

# On the Use of Context for Predicting Citation Worthiness of Sentences in Scholarly Articles

Rakesh Gosangi, Ravneet Arora, Mohsen Gheisarieha,  
Debanjan Mahata, Haimin Zhang

Bloomberg, New York

{rgosangi, rarora62, mgheisarieha, dmahata}@bloomberg.net,  
raymondzhang@gmail.com

## Abstract

In this paper, we study the importance of context in predicting the citation worthiness of sentences in scholarly articles. We formulate this problem as a sequence labeling task solved using a hierarchical BiLSTM model. We contribute a new benchmark dataset containing over two million sentences and their corresponding labels. We preserve the sentence order in this dataset and perform document-level train/test splits, which importantly allows incorporating contextual information in the modeling process. We evaluate the proposed approach on three benchmark datasets. Our results quantify the benefits of using context and contextual embeddings for citation worthiness. Lastly, through error analysis, we provide insights into cases where context plays an essential role in predicting citation worthiness.

## 1 Introduction and Background

Citation worthiness is an emerging research topic in the natural language processing (NLP) domain, where the goal is to determine if a sentence in a scientific article requires a citation<sup>1</sup>. This research has potential applications in citation recommendation systems (Strohman et al., 2007; Küçüktunç et al., 2014; He et al., 2010), and is also useful for scientific publishers to regularize the citation process. Providing appropriate citations is critical to scientific writing because it helps readers understand how the current work relates to existing research.

Citation worthiness was first introduced by (Sugiyama et al., 2010), where the authors formulated as a sentence-level binary classification task solved using classical machine learning techniques like Support Vector Machines (SVMs). Subsequent works from (Färber et al., 2018; Bonab et al., 2018) use similar approach but employ deep

learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). More recently, Zeng et al. (Zeng and Acuna, 2020) proposed a Bidirectional Long short-term memory (BiLSTM) based architecture and demonstrated that context, specifically the two adjacent sentences, can help improve the prediction of citation worthiness.

Citation worthiness is closely related to citation recommendation (suggest a reference for a sentence in a scientific article), which is often approached as a ranking problem solved using models that combine textual, contextual, and document-level features (Strohman et al., 2007; He et al., 2010). More recent works employ deep learning models (Huang et al., 2015; Ebesu and Fang, 2017) and personalization (Cai et al., 2018; Yang et al., 2018; Cai et al., 2018). Citation analysis (Athar and Teufel, 2012) and citation function (Teufel et al., 2006; Li et al., 2013; Hernandez-Alvarez et al., 2017), other closely related domains, aim to predict the sentiment and motivation of a citation respectively. Researchers have used many supervised approaches like sequence labeling (Abu-Jbara et al., 2013), structure-based prediction (Athar, 2011), and multi-task learning (Yousif et al., 2019) to address these problems.

In this paper, we want to investigate two research question about citation worthiness. First, we posit that citation worthiness is not purely a sentence-level classification task because the surrounding context could influence if a sentence requires a citation. This context could include not only adjacent sentences but also information about the section titles, paragraphs, and other included citations. Previous work (Bonab et al., 2018) has explored using the adjacent two sentences; we predict that citation worthiness models would improve with access to more contextual information. To pursue this hypothesis, we propose two new formulations: (a) sentence pair classification and (b) sentence se-

<sup>1</sup>For example, in the first excerpt in Table 1, the goal is to predict that the first, third, and fourth sentences would require citations but the second does not.

The baseline TTS systems in the project utilize the HTS toolkit which is built on top of the HTK framework [Cite]. The HMM-based TTS systems have been developed for Finnish, English, Mandarin and Japanese [No Cite]. The systems include an average voice model for each language trained over hundreds of speakers taken from standard ASR corpora, such as Speecon [Cite]. Using speaker adaptation transforms, thousands of new voices have been created and new voices can be added using a small number of either supervised or unsupervised speech samples [Cite].

