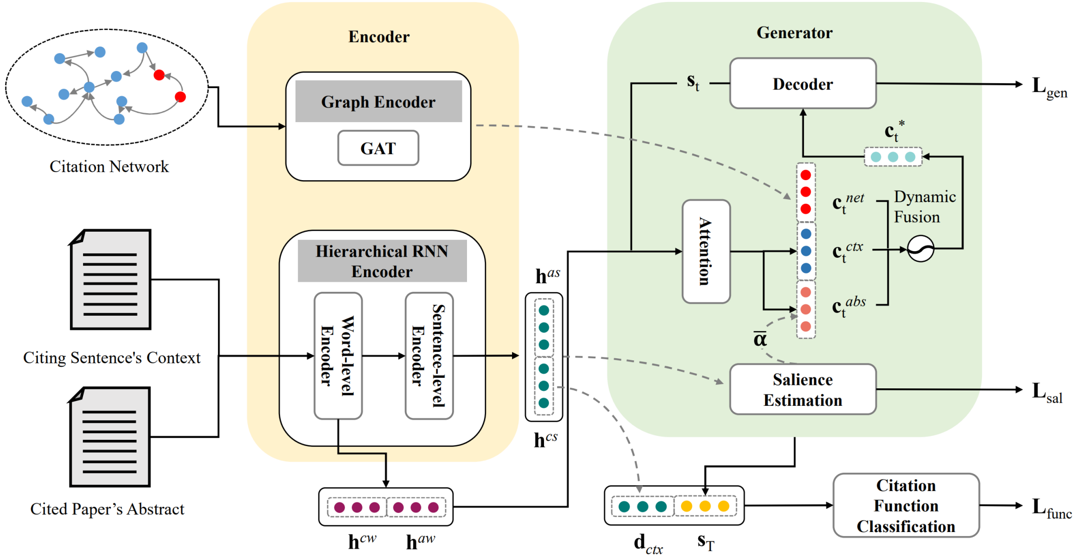


Figure 2: The overall architecture of the proposed framework. The encoder is used to encode the citation network and papers' (both citing and cited) text. The generator estimates salience of sentence in the cited paper's abstract, and utilizes this information for text generation. The framework is additionally trained with citation functions.

4.2.1 Decoder

The decoder is a unidirectional LSTM conditioned on all encoded hidden states. The attention distribution is calculated as in (Bahdanau et al., 2015).

Since we considered both the citing sentence's context and the cited paper's abstract on the source side, we applied the attention mechanism to $\{\mathbf{h}_i^{cw}\}$ and $\{\mathbf{h}_j^{aw}\}$ separately to obtain two attention vectors $\mathbf{a}_t^{\text{ctx}}$, $\mathbf{a}_t^{\text{abs}}$, and their corresponding context vectors $\mathbf{c}_t^{\text{ctx}}$, $\mathbf{c}_t^{\text{abs}}$ at the step t. We then aggregated input context $\mathbf{c}_t^*$ from the citing sentence's context, the cited paper's abstract, and background knowledge by applying a dynamic fusion operation based on modality attention as described in (Moon et al., 2018b,a), which selectively attenuated or amplified each modality based on their importance to the task:

$$[\text{att}_{\text{ctx}}; \text{att}_{\text{abs}}; \text{att}_{\text{net}}] = \sigma(\mathbf{W}_m[\mathbf{c}_t^{\text{ctx}}; \mathbf{c}_t^{\text{abs}}; \mathbf{c}_t^{\text{net}}] + \mathbf{b}_m), \tag{1}$$

$$\tilde{\text{att}}_m = \frac{\exp(\text{att}_m)}{\sum_{m' \in \{\text{abs,ctx,net}\}} \exp(\text{att}_{m'})}, \tag{2}$$

$$\mathbf{c}_t^* = \sum_{m \in \{\text{abs,ctx,net}\}} \tilde{\text{att}}_m \mathbf{c}_t^m, \tag{3}$$

where $\mathbf{c}_t^{\text{net}} = [\mathbf{h}_p^n; \mathbf{h}_q^n]$ represents the learned background knowledge for papers p and q, and is kept constant during all decoding steps t, and $[\text{att}_{\text{ctx}}; \text{att}_{\text{abs}}; \text{att}_{\text{net}}]$ is the attention vector.

