Additional Context Helps! Leveraging Cited Paper Information to Improve Citation Classification

Kamal Kaushik Varanasi[1], Tirthankar Ghosal[2], and Valia Kordoni[3]

[1] kamalkaushikv@gmail.com, [2] tirthankar.pcs16@iitp.ac.in
Indian Institute of Technology Patna, Bihar (India)
[3] evangelia.kordoni@anglistik.hu-berlin.de
Humboldt Universität zu Berlin (Germany)

Abstract

With the rapid growth in research publications, automated solutions to tackle scholarly information overload is growing more relevant. Correctly identifying the intent of the citations is one such task that finds applications ranging from predicting scholarly impact, finding idea propagation, to text summarization to establishing more informative citation indexers. In this in-progress work, we leverage the cited paper's information and demonstrate that this helps in the effective classification of citation intents. We propose a neural multi-task learning framework that harnesses the structural information of the research papers and the relation between the citation context and the cited paper for citation classification. Our initial experiments on three benchmark citation classification datasets show that with incorporating cited paper information (title), our neural model achieves a new state of the art on the ACL-ARC dataset with an absolute increase of **5.3%** in the F1 score over the previous best model. Our approach also outperforms the submissions made in the 3C Shared task: Citation Context Classification with an increase of **8%** and **3.6%** over the previous best Public F1-macro and Private F1-macro scores respectively.

Introduction

Citations are crucial in analyzing scientific works and for understanding the link between different research articles. They act as trackers of the direction of research in a field and as an important measure in understanding the impact of research articles, venues, researchers, etc. Citations may also have different nature. Authors may cite a research publication in different ways. For example - a citation might indicate motivation or usage of a method from a previous work or a comparison of results of various works. So, identification of the intent behind a citation is crucial for automated analysis of academic literature. Most of the research works in the field of citation classification provide too fine-grained citation categories, example-(Stevens and Giuliano (1965); Moravcsik and Murugesan (1975)), so only a handful of these are used for automated analysis of the scientific publications. To overcome these problems, Jurgens et al. (2018) proposed a six category classification scheme. Then, Cohan et al. (2019) used a different scheme that had only three classification categories. More recently, Pride et al. (2020) proposed a classification scheme similar to Jurgens et al. (2018).

Table 1. Examples of citations with cited paper titles and intents.

Citation context	*Cited Paper Title*	*True Label*
She evaluates 3,000 German verbs with a token frequency between 10 and 2,000 against the Duden (@@CITATION).	duden—das stilworterbuch duden—the style dictionary	BACKGROUND

Jurgens et al. (2018) used a set of engineered features like (1) Pattern based features (2) Topic based features (3) Prototypical Argument features for this task. While recently, Cohan et al. (2019) argued that features based on the structural properties related to scientific literature are more effective than the predefined hand engineered domain-dependent features or external resources. We argue that in addition to leveraging the structural information related to the scientific discourse, utilizing the cited paper information as additional context can significantly improve the performance. In the example from table 1, it is evident that the instance seems less ambiguous and easier to classify after accessing the cited paper title information in addition to the citation context. To tackle these problems, we propose a Multi Task Learning framework that incorporates three scaffolds, including a cited paper title scaffold that leverages the relationship between the citation context and the cited paper title. The other two scaffolds are the structural scaffolds to leverage the relationship between the structure of the research papers and the intent of the citations. These two scaffolds are inspired by the work done in Cohan et al. (2019). We explain these scaffolds in detail in Table 4 under the Model Section.

Dataset Description

Table 2 shows the classification categories of different datasets and Table 3 shows the corresponding data statistics. Please note that we retrieve the cited paper title scaffold data from the target datasets and the SciCite dataset includes the data corresponding to the structural scaffolds.

Table 2.Intent categories of different datasets.

Dataset	*Citation Intent Categories*
SciCite	BACKGROUND, METHOD, RESULT_COMPARISON
ACL-ARC/3C Challenge dataset	BACKGROUND, USES, COMPARE_CONTRAST, MOTIVATION, EXTENSION, FUTURE.
Section title scaffold data (91412 instances)	INTRODUCTION, CONCLUSION, EXPERIMENTS, METHOD, RELATED WORK
Citation worthiness scaffold data (73484 instances)	TRUE, FALSE

Table 3.Cross-study comparison of different datasets.

