Scientometrics (2024) 129:767–783
https://doi.org/10.1007/s11192-023-04811-5

A Deep Multi-Tasking Approach Leveraging on Cited-Citing Paper Relationship For Citation Intent Classification

Tirthankar Ghosal[1,4] · **Kamal Kaushik Varanasi**[2] · **Valia Kordoni**[3]

Received: 9 November 2021 / Accepted: 19 October 2022 / Published online: 13 December 2023

Abstract
Citations are crucial artifacts to provide additional information to the reader to comprehend the research under concern. There are different roles that citations play in scientific discourse. Correctly identifying the intent of the citations finds applications ranging from predicting scholarly impact, finding idea propagation, to text summarization. With the rapid growth in scientific literature, the need for automated methods to classify citations is now growing intense. However, we can only fully understand the intent of a citation if we look at the citation context in the citing paper and also the primary purpose of the cited article. In this work, we propose a neural multi-task learning framework that harnesses the structural information of the research papers and the cited paper's information for the effective classification of citation intents. We analyze the impact of three auxiliary tasks on the performance of our approach for citation classification. Our experiments on three benchmark citation classification datasets show that incorporating cited paper information (title) shows that our deep 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. We also achieve comparable performance with respect to the best-performing systems in the SDP 2021 3C Shared task on Citation Context Classification. We make our codes available at https://github.com/Tirthankar-Ghosal/citationclassification-SCIM

Keywords Citation classification · Citation context · Deep learning

Introduction

Citations are crucial in a research paper and the scholarly community for various reasons, including scientific and administrative. Over the years, citation analysis techniques are used to track research in a field, discover evolving research topics (Morris et al., 2003; Upham & Small, 2010; Small et al., 2014, 2017; Chaker et al., 2021), and measure the impact of research articles, venues, researchers, etc (Li & Ho, 2008; Zhang &

Tirthankar Ghosal and Kamal Kaushik Varanasi have contributed equally to this work.

The first author performed this work while he was in Institute of Formal And Applied Linguistics, Charles University, Malostranské náměstí 25, 11800 Prague, Czech Republic.

Extended author information available on the last page of the article

Wu, 2021; Waltman, 2016; Hernandez-Alvarez et al., 2017). Citations help analyze the link between different research articles, identify research gaps, and stem up new ideas. Authors use citations to frame their contributions and connect to an intelligent lineage (Latour, 1987). Authors cite other works for a number of reasons including demonstrating knowledge of the field, establishing the placement of the citing work in the field, comparing and criticizing other works, gaining support for their claims, and attributing contributions of seminal work by pioneers in the field (Hernandez-Alvarez et al., 2017). The automatic recognition of the rhetorical function of citations in scientific text has many applications, from improvement of impact factor calculations to text summarization and more informative citation indexers (Teufel et al., 2006). With the growing intrusion of Artificial Intelligence techniques in scholarly document processing (Chandrasekaran et al., 2020), automated analysis of the scientific discourse leveraging on the intent of citations is an exciting direction to investigate. However, not all citations are created equal, nor do they play similar roles. Citations have different intents depending upon the citation context, the section under which they occur, etc. For example, they might indicate the usage of a method or getting motivation from previous work, or authors might use them to compare the methodology of different works. Most of the works for this task are feature based, where the authors use a set of predefined hand-engineered features. But recently, the authors in Cohan et al. (2019) have stressed more on the importance of the structural signals available in the data that are based on the structural properties of a scientific work for this task. We adopted the idea, and formulated a novel approach for the task. We opine that *researchers make the purpose of the cited paper explicit when they cite it. The purpose of a paper is usually manifested in the paper title or the abstract. Hence, using the citation context in the citing paper and the purpose of the cited paper (title or abstract) seems an interesting direction to probe for understanding the intent of the citation*. Example in Table 1 demonstrates the evidence that the citation seems less ambiguous and easier to classify after accounting for the cited paper title information in addition to the citation context. In this work, we show that by utilizing the cited paper information in addition to the information from structural signals, we can learn better representations for solving the task in hand.