To enable our model to copy words from both the citing sentence's context and the cited paper's abstract, we calculated the generation probability and copy probabilities as follows:

$$\begin{aligned}[p_{\text{gen}}, p_{\text{copy1}}, p_{\text{copy2}}] &= \text{softmax}(\mathbf{W}_{\text{ctx}}\mathbf{c}_t^{\text{ctx}} \\ &+ \mathbf{W}_{\text{abs}}\mathbf{c}_t^{\text{abs}} + \mathbf{W}_{\text{net}}\mathbf{c}_t^{\text{net}} + \mathbf{W}_{\text{dec}}\mathbf{s}_t \\ &+ \mathbf{W}_{\text{emb}}e(w_{t-1}) + \mathbf{b}_{ptr}),\end{aligned} \tag{4}$$

where p_{gen} is the probability of generating words, p_{copy1} is the probability of copying words from the citing sentence's context, p_{copy2} is the probability of copying words from the cited paper's abstract, $\mathbf{s}_t$ represents the hidden state of the decoder at step t, and $e(w_{t-1})$ indicates the input word embedding. Meanwhile, the context vector c_t^*, which can be seen as an enhanced representation of source-side information, was concatenated with the decoder state s_t to produce the vocabulary distribution P_{vocab}:

$$P_{\text{vocab}} = \text{softmax}(\mathbf{V}'(\mathbf{V}[\mathbf{s}_t; \mathbf{c}_t^*] + \mathbf{b}) + \mathbf{b}'). \tag{5}$$

Finally, for each text, we defined an extended vocabulary as the union of the vocabulary and all words appearing in the source text, and calculated the probability distribution over the extended vocabulary to predict words w:

$$\begin{aligned}P(w) =& p_{\text{gen}}P_{\text{vocab}}(w) + p_{\text{copy1}} \sum_{i:cw_i=w} a_{t,i}^{\text{ctx}} \\ &+ p_{\text{copy2}} \sum_{i:aw_i=w} a_{t,i}^{\text{abs}}.\end{aligned} \tag{6}$$

4.2.2 Salience Estimation

The estimation of the salience of each sentence that occurs in a cited paper's abstract was used to identify what information needed to be concentrated for the generation. We assumed a sentence's salience to depend on the citing paper such that the same sentences from one cited paper can have different salience in the context of different citing papers. Hence, we represented this salience as a conditional probability $P(s_i|D_{\text{src}})$, which can be interpreted as the probability of picking sentence s_i from a cited paper's abstract given the citing paper D_{src}.

We first obtained the document representation $\mathbf{d}_{\text{src}}$ of a citing paper as the average of all its abstract's sentence representations. Then, for calculating salience, which is defined as $P(s_i|D_{\text{src}})$, we designed an attention mechanism that assigns a weight α_i to each sentence s_i in a cited paper's abstract D_{tgt}. This weight is expected to be large if the semantics of s_i are similar to $\mathbf{d}_{\text{src}}$. Formally, we have:

$$\alpha_i = \mathbf{v}^T \tanh(\mathbf{W}_{\text{doc}}\mathbf{d}_{\text{src}} + \mathbf{W}_{\text{sent}}\mathbf{h}_i^{as} + \mathbf{b}_{\text{sal}}), \tag{7}$$

$$\tilde{\alpha}_i = \frac{\alpha_i}{\sum_{s_k \in D_{tgt}} \alpha_k}, \tag{8}$$

where $\mathbf{h}_i^{as}$ is the i^{th} sentence representation in the cited paper's abstract, $\mathbf{v}$, $\mathbf{W}_{\text{doc}}$, $\mathbf{W}_{\text{sent}}$ and $\mathbf{b}_{\text{sal}}$ are learnable parameters, and $\tilde{\alpha}_i$ is the salience score of the sentence s_i.