Dataset	*Papers*	*Annotated by*	*Citations*	*Intent categories*	*Discipline(s)*
SciCite	6,627	Volunteers	11,020	3	Comp. Sci/Medicine
3C Challenge	883	Paper authors	3,000	6	Multi-disciplinary
ACL-ARC	185	Domain Experts	1,989	6	Comp. Science

Model

We propose a Multitask Learning (Caruana, 1997) model with the main task of citation intent classification along with a total of three auxiliary tasks (scaffolds). The knowledge acquired by the auxiliary tasks helps the model to learn optimal parameters for the main task.

Table 4. Scaffolds in our Multi-tasking Approach

Scaffolds	*Description*
Section Title	This task is related to predicting the section under which the citation occurs, given a citation context. In general, researchers follow a standard order while presenting their scientific work in the form of sections. Citations may have different nature according to the section under which they are cited. Hence, the intent of the citation and the section are related to each other. For example, the results-comparison related citations are often cited under the Results section.
Citation Worthiness	This task is related to predicting whether a sentence needs a citation or not, i.e., it is the task of classifying whether a sentence is a citation text or not.
Cited Paper Title	Sometimes a citation context might be ambiguous, making it difficult to predict the intent of the citation correctly. In such cases, information from the cited paper like the abstract of the paper, title of the paper, etc. may provide some additional context that can assist in identifying the appropriate intent behind that citation. This auxiliary task helps the model to learn these nuances by leveraging the relationship between the citation context and the cited paper. We use a concatenated vector of citation context and the cited paper title fields from the target dataset as the input for this task.

We use these auxiliary tasks only while training/fine-tuning the model. Our model architecture is shown in Figure 1.

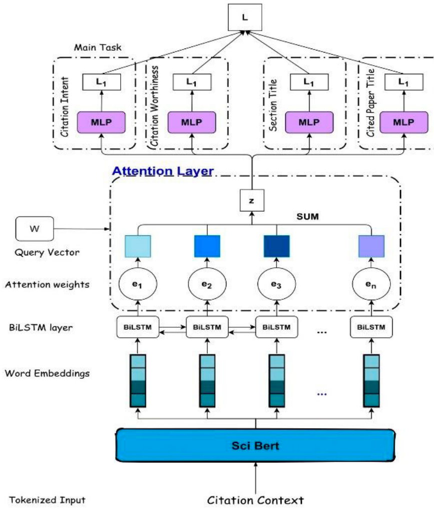

Figure 1. The architecture of our proposed model. The main task MLP is for prediction of citation intents (top left) followed by three MLPs for section title, citation worthiness, and cited paper title scaffolds

Model Structure

Let C be the tokenized citation context of size n. We pass it onto the SciBERT (Beltagy, 2019) model with pre-trained weights to get the word embeddings of size (n,d_1) i.e. we have the output as x = { x_1, x_2, x_3,……..x_n } where $x_i \in R^{d1}$. Then we use a Bidirectional long short-term memory (BiLSTM) network with a hidden size d_2 to get an output vector h of size (n,2d_2). We pass h to the dot-product attention layer with query vector w to get an output vector z which represents the whole input sequence.

$$h_i = [LSTM(x,i); LSTM(x,i)] \qquad (1)$$

$$\alpha_i = softmax(w^T h_i) \qquad (2)$$

Here, α_i represents the attention weights.

$$z = \sum_{i=1}^{n} \alpha_i h_i \qquad (3)$$

Now, we pass the attention representation vector z to m MLPs related to the m tasks with $Task_1$ as the main task and $Task_i$ as the m-1 scaffold tasks, where i ∈ [2, m], to get an output vector y = { y_1, y_2, y_3,.......y_m } :

$$y_i = softmax(MLP_i(z)) \qquad (4)$$

For each task, we use a Multi Layer Perceptron (MLP) followed by a softmax layer to obtain the class with the highest class probability. The parameters of a task's MLP are the specific parameters of that task and the parameters in the lower layers (parameters till the attention layer) are the shared parameters.

Training

We train our model only on the SciCite dataset. Then we fine-tune on the target dataset (ACL-ARC or 3C Challenge). We use the pre-trained scibert scivocab uncased model trained on a corpus of 1.14M papers and 3.1B tokens to get the 768-dimensional Word Embeddings. While training on the SciCite dataset, we only train the two structural scaffolds which are - 1. Citation Worthiness scaffold, 2. Section Title scaffold, along with the main task. While fine-tuning on the target datasets, we use the Cited paper title scaffold only, while freezing the task specific parameters of the other two scaffolds, learned during the training on the SciCite dataset. We compute the loss function as:

$$L = \sum_{(x,y) \in D_1} L_1(x,y) + \sum_{i=2}^{n} \lambda_i \sum_{(x,y)\, \epsilon\, D_i} L_i(x,y) \qquad (5)$$

where D_i is the labeled dataset corresponding to $task_i$, λ_i is the hyperparameter that specifies the sensitivity of the model to each specific task, L_i is the loss corresponding to $task_i$. In each training epoch, we take a batch with equal number of instances from all the tasks and calculate the loss as specified in Equation 5, where $L_i = 0$ for all the instances of other tasks, $task_k$ where $k \neq i$. Then, we perform back propagation and update the parameters using the AdaDelta optimizer with gradient clipping.