The LexRank method addressed was very successful in generic multi-document summarization [Cite]. A topic-sensitive LexRank is proposed [Cite]. As in LexRank, the set of sentences in a document cluster is represented as a graph, where nodes are sentences and links between the nodes are induced by a similarity relation between the sentences [No Cite].

Table 1: The excerpts are obtained from (Kurimo et al., 2010) and (Chali et al., 2009). The actual citations have been removed but we include citation worthiness labels.

quence modeling. For the latter formulation, we propose a new hierarchical architecture, where the first layer provides sentence-level representations, and the second layer predicts the citation worthiness of the sentence sequence. We also introduce a new dataset mostly because the prior datasets (Bonab et al., 2018; Zeng and Acuna, 2020) do not have sufficient contextual information to study this research question.

The second research objective is to understand if contextual embedding models would help citation worthiness. Recent developments in language modeling, specifically contextual embedding models (Liu et al., 2019; Devlin et al., 2019), have already demonstrated significant improvements in various NLP research tasks. We expect to observe similar gains in citation worthiness. Following is a summary of the main contributions of this work:

- We propose two new formulations for citation worthiness: sentence-pair classification and sentence sequence modeling.
- We contribute a new dataset containing significantly more context, and we expect it to serve as another benchmark.
- Through rigorous experimental work, we demonstrate the benefits of sequential modeling and contextual embeddings for citation worthiness.
- We obtain new state-of-the-art results on three benchmark datasets.

## 2 Methodology

### 2.1 Problem Statement

Let  $d = \{s_1, s_2, \dots, s_n\}$  be a scientific article, where  $s_i$  is the  $i^{th}$  sentence. The problem of citation worthiness is to assign each sentence  $s_i$  one of two possible labels  $L = \{l_c, l_n\}$ , where  $l_c$  denotes that sentence requires a citation and  $l_n$  means otherwise. We present three different formulations here to investigate our main research objectives.

### 2.2 Sentence Classification

Our first formulation (Figure 1 (a)) approaches citation worthiness as a sentence level classification task similar to prior works of (Bonab et al., 2018; Färber et al., 2018). Given a sentence  $s_i$ , we map it to a fixed-size dense vector  $x_i$  using contextual embedding models (e.g. BERT (Devlin et al., 2019)). We then feed  $x_i$  to a feed-forward layer to obtain the citation worthiness label. We fine-tune this entire architecture by optimizing the weights of the final layer.

### 2.3 Sentence-Pair Classification

Our second approach (Figure 1 (b)) is to formulate citation worthiness as a sentence-pair classification task, where the pair consists of the given sentence and a sentence-like representation of the context. Namely, for a given sentence  $s_i$ , we define the context  $c_i$  as the concatenation of the previous  $s_{i-1}$ ,  $s_i$ , and the next sentence  $s_{i+1}$ :

$$c_i = [s_{i-1}; s_i; s_{i+1}] \quad (1)$$

We then concatenate  $s_i$  with  $c_i$  separated by the [SEP] token, pass it through the embedding layer to obtain a vector representation  $x_i$ . This vector is then passed through a feed-forward layer to obtain the class label. This approach is similar to (Zeng and Acuna, 2020), where the authors used Glove embeddings (Pennington et al., 2014) to obtain sentence representations, and BiLSTMs for context representations. This formulation has also been used previously for question-answering (Devlin et al., 2019) and passage re-ranking (Nogueira and Cho, 2019). In our sentence-pair classification approach, we defined  $c_i$  to include only two adjacent sentences, but it could easily include more. However, if we included too many sentences, the context might be too long for most transformer-based models.