We then used the estimated salience of sentences in the cited paper's abstract to update the word-level attention of the cited paper's abstract $\{\mathbf{h}_j^{aw}\}$ so that the decoder can focus on these important sentences during text generation. Considering that the estimated salience $\tilde{\alpha}_i$ is a sentence weight, we determined each token in a sentence to share the same value of $\tilde{\alpha}_i$. Accordingly, the new attention $\overline{\mathbf{a}_t^{\text{abs}}}$ of the cited paper's abstract became $\overline{\mathbf{a}_t^{\text{abs}}} = \tilde{\alpha}_i \mathbf{a}_t^{\text{abs}}$. After normalizing $\overline{\mathbf{a}_t^{\text{abs}}}$, the context vector $\mathbf{c}_t^{\text{abs}}$ was updated accordingly.

4.3 Model Training

During model training, the objective of our framework covers three parts: generation loss, salience estimation loss, and citation function classification.

4.3.1 Generation Loss

The generation loss was based on the prediction of words from the decoder. We minimized the negative log-likelihood of all target words w_t^* and used them as the objective function of generation:

$$\mathcal{L}_{\text{gen}} = -\sum_t \log P(w_t^*). \tag{9}$$

4.3.2 Salience Estimation Loss

To include extra supervision into the salience estimation, we adopted a ROUGE-based approximation (Yasunaga et al., 2017) as the target. We assume citing sentences to depend heavily on salient sentences from the cited papers' abstracts. Based on this premise, we calculated the ROUGE scores between the citing sentence and sentences in the corresponding cited paper's abstract to obtain an approximation of the salience distribution as the ground-truth. If a sentence shared a high ROUGE score with the citing sentence, this sentence would be considered as a salient sentence because the citing sentence was likely to be generated based on this sentence, while a low ROUGE score implied that this sentence may be ignored during the generation process due to its low salience. Kullback–Leibler divergence was used as our loss function for enforcing the output salience distribution to be close to the normalized ROUGE score distribution of sentences in the cited paper's abstract:

$$\mathcal{L}_{\text{sal}} = D_{KL}(\mathcal{R} \| \tilde{\boldsymbol{\alpha}}), \tag{10}$$

$$\mathcal{R}_i = \frac{\beta r(s_i)}{\sum_{s_k \in D_{\text{tgt}}} \beta r(s_k)}, \tag{11}$$

where $\tilde{\boldsymbol{\alpha}}, \mathcal{R} \in \mathbb{R}^m$, $\mathcal{R}_i$ refers to the scalar indexed i in $\mathcal{R}$ $(1 \leq i \leq m)$, and $r(s_i)$ is the average of ROUGE-1 and ROUGE-2 F1 scores between the sentence s_i in the cited paper's abstract and the citing sentence. We also introduced a hyper-parameter β as a constant rescaling factor to sharpen the distribution.

4.3.3 Citation Function Classification

We added a supplementary component to enable the citation function classification to be trained with the generator, aiming to make the generation conscious of *why to cite*. Following a prior general pipeline of citation function classification (Cohan et al., 2019; Zhao et al., 2019), we first concatenated the last hidden state $\mathbf{s}_T$ of the decoder, which we considered as a representation of the generated citing sentence, with the document representation $\mathbf{d}_{\text{ctx}}$ of the citing sentence's context. Here, $\mathbf{d}_{\text{ctx}}$ was calculated as the average of its sentence representations. We then fed the concatenated representation into an

MLP followed by the softmax function to predict the probability of the citation function $\hat{\mathbf{y}}_{\text{func}}$ for the generated citing sentence. Cross-entropy loss was set as the objective function for training the classifier with the ground truth label $\mathbf{y}_{\text{func}}$, which is a one-hot vector:

$$\mathcal{L}_{\text{func}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{\text{func}}^{i}(j)\log \hat{y}_{\text{func}}^{i}(j), \quad (12)$$

where N refers to the size of training data and K is the number of different citation functions.

Finally, all aforementioned losses were combined as the training objective of the whole framework:

$$\mathcal{J}(\theta) = \mathcal{L}_{\text{gen}} + \lambda_{\mathcal{S}}\mathcal{L}_{\text{sal}} + \lambda_{\mathcal{F}}\mathcal{L}_{\text{func}}, \quad (13)$$

where $\lambda_{\mathcal{S}}$ and $\lambda_{\mathcal{F}}$ are the hyper-parameters to balance these losses.