In this work, our contribution towards citation intent classification via leveraging cited-citing paper relationship are:

- We propose a deep multi-task learning (MTL) framework with three auxiliary tasks (scaffolds) and representations learned from a contextualized language model trained on scientific articles (SciBERT (Beltagy et al., 2019)). We introduce a new auxiliary task, the cited paper title scaffold, that leverages the relationship between the citation context and the cited paper title.
- We exhibit an increase in performance with an absolute point of 5.3% F1 from the previous state-of-the-art (Cohan et al., 2019). The proposed approach achieves 73.2% F1 score on the Anthology Reference Corpus (ACL-ARC) citations benchmark.

Essentially, we use Natural Language Processing and Machine Learning to include information from the citation context of the citing paper and the purpose of the cited paper (essentially cited paper's title) for classifying citation purposes. Our current work draws motivation from structural scaffolds in Cohan et al. (2019) and builds upon our earlier work (Varanasi et al., 2021) published as a short paper in ISSI 2021.

The paper is organized as follows. In "Introduction" section, we discuss the related work. In "Dataset description" section, we describe the datasets that we use for our

Table 1 Example to show how cited paper title aids in understanding the citation intent

Citation context	Cited paper title	True Label
She evaluates 3000 German verbs with a token frequency between 10 and 2000 against the Duden (@@CITATION)	Duden–das stilworterbuch duden–the style dictionary	BACKGROUND

experiments. In "Proposed approach" section, we discuss about our proposed approach for this task. In "Experiments" section, we discuss about the experimental details and the baselines. In "Results" section, we discuss about the results of our experiments and analyze them. Finally, "Conclusion and future work" section contains our conclusions and our future plans.

Related works

Research on different schemes for citation classification is popular with the scientometrics community. Most of these studies provide fine-grained citation categories as in Garfield et al. (1965), Moravcsik and Murugesan (1975), Teufel et al. (2006), so they are rarely used for automated analysis of the scientific publications. To overcome these problems, Jurgens et al. (2018) proposed a six-category classification scheme. Then, in 2019, Cohan et al. (2019) used a different schema with only three categories to devise more computationally efficient methods. More recently, Pride and Knoth (2020) proposed the academic citation typing (ACT) dataset that follows a classification scheme similar to Jurgens et al. (2018) with the only difference being the addition of an extra layer to the compare/contrast category. The addition of this sub-class is to show similarities, differences or disagreement.

One of the early contributions for automated classification of citation intents was from Garzone and Mercer (2000), a rule-based system where the authors used a classification scheme with 35 categories. Later on, works included using machine learning systems based on the linguistic patterns of the scientific works. For example, the use of “cue phrases” along with fine-grained location features such as the location of citation within the paragraph and the section in Teufel et al. (2006). Jurgens et al. (2018) engineered pattern-based features, topic-based features, and prototypical argument features for the task. Recently, Cohan et al. (2019) proposed 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. To this end, we propose a deep MTL framework with three scaffolds. We explain more about our model architecture in the following sections.

Dataset description

We use three benchmark datasets from the NLP community for the task. Table 2 shows the data statistics related to various datasets.

SciCite

Cohan et al. (2019) introduced a citation intents dataset that provide a concise classification scheme with three intent categories: BACKGROUND, METHOD and RESULT_COMPARISON. The authors propose this classification scheme by merging multiple categories listed in Jurgens et al. (2018) into the BACKGROUND category. They argued that their scheme is general and naturally fits in scientific discourse in multiple domains, unlike the other ones that are domain specific.

Please note that the SciCite dataset includes the data corresponding to the structural scaffolds: Section Title Prediction (91412 instances) with five labels—*Introduction, Conclusion, Experiments, Method, and Related Work*, and the Citation Worthiness Prediction (73484 instances) with two labels—*True, False*.