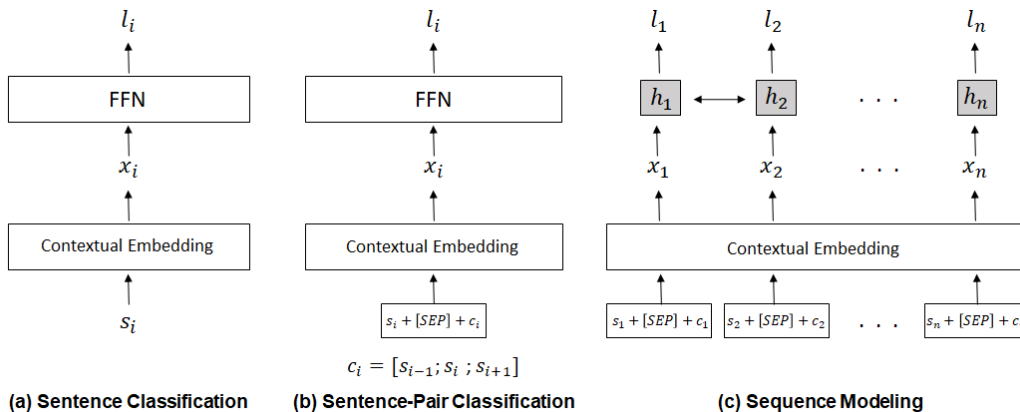


Figure 1: The three proposed architectures.

## 2.4 Sentence Sequence Modeling

The third formulation addresses citation worthiness as a sequence labeling task solved using a hierarchical BiLSTM architecture (Figure 1 (c)). We first map each sentence  $s_i$  and context  $c_i$  (eq. 1) to a fixed-size dense vector  $x_i$  using the same approach as in section 2.3. Thus the given document  $d$  is represented as a sequence of vectors  $x = \{x_1, x_2, \dots, x_n\}$ . We then feed these vectors to a BiLSTM model to capture the sequential relations between the sentences. The hidden state of the BiLSTM  $h_i$  provides a vector representation for sentence  $s_i$  that incorporates information from the surrounding sentences. Thus the sequence modeling approach captures long term dependencies between the sentences without us needing to explicitly encode them as extra features. We then use a feed-forward layer to map the BiLSTM output to the citation worthiness labels. We have also experimented with using only the sentence  $s_i$  to construct the vector  $x_i$ . However, we observed that using the context  $c_i$  helped improve the model performance.

## 3 Experimental Work

### 3.1 Datasets

Prior works in citation worthiness presented two benchmark datasets: SEPID-cite (Bonab et al., 2018) and PMOA-cite (Zeng and Acuna, 2020). SEPID-cite contains 1,228,053 sentences extracted from 10,921 articles<sup>2</sup> but does not contain the source of the sentences (e.g. paper id) or the sentence order. PMOA-cite contains 1,008,042 sentences extracted from 6,754 papers from PubMed open access. PMOA-cite also contains the pre-

ceding sentence, the next sentence, and the section header. However, the authors of PMOA-cite did data splits at sentence-level, which means sentences from the same research paper could be part of train and test datasets. Since we cannot use either one of these datasets directly for sequence modeling, we chose to process the ACL Anthology Reference Corpus (Bird et al., 2008)<sup>3</sup> (ACL-ARC) while preserving the sentence order, and then split the data at document-level.

The latest version of ACL-ARC (Bird et al., 2008), released in 2015, contains 22,878 articles. Each article here contains the full text and metadata such as author names, section headers, and references. We first processed this corpus to exclude all articles without abstracts because they typically were conference cover sheets. Then, for each section in an article, we extracted paragraph information based on newlines. Then, we split the paragraphs into constituent sentences<sup>4</sup> and processed the sentences to obtain citation labels based on regular expressions (Appendix A). We then sanitized the sentences to remove all the citation patterns.