5 Experiments

5.1 Metrics and Baselines

Following previous work, we report ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest common subsequence) scores to evaluate the generated citing sentences (Lin, 2004). Implementation details are shown in the Appendix, Section A.2. We also report ROUGE F1 score on our dataset. Finally, we compare our model to competitive baselines:

- **PTGEN** (See et al., 2017): is the original pointer-generator network.
- **EXT-Oracle** (Xing et al., 2020): selects the best possible sentence from the abstract of a cited paper that gives the highest ROUGE w.r.t. the ground truth. This method can be seen as an upper bound of extractive methods.
- **PTGEN-Cross** (Xing et al., 2020): enhances the original pointer-generator network with a cross attention mechanism applied to the citing sentence's context and the cited paper's abstract.

Additionally, we report results from using several extractive methods that have been used for summarization tasks[3], including:

- **LexRank** (Erkan and Radev, 2004): is an unsupervised graph-based method for computing relative importance of extractive summarization.
- **TextRank** (Mihalcea and Tarau, 2004): is an unsupervised algorithm where sentence importance scores are computed based on eigenvector centrality within weighted-graphs.

5.2 Experimental Results

As the results in Table 2 show, our proposed framework (BACO) outperformed all of the considered baselines. BACO achieved scores of 32.54 (ROUGE-1), 9.71 (ROUGE-2), and 24.90 (ROUGE-L). We also observed that the extractive methods performed comparatively poorly and notably worse than the abstractive methods. All abstractive methods did better than EXT-Oracle; a result different from performance on other summarization tasks, such as news document summarization. We think that this deviation from prior performance outcomes is because citing sentence in the domain of scholarly papers contain new expressions when referring to cited papers, which requires high-level summarizing or paraphrasing of cited papers instead of copying sentences verbatim from cited papers. Our results suggest that extractive methods may not be suitable for our task.

Among the extractive methods we tested, we observed EXT-Oracle to be superior to others, which aligns with our expectation of EXT-Oracle to serve as an upper bound of extractive methods. For abstractive methods, our framework achieved about 2.57 points improvement on ROUGE-2 F1 score compared to PTGEN-Cross. We assume two reasons for this improvement: First, BACO uses richer text features, e.g., *what to cite* (sentence salience estimation) and *why to cite* (citation function classification), that provide useful information for this task. Second, we included structural information from the citation network, which might offer supplemental *background knowledge* about a field that is not explicitly covered by the given cited and citing papers.

5.3 Ablation Study

We performed an ablation study to investigate the efficacy of the three main components in our framework: (1) we removed the node features (papers) that are output from the graph encoder to test the effectiveness of *background knowledge*; (2) we removed the predicted salience of sentences in the abstracts of cited papers to assess the effectiveness of one part of *content* (*what to cite*); and (3) we removed the training of citation function classification and only trained the generator to test the effectiveness of the other part of *content* (*why to cite*). As the removal of node features of papers re-

[3] We apply extractive methods on the cited paper's abstract to extract one sentence as the citing sentence.

Models	R-1	R-2	R-L
Extractive			
LexRank	11.96	1.04	9.69
TextRank	12.35	1.19	10.04
EXT-Oracle	22.60	4.21	16.83
Abstractive			
PTGEN	24.60	6.16	19.19
PTGEN-Cross*	27.08	7.14	20.61
BACO	**32.54**	**9.71**	**24.90**

Table 2: Experimental results for our framework and comparative models. *indicates our re-implementation.

Models	R-1	R-2	R-L
BACO	**32.54**	**9.71**	**24.90**
-w/o BK	31.02	8.90	22.46
-w/o SA	31.43	7.51	22.77
-w/o CF	30.84	8.67	23.31

Table 3: Ablation study for different components of our framework. w/o = without, BK = background knowledge, SA = salience estimation, CF = citation function.

duces the input to the dynamic fusion operation for the context vector (Equation 1), we changed Equation 2 to a sigmoid function so that the calculated attention becomes a vector of size 2 when combining the context vectors of the citing sentence's context and the cited paper's abstract.