ACL-ARC

In 2018, Jurgens et al. (2018) introduced the ACL-ARC citation function dataset for citation classification based on a six category classification scheme. The classification categories are described in Table 3. Note that as mentioned earlier, unlike Pride et al. (2020), the Jurgens et al. (2018) classification scheme does not include an extra layer with sub-classes in the compare/contrast category. Kindly refer to Table 2 in Jurgens et al. (2018) for the citation class distribution. We see that labels MOTIVATION (98 instances), CONTINUATION (73), and FUTURE (68) are relatively scarce in comparison to BACKGROUND (1021), USES (365), and COMPARES OR CONTRASTS (344).

3C challenge dataset

The 3C Shared Task as part of the Scholarly Document Processing workshop 2021 (Beltagy et al., 2021) hosted a community challenge for Citation Context Classification (3C). The competition used a part of the ACT dataset (Pride & Knoth, 2020) that we refer to here as the 3C Challenge dataset. The 3C challenge was motivated towards multiclass classification of citation contexts based on purpose with categories—BACKGROUND, USES, COMPARES & CONTRASTS, MOTIVATION, EXTENSION, and FUTURE. The dataset consists of 3000 training instances and 1000 testing instances. The test data is not publicly available, so we mention the results we get after submitting our test data predictions on the Kaggle competition platform[1].

[1] https://www.kaggle.com/c/3c-shared-task-purpose-v2.

Table 2 Citation classification dataset details used in this study

Dataset	Papers	Annotated by	Citations	Intents	Discipline(s)
SciCite	6627	Volunteers	11,020	3	Comp. Sci/Medicine
3C Challenge	883	Paper authors	3000	6	Multi-disciplinary
ACL-ARC	185	Domain Experts	1989	6	Comp. Science

Table 3 Citation classification scheme followed in the ACL-ARC and the ACT (3C Challenge) datasets

Category	Description
Background	The cited paper provides relevant Background information or is part of the body of literature
Uses	The citing paper uses the methodology or tools created by the cited paper
Compare contrast Similarities Differences Disagreement	The citing paper express similarities or differences too, or disagrees with, the cited paper
Motivation	The citing paper is directly motivated by the cited paper
Extension	The citing paper extends the methods, tools or data etc. of the cited paper
Future	The cited paper may be a potential avenue for future work

Proposed approach

We propose a Multitask learning framework (Caruana, 1997) with the main task of citation intent classification along with three auxiliary tasks. These tasks help the model to learn optimal parameters for better performance on the main task. We retain the two structural scaffolds as proposed by Cohan et al. (2019). These auxiliary tasks are related to the structural properties of the scientific papers. They help the model to incorporate the structural information available in scientific documents into the citation intents. The scaffolds that we use are explained below. Note that the first two scaffolds are the structural scaffolds.

Section title scaffold task

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 scaffold task

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.

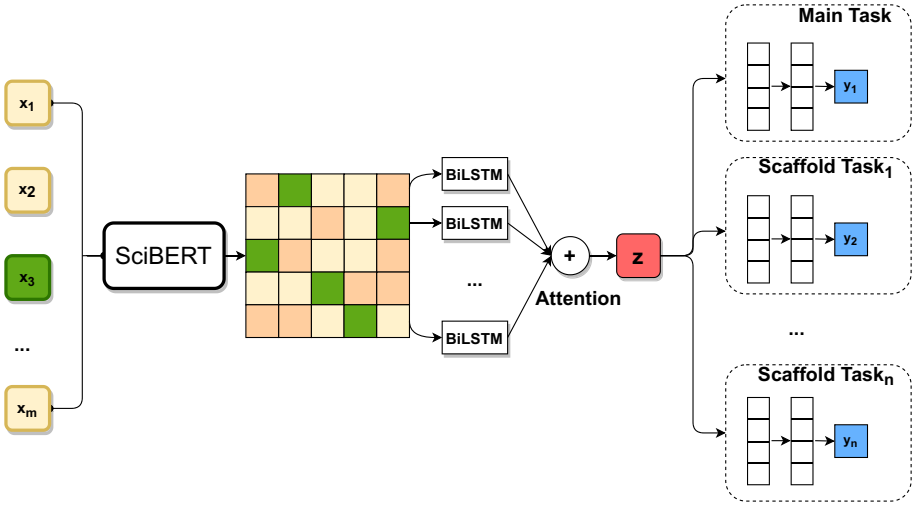

Fig. 1 Our proposed model structure for citation classification. Our main task MLP is for prediction of citation intents (top right) followed by MLPs for the auxiliary tasks

Cited paper title scaffold

Sometimes a citation context might not be enough to correctly predict the intent of the citation. 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 intent behind the 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. The target labels are the same as the main task labels.