The resulting new corpus (ACL-cite<sup>5</sup>) contained 2,706,792 sentences from 17,440 documents of which 305,733 sentences (11.3%) had citations. Lastly, we performed document-level splits: training (10,464 docs, 1,625,268 sentences), validation (3,487 docs, 539,085 sentences), and test (3,487 docs, 542,081 sentences). To validate our citation regular expressions, we manually annotated a random sample of 500 sentences and observed only one error in the extracted labels. Table 3 provides some basic statistics on the three datasets.

<sup>3</sup><https://www.aclweb.org/anthology/>

<sup>4</sup><https://github.com/fnl/segtok>

<sup>5</sup>Dataset available at: <https://zenodo.org/record/4651554>

<sup>2</sup><http://pars.ie/lr/sep-id-corpus>

Model	P	R	F1	W-F1
<b>SEPID-cite</b>				
CNN + Glove (f)	0.196	0.269	0.227	N/A
CNN + w2v (b)	0.448	0.405	0.426	N/A
BiLSTM + Glove (z)	<b>0.720</b>	0.391	0.507	N/A
SC + Roberta	0.674	<b>0.595</b>	<b>0.629</b>	0.950
<b>PMOA-cite</b>				
BiLSTM + Glove (z)	0.883	0.795	0.837	N/A
SC + Roberta	<b>0.887</b>	<b>0.814</b>	<b>0.849</b>	0.942
BiLSTM + Glove + context (z)	<b>0.907</b>	0.811	0.856	N/A
SPC + Roberta	0.905	<b>0.843</b>	<b>0.873</b>	0.962
<b>ACL-cite</b>				
SC + Roberta	0.782	0.496	0.607	0.921
SPC + Roberta	0.820	0.562	0.667	0.932
SSM (4 sentences) + Roberta	<b>0.821</b>	0.635	0.716	0.940
SSM (8 sentences) + Roberta	0.816	0.643	0.719	0.940
SSM (16 sentences) + Roberta	0.809	0.654	0.723	0.941
SSM (16 sentences + section) + Roberta	0.813	<b>0.662</b>	<b>0.730</b>	<b>0.942</b>

Table 2: SC: Sentence Classification, SPC: Sentence-Pair Classification, SSM: Sentence Sequence Modeling. (f) denotes that the numbers reported in (Färber et al., 2018), (b) in (Bonab et al., 2018), and (z) in (Zeng and Acuna, 2020). P, R, and F1 are precision, recall, and F1 scores for the cite class ( $l_c$ ). W-F1 is the weighted-F1 on entire test dataset.

	SEPID-cite	PMOA-cite	ACL-cite
Articles	10,921	6,754	17,440
Sections	-	32,198	130,603
Paragraphs	-	202,047	934,502
Sentences	1,228,053	1,008,042	2,706,792
S w/o c	1,142,275	811,659	2,401,059
S w c	85,778	196,383	305,733
A c/s	131	132	141
A w/s	22	20	22

Table 3: Statistics of the three datasets. S w/o c: Sentences without citation, S w c: Sentences with citation, A c/s: Average characters per sentence, A w/s: Average words per sentence.

### 3.2 Experimental settings

We applied the sentence classification (SC) model on all three datasets, sentence-pair classification (SPC) model on PMOA-cite and ACL-cite, and sentence sequence modeling (SSM) approach on ACL-cite. This is because SEPID-cite does not have any context to apply SPC or SSM, and PMOA-cite does not have sufficient context for SSM.

To obtain sentence representations, we also explored the idea of pooling word-level embeddings obtained using CNNs. However, we observed no significant difference in the model performance when compared to using the [CLS] token. We also experimented with the choice of contextual embeddings: BERT (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), Roberta (Liu et al., 2019), and XLnet (Yang et al., 2019) and observed that the Roberta model consistently gave the best results; therefore, we only report those numbers.