Table 3 presents the results of the ablation study. We observed the ROUGE-2 F1 score to drop by 0.81 after the removal of the nodes (papers) feature. This indicates that considering *background knowledge* in a structured representation is useful for citing sentence generation. The ROUGE-2 F1 score dropped by 2.20 after disregarding salience of sentences in the cited paper. This implies that sentence-level salience estimation is beneficial, and it can be used to identify important sentences during the decoding phase so that the decoder can pay higher attention to those sentences. This process might also align with how scholars write citing sentences: they focus on specific parts or elements of cited papers, e.g., methods or results, and do not consider all parts equally when writing citing sentences. Lastly, the ROUGE-2 F1 score dropped by 1.04 after the removal of citation function classification; indicating that this feature is also helpful to the text generation task. We conclude that for a citing sentence generation, considering and training a model on background knowledge, sentence salience, and citation function improves the performance.

	Gold	BACO	PTGEN-Cross
Fluency	4.91	**3.64**	3.52
Relevance	4.86	**3.07**	2.64
Coherence	4.88	**2.77**	2.61
Overall	4.79	**2.95**	2.69

Table 4: Human evaluation results.

5.4 Case study

We present an illustrative example generated by our re-implementation of PTGEN-Cross versus by BACO, and compare both to ground truth (see Appendix, Section A.3). The output from BACO showed a higher overlap with the ground truth, specifically because it included background that is not explicitly covered in the cited paper. Furthermore, our output contained the correct citation function ("... have been shown to be effective"), which was present in the ground truth, but missing in PTGEN-Cross's output.

5.5 Human Evaluation

We sampled 50 instances from the generated texts. Three graduate students who are fluent in English and familiar with NLP were asked to rate citing sentences produced by BACO and the re-implemented PTGEN-Cross with respect to four aspects on a 1 (very poor) to 5 (excellent) point scale: fluency (whether a citing sentence is fluent), relevance (whether a citing sentence is relevant to the cited paper's abstract), coherence (whether a citing sentence is coherent within its context), and overall quality. Every instance was scored by the three judges, and we averaged their scores (Table 4). Our results showed that citing sentences generated by BACO score were generally better than output by PTGEN-Cross (e.g., Relevance score: BACO=3.07; PTGEN-Cross=2.64). This finding provided further evidence for the effectiveness of including the features we used for this task.

6 Conclusions and Future Work

We have brought together multiple pieces of information from and about cited and citing papers to improve citing sentence generation. We integrated them into **BACO**, a **BA**ckground knowledge- and

COntent-based framework for citing sentence generation, which learns and uses information that relate to (1) *background knowledge*; and (2) *content*. Extensive experimental results suggest that our framework outperforms competitive baseline models.

This work is limited in several ways. We only demonstrated the utility of our model within the standard RNN-based seq2seq framework. Secondly, our citation functions scheme only contained valence-based items. Finally, while this method is intended to support scholars in practicing strategic note taking on prior work with respect to a new literature review or research project, we did not evaluate the usefulness or effectiveness of this training option for researchers.

In future work, we plan to investigate the adaptation of our framework into more powerful models such as Transformer (Vaswani et al., 2017). We also hope to extend our citation functions scheme beyond valence of the citing sentences to more fine-grained categories, such as those outlined in Moravcsik and Murugesan (1975) and Lipetz (1965).

Impact Statement

This work is intended to support scholars in doing research, not to replace or automate any scholarly responsibilities. Finding, reading, understanding, reviewing, reflecting upon, and properly citing literature are key components of the research process and require deep intellectual engagement, which remains a human task. The presented approach is meant to help scholars to see examples for how to strategically synthesize scientific papers relevant to a certain topic or research problem, thereby helping them to cope with information overload (or "research deluge") and honing their scholarly writing skills. Additional professional responsibilities also still apply, such as not violating intellectual property/ copyright issues.