Experiments

Baselines

We have worked on multiple baseline models to compare their performance on the ACL-ARC and the 3C Challenge datasets.

Table 5. Baselines and Proposed System.

Baselines	*Description*
BiLSTM+Attention (with SciBERT)	This baseline has a similar structure as our proposed model until the attention layer. It only has one MLP related to the main task and optimizes the network for the main loss.
3C Shared Task Submission 1	This system submission has achieved the best submission results on the 3C Challenge dataset in the Kaggle 3C Shared Task. The model is a Passive Aggressive Classifier with a concatenated vector (including the citing paper title, cited paper title and the citation context) as the input.
Cohan et al. (2019)	The model has reported state-of-the-art results on the ACL-ARC dataset. It incorporates a multi task learning framework with two structural scaffolds predicting the section title and citation worthiness, given the citation context.
Representation Model	The model framework for this baseline incorporates the concatenation of two representation vectors which is passed on to a MLP for classification. We get the first representation from the attention layer of the pretrained first baseline by passing citation context and the cited title as input. We use the pre-trained Cohan et al. (2019) model, trained on SciCite to get the predicted labels on the target dataset. Then, we infuse this external knowledge with the citation context and pass it to the first baseline to obtain the second attention layer representation.
Late Fusion Model	This baseline model has a similar structure to that of the first baseline. We use the pre-trained Cohan et al. (2019) model, trained on SciCite to get the citation intent, section title and the citation worthiness labels. We concatenate these labels with the output of the attention layer of this baseline and pass it to a MLP for prediction.

Table 6. Results on the ACL-ARC and the 3C Challenge datasets. The first two columns (Macro F1 score and Accuracy) correspond to the results on the ACL-ARC dataset. The last two columns (Public and Private F1 scores) are the results on the 3C Challenge dataset.

Category	*ACL-ARC*		*3C Challenge dataset*	
	Macro F1 Score	*Accuracy*	*Public F1*	*Private F1*
BiLSTM + Attention (with SciBERT)	57.1	63.3	27.8	23.9
Cohan et al. (2019)	67.9	76.2	22.4	**25.2**
Kaggle 3C Shared Task Submission 1	-	-	21.5	20.6
Our Model	**73.2**	**77.0**	**29.5**	24.2
Representation Model	38.2	54.7	20.6	23.1
Late Fusion Model	48.3	61.9	22.4	22.4

Results

Our results for the ACL-ARC and the 3C challenge datasets are shown in Table 6. We observe that as compared to the first baseline "BiLSTM+Attention (with SciBERT)", Cohan et al. (2019) achieves an F1 macro score of 67.9 (Δ = 10.8) and a validation accuracy of 76.2 (Δ =

12.9) on the ACL-ARC dataset indicating the fact that leveraging the structural information of a research work helps the model to learn more effectively. Out of all the other baselines, our model achieves the best results with an F1 score of 73.2, a significant improvement over the previous state of the art results of Cohan et al. (2019) (Δ = 5.3) and a validation accuracy of 77.0 (Δ = 0.8) on the ACL-ARC dataset. This clearly demonstrates the efficacy of using the three scaffolds in a transfer learning framework for this task. For the last two baselines that are mainly based on fusing external knowledge obtained by using the pre trained Cohan et al. (2019) model, we find a significant dip in the performance. This suggests that this external knowledge does not provide any useful signals beyond what the first baseline already learns from the data. The results on the 3C Challenge dataset also show similar patterns.

Analysis

To gain more insight into how the scaffolds are helping the model, we consider examples from the ACL-ARC and the 3C Challenge datasets and compare the predictions of the simple baseline *'BiLSTM+Attention (with SciBERT)'*, the previous state of the art *'Cohan et al. (2019)'*, and our best-proposed model *'BiLSTM+Attention (with SciBERT)+three scaffolds'*. In table 7, the first example is from the ACL-ARC dataset with true label *COMPARE*, the simple baseline and the Cohan et al. (2019) incorrectly predict it as *MOTIVATION* and *BACKGROUND* respectively, whereas our model predicts it correctly. The simple baseline does not include any scaffold, while the Cohan et al. (2019) and our model incorporate a multi task learning framework. Note that our model is similar to that of Cohan et al. (2019) but includes three scaffolds (two structural scaffolds + cited paper title scaffold). The second example is from the 3C Challenge dataset where the true label is *BACKGROUND*, the Cohan et al. (2019) model is probably distracted by the phrase "use", so it classifies it incorrectly as *USE*, whereas our model correctly classifies it. Note that our model also consists of additional information from the cited paper (title) which provides additional context, thus helping it to classify better.