We used a batched training approach for the SSM models: split each article into sequences of  $m$  with an overlap of  $m/2$  sentences. For example, consider a document with 32 sentences and  $m = 16$ , we create three training sequences; first sequence: sentences 1 to 16, second sequence: sentences 9 to 24, and so on. During inference, for a given sentence, we include the preceding  $m/2$  sentences and the succeeding  $m/2 - 1$  sentences<sup>6</sup>. We trained and evaluated models at different values of  $m = 4, 8, 16$ . We trained all the models using the Adam optimizer with a batch size of 16, a learning rate of  $1e-5$ , a maximum of 4 epochs to optimize for cross-entropy loss. The hidden layers in the BiLSTM models were set to 128 units. The models were trained on a GPU machine with 6 cores and each training epoch took approximately 4 hours. More details on the experimental settings are available in the Appendix.

### 3.3 Results

Table 2 summarizes the results in terms of the precision, recall, F1 score for  $l_c$ , and overall weighted F1 score. The baseline numbers reported here are either from prior works (Färber et al., 2018; Bonab et al., 2018; Zeng and Acuna, 2020) or based on architectures very similar to those used in these prior works. On the SEPID-cite dataset, our SC model obtained significantly better performance than the state-of-the-art results from (Zeng and

<sup>6</sup>We used zero-padding in cases with insufficient context, e.g., beginning or end of a document.

Section	SC			SSM		
	P	R	F1	P	R	F1
Abstract	0.340	0.505	0.407	0.543	0.576	<b>0.559</b>
Acknowledgments	0.874	0.361	0.511	0.759	0.480	<b>0.588</b>
Conclusion	0.585	0.459	0.515	0.711	0.560	<b>0.626</b>
Evaluation	0.770	0.538	0.633	0.808	0.659	<b>0.726</b>
Introduction	0.833	0.514	0.636	0.831	0.645	<b>0.726</b>
Methods	0.791	0.525	0.631	0.803	0.650	<b>0.718</b>
Related Work	0.901	0.707	0.792	0.918	0.827	<b>0.870</b>

Table 4: Comparison of the F1 score of SC and SSM models by section.

Acuna, 2020) with the F1 score increasing by more than 12%. On the PMOA-cite dataset, we obtain an F1 gain of 1.2% for sentence-level and 1.7% for contextual models. We indicate that the numbers from (Zeng and Acuna, 2020) use additional hand-crafted contextual features, including labels of surrounding sentences, but our models only use textual features.

The results on the ACL-cite dataset clearly show the importance of context in this domain. The use of surrounding two sentences boosted the performance by nearly 6% points, and the performance continues to improve with added context increasing by another 5.6% points for 16 sentences. The model performance improves by another 0.7% with the inclusion of section headers in the context.

Table 4 compares the performance of the SC and SSM models for different sections in the papers. The F1 score improves for all but most prominent for *Abstract* and *Conclusion* sections because of significant improvements in the precision.

### 3.4 Subjective Analysis

We observed some interesting trends during the error-analysis of the SC and SSM models. We categorized these trends into three groups and selected an example from each to illustrate the impact of context (Table 1).

- *Prior works*: In the first excerpt, the last sentence could be interpreted as the author’s contribution if no context was available. The preceding sentences in the paragraph seem to help the model understand that this sentence requires a citation because it refers to prior work.
- *Sections*: In the second excerpt, the second sentence could be interpreted as an introduction or conclusion. Once again, the context provides information to infer the section correctly and, therefore, the correct label.
- *Topic sentences*: Context is essential to understand if a sentence is the first statement about a topic, typically when researcher provide citations,

or continuation of a discussion. In the second excerpt, the model does not predict  $l_c$  for the last sentence because the authors already introduced the concept *LexRank* in previous sentences.

## 4 Conclusions

In this paper, we study the impact of context and contextual models on citation worthiness. We propose two new formulations for this problem: sentence-pair classification and sentence sequence modeling. We contribute a new benchmark dataset with document-level train/dev/test splits, which enables to incorporate contextual information better. We propose a hierarchical BiLSTM approach for sequence modeling, but we could also consider a transformer-based approach and further improve with a CRF layer. Likewise, we also want to consider some of the newer language models (Zaheer et al., 2020; Beltagy et al., 2020) that handle longer sentences.