We believe that this work does not present foreseeable negative societal consequence. While not intended, our method may be misused for the automated generation of parts of literature reviews. We strongly discourage this misuse as it violates basic assumptions about scholarly diligence, responsibilities, and expectations. We advocate for our method to be used as a scientific writing training tool.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Kimitaka Asatani, Junichiro Mori, Masanao Ochi, and Ichiro Sakata. 2018. Detecting trends in academic research from a citation network using network representation learning. *PloS one*, 13(5):e0197260.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Mark W Bruner, Karl Erickson, Brian Wilson, and Jean Côté. 2010. An appraisal of athlete development models through citation network analysis. *Psychology of sport and exercise*, 11(2):133–139.

Jingqiang Chen and Hai Zhuge. 2016. Summarization of related work through citations. In *12th International Conference on Semantics, Knowledge and Grids (SKG)*. IEEE.

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

Kaihua Chen and Jiancheng Guan. 2011. A bibliometric investigation of research performance in emerging nanobiopharmaceuticals. *Journal of Informetrics*, 5(2):233–247.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

John Cullars. 1990. Citation characteristics of italian and spanish literary monographs. *The Library Quarterly*, 60(4):337–356.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9):1820–1833.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.

Jianhua Hou, Xiucai Yang, and Chaomei Chen. 2018. Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics*, 115(2):869–892.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: an optimization approach. In *Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xin Huang, Chang-an Chen, Changhuan Peng, Xudong Wu, Luoyi Fu, and Xinbing Wang. 2018. Topic-sensitive influential paper discovery in citation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*.

Xuerong Li, Han Qiao, and Shouyang Wang. 2017. Exploring evolution and emerging trends in business model study: a co-citation analysis. *Scientometrics*, 111(2):869–887.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ben-Ami Lipetz. 1965. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16(2):81–90.

Kobra Mansourizadeh and Ummul K. Ahmad. 2011. Citation practices among non-native expert and novice scientific writers. *Journal of English for Academic Purposes*, 10(3):152–161.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018a. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018b. Multimodal named entity recognition for short social media posts. In *Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Spencer Moore, Alan Shiell, Penelope Hawe, and Valerie A Haines. 2005. The privileging of communitarian ideas: citation practices and the translation of social capital into public health research. *American Journal of Public Health*, 95(8):1330–1337.

Michael J Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86–92.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Horacio Saggion, Alexander Shvets, Àlex Bravo, et al. 2020. Automatic related work section generation: experiments in scientific document abstracting. *Scientometrics*, 125(3):3159–3185.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Linda C Smith. 1981. Citation analysis. *Library Trends*, 30(3):83–106.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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 *Proceedings of 31st International Conference on Neural Information Processing Systems (NeurIPS)*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Yongzhen Wang, Xiaozhong Liu, and Zheng Gao. 2018. Neural related work summarization with a joint context-driven attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Howard D White. 2004. Citation analysis and discourse analysis revisited. *Applied linguistics*, 25(1):89–116.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R 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 *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of 21st Conference on Computational Natural Language Learning (CoNLL)*.

Hanlin You, Mengjun Li, Keith W Hipel, Jiang Jiang, Bingfeng Ge, and Hante Duan. 2017. Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics*, 111(1):297–315.

He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A context-based framework for modeling the role and function of on-line resource citations in scientific literature. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

A Appendices

A.1 Experiments for Citation Function Labeling Model

To test our citation function labeling model, we applied 10-fold cross-validation to our training dataset with 800 citing sentences. We then tested our trained model on the test data with 400 sentences, which we refer to as the external test set.

We set the hidden size of the MLP in our labeling model to 256, and adopted a dropout with a rate of 0.2. For the optimizer, an Adam (Kingma and Ba, 2015) with a learning rate of 2e-3 was used. The batch size was set to 8. We used F1 score for evaluating labeling accuracy. Since there was imbalance among the distributions of labels, we choose the micro-F1 score specifically. The results are shown in Table 5. After training, we used the trained model to label the rest of the data (84,376 instances) for further training the citing sentence generation model. The final dataset contains 85,576 instances. Following (Xing et al., 2020), we used the above-mentioned 400 citing sentences as the test set, and combined the 800 citing sentences with the rest of the model-labelled instances as our training set. The average length of citing sentences in the training and test data is 28.72 words and 26.45 words, respectively.