Table 7. A sample of predictions of the models on examples from the ACL-ARC and the 3C Challenge datasets.

Example	*Model*	*Predicted Label*	*True Label*
The advantage of tuning similarity to the application of interest has been shown previously by CITATION.	BiLSTM + Attention (with SciBERT)	MOTIVATION	COMPARE
	Cohan et al.	BACKGROUND	COMPARE
	Our Model	COMPARE	COMPARE
Others use concepts such as expansion and contraction (Mattsson, 1987); extension and consolidation CITATION and splitting and joining (Hertz, 1996)	Cohan et al.	USE	BACKGROUND
	Our Model	BACKGROUND	BACKGROUND
We experiment with four learners commonly employed in language learning: Decision List (DL): We use the DL learner as described in CITATION, motivated by its success in the related tasks of word sense disambiguation (Yarowsky , 1995) and NE classification (Collins and Singer , 1999) .	Our Model	USE	MOTIVATION

Error Analysis

We investigate the type of errors made by our proposed model on the two datasets. We found it surprising to note that in case of the ACL-ARC dataset, the model has more tendency to make false positive errors in the *COMPARE* category, although it being the second most dominating category. Whereas in the case of the 3C Challenge dataset, it makes many false positive errors in the *BACKGROUND* category. To overcome this problem of overfitting, we decided to use some oversampling techniques like SMOTE, but we still did not get any significant improvements. Figure 2 shows the confusion matrix of our best model on the two datasets. We also found out that some errors are due to ambiguity in the citation context as well as the title of the cited paper. We can avoid them by providing some additional context apart from the cited paper title information (for example, providing abstract from the cited paper, etc). In the last example from Table 7, the model is probably distracted by the phrases "We use" and "as described in CITATION", leading to an inference that there is a usage of a method from the cited paper, instead of considering the latter part of the sentence that describes the motivation. This is likely due to the small number of training instances in the *MOTIVATION* category, preventing the model from learning such subtle details.

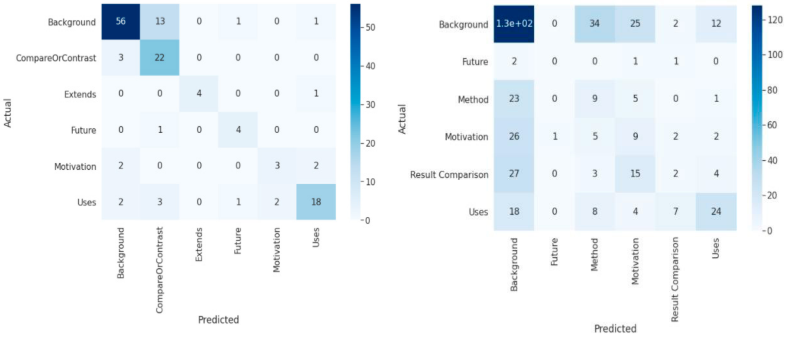

Figure 2. Confusion matrix showing the classification errors of our best model on the ACL-ARC (test size: 139) and the 3C Challenge datasets (test size: 400) respectively.

Conclusion and Future Work

In this work, we demonstrate that the structural information related to a research paper and additional context (title information) of the cited paper can be leveraged to effectively classify the intent of the citations. A future line of research would be to use the abstract of the cited paper as further contextual information for the task and also to investigate alternative approaches for solving the issue of overfitting on the 3C Challenge dataset.

References

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *EMNLP*.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28, 41–75.

Cohan, A., Ammar, W., Zuylen, M. & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *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) (*NAACL*) (2019), 3586–3596.

Jurgens, D., Kumar, S., Hoover R., McFarland D. & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *TACL*, 6, 391-406.

Moravcsik, M. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science*, 5(1), 86–92.

Pride, D. & Knoth, P. (2020). An Authoritative Approach to Citation Classification. In: *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*, 1-5 Aug 2020, Virtual China.

Stevens, M. & Giuliano, V. (1965). *Statistical Association Methods for Mechanized Documentation: Symposium Proceedings, Washington, 1964*, volume 269. US Government Printing Office.