We expect citation worthiness would be an important part of developing writing assistants for scientific documents. We studied the citation worthiness of sentences in scholarly articles in this paper, but we believe these findings are relevant to other domains like news, Wikipedia, and legal documents.

## References

- Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. [Purpose and polarity of citation: Towards NLP-based bibliometrics](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.
- Awais Athar. 2011. [Sentiment analysis of citations using sentence structure-based features](#). In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

- Awais Athar and Simone Teufel. 2012. [Context-enhanced citation sentiment detection](#). In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601, Montréal, Canada. Association for Computational Linguistics.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. [The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Hamed Bonab, Hamed Zamani, Erik Learned-Miller, and James Allan. 2018. Citation worthiness of sentences in scientific reports. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1061–1064.
- X. Cai, J. Han, W. Li, R. Zhang, S. Pan, and L. Yang. 2018. [A three-layered mutually reinforced model for personalized citation recommendation](#). *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6026–6037.
- Xiaoyan Cai, Junwei Han, and Libin Yang. 2018. [Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Yllias Chali, Shafiq R. Joty, and Sadid A. Hasan. 2009. Complex question answering: Unsupervised learning approaches and experiments. *Journal of Artificial Intelligence Research*, 35(1):1–47.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Travis Ebesu and Yi Fang. 2017. [Neural citation network for context-aware citation recommendation](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*, page 1093–1096, New York, NY, USA. Association for Computing Machinery.
- Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. To cite, or not to cite? detecting citation contexts in text. In *Advances in Information Retrieval*, pages 598–603, Cham. Springer International Publishing.
- Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. [Context-aware citation recommendation](#). In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, page 421–430, New York, NY, USA. Association for Computing Machinery.
- Myriam Hernandez-Alvarez, José M Gomez Soriano, and Patricio Martínez-Barco. 2017. Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4):561.
- Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A neural probabilistic model for context based citation recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, page 2404–2410. AAAI Press.
- Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V. Çatalyürek. 2014. [Diversifying citation recommendations](#). *ACM Trans. Intell. Syst. Technol.*, 5(4).
- Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, Matthew Gibson, Yong Guan, Teemu Hirsimäki, Reima Karhila, Simon King, Hui Liang, Keiichi Oura, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi-Jian Wu, and Junichi Yamagishi. 2010. Personalising speech-to-speech translation in the emime project. In *Proceedings of the ACL 2010 System Demonstrations, ACLDemos '10*, page 48–53, USA. Association for Computational Linguistics.
- Xiang Li, Yifan He, Adam Meyers, and Ralph Grishman. 2013. [Towards fine-grained citation function classification](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 402–407, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Trevor Strohman, W. Bruce Croft, and David Jensen. 2007. [Recommending citations for academic papers](#). In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07*, page 705–706, New York, NY, USA. Association for Computing Machinery.
- K. Sugiyama, T. Kumar, M. Kan, and R. C. Tripathi. 2010. [Identifying citing sentences in research papers using supervised learning](#). In *2010 International Conference on Information Retrieval Knowledge Management (CAMP)*, pages 67–72.
- Simone Teufel, Advait Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06*, page 103–110, USA. Association for Computational Linguistics.
- L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, L. Guo, and T. Dai. 2018. [A lstm based model for personalized context-aware citation recommendation](#). *IEEE Access*, 6:59618–59627.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.
- Abdallah Yousif, Zhendong Niu, James Chambua, and Zahid Younas Khan. 2019. [Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification](#). *Neurocomputing*, 335:195–205.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. [Big bird: Transformers for longer sequences](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Tong Zeng and Daniel E. Acuna. 2020. [Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models](#). *Scientometrics*, 124(1):399–428.