	Precision	**Recall**	**F1-score**
cross	95.43	95.43	95.43
test	91.59	91.59	91.59

Table 5: The results of the citation function labeling model for cross-validation (denoted as cross) and on the external test data (denoted as test)

A.2 Implementation Details

We used pre-trained *Glove* vectors (Pennington et al., 2014) to initialize word embeddings with the vector dimension 300 and followed Veličković et al. (2018) to initialize the node features for the graph encoder as a bag-of-words representation of the paper's abstract. The hidden state size of LSTM was set to 256 for the encoder, and 512 for the decoder. An AdaGrad (Duchi et al., 2011) optimizer was used with a learning rate of 0.15 and an initial accumulator value of 0.1. We picked 64 as the batch size for training. For the rescaling factor β in Equation 11, we chose 40 based on the results reported in Yasunaga et al. (2017). We also used ROUGE scores (Lin, 2004) for quantitative evaluation, and reported the F_1 scores of ROUGE-1, ROUGE-2 and ROUGE-L for comparing BACO to alternative models.

A.3 Case Study

We present an example generated by our re-implementation of the baseline PTGEN-Cross, and our framework in Table 6. Note that we used `#REFR` to mark the citation of the cited paper. The reference signs to other papers are masked as `#OTHEREFR`. The `#CITE` in context indicates the position where the citing sentence should be inserted. The output of our framework has a higher overlap with the ground truth than the output from PTGEN-Cross. Please note that our framework was able to infer that the mentioned methods in the generated citing sentence are "supervised", and we believe that this knowledge was gained from the citation network where other neighboring cited papers explicitly mentioned "supervised methods". Also, the generated citing sentence from our framework showed a positive citation function (... *have been shown to be effective*) as the ground truth, while PTGEN-Cross's output expressed the wrong citation function (neutral). We think the underlying reason for this difference in outputs may be that our joint training of citing sentence generation and citation function classification, which forced our framework to recognize the corresponding citation function during the generation and further improved the performance.

Citing Paper's Abstract	The state-of-the-art methods used for relation classification are primarily based on statistical machine learning, and their performance strongly depends on the quality of the extracted features. The extracted features are often derived from the output of pre-existing natural language processing (NLP) systems, which leads to the propagation of the errors in the existing tools and hinders the performance of these systems. In this paper, we exploit a convolutional deep neural network (DNN) to extract lexical and sentence level features. our method takes all of the word tokens as input without complicated pre-processing. First, the word tokens are transformed to vectors by looking up word embeddings. Then, lexical level features are extracted according to the given nouns. Meanwhile, sentence level features are learned using a convolutional approach. These two level features are concatenated to form the final extracted feature vector. Finally, the features are fed into a softmax classifier to predict the relationship between two marked nouns. The experimental results demonstrate that our approach significantly outperforms the state-of-the-art methods.
Cited Paper's Abstract	We present a novel approach to relation extraction, based on the observation that the information required to assert a relationship between two named entities in the same sentence is typically captured by the shortest path between the two entities in the dependency graph. Experiments on extracting top-level relations from the ace (automated content extraction) newspaper corpus show that the new shortest path dependency kernel outperforms a recent approach based on dependency tree kernels.
Context	The task of relation classification is to predict semantic relations between pairs of nominals and can be defined as follows: given a sentence S with the annotated pairs of nominals e and e , we aim to identify the relations between e and e `#OTHREFR`. There is considerable interest in automatic relation classification, both as an end in itself and as an intermediate step in a variety of NLP applications. `#CITE`. Supervised approaches are further divided into feature-based methods and kernel-based methods. Feature-based methods use a set of features that are selected after performing textual analysis. They convert these features into symbolic IDs, which are then transformed into a vector using a paradigm that is similar to the bag-of-words model.
Ground Truth	The most representative methods for relation classification use supervised paradigm ; such methods have been shown to be effective and yield relatively high performance `#OTHREFR`; `#REFR`.
PTGEN-Cross	There has been a wide body of approaches to predict relation extraction between nominals `#OTHREFR`; `#REFR`.
BACO	Most methods for relation classification are supervised and have been shown to be effective `#OTHREFR`; `#REFR`.

Table 6: Example output citing sentences