Model architecture

In this section, we explain the architecture of our MTL framework. We use these auxiliary tasks only while training/fine-tuning the model for the main task. The overview of the model is shown in Fig. 1.

Let C be the tokenized citation context of size m. We pass it onto the SciBERT (Beltagy, 2019). Beltagy et al. (2019) model with pre-trained weights to get the word embeddings of size (m, d_1) i.e. we have the output as $v = \{v_1, v_2, v_3, ...v_m\}$ where $v_i \in R^{d_1}$. Then we use a Bidirectional long short-term memory (Hochreiter & Schmidhuber, 1997) (BiLSTM) network with a hidden size d_2 to get an output vector h of size $(m, 2d_2)$.

$$h_i = [\text{LSTM}(x, i);\text{LSTM}(x, i)] \tag{1}$$

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,

$$\alpha_i = \text{softmax}(w^T h_i / d_2) \tag{2}$$

Here, α_i represents the attention weights.

$$z = \sum_{i=1}^{m} \alpha_i h_i \tag{3}$$

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.

We pass the vector z to n MLPs related to the n tasks with task_1 as the main task and task_i as the n-1 scaffold tasks, where $i \in [2, n]$, to get an output vector $y = \{y_1, y_2, y_3, ...y_n\}$.

$$y_i = \text{softmax}(\text{MLP}_i(z)) \tag{4}$$

Training

In this section, we describe the training in two stages. Note that we use the Citation intenent classification dataset (SciCite dataset) only for improving our performance on the target datasets. In our experiments, we use the ACL-ARC and the 3C Challenge datasets as the target datasets.

- *Training on the SciCite dataset* We only use the two structural scaffolds which are (1) Citation Worthiness scaffold, (2) Section Title scaffold, while turning off the Cited Paper Title scaffold (i.e. we freeze the parameters related to the MLP of this task).
- *Fine-tuning on the target datasets* We use the Cited paper title scaffold only while turning off the other two scaffolds (freezing the task specific parameters of the other two scaffolds).

We compute the loss function as :

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

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 formulate a batch with an equal number of instances from all the tasks and calculate the loss as specified in Eq. (5), where $L_i = 0$ for all the instances of other tasks, task_k where $k \neq i$. Then, we perform backpropagation and update the parameters using the AdaDelta optimizer.

Experiments

Hyperparameter details

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. Then, we use a single layer BiLSTM with a hidden size of 50 for each direction. For each task, we use an

MLP layer with 20 hidden nodes, a dropout layer between the input and the hidden layer with a dropout rate = 0.2 (Srivastava et al., 2014) in case of training on SciCite, while a dropout rate = 0.3 in case of fine tuning, and a RELU (Nair & Hinton, 2010) activation layer. For training on SciCite, we use hyperparameters λ_i as: λ_1 (section title scaffold) = 0.05, λ_2 (citation worthiness scaffold) = 0.1, λ_3 (cited paper title scaffold) = 0. For fine-tuning on the target datasets, we use: λ_1 (section title scaffold) = 0, λ_2 (citation worthiness scaffold) = 0, λ_3 (cited paper title scaffold) = 0.1. We determine the λ_i and the other hyper-parameters on the basis of performance of the model on the validation data. We use a batch size of 12 for SciCite and 8 for the target (3C Challenge/ACL-ARC) datasets. We also use SMOTE (Chawla et al., 2002) oversampling while fine tuning on the target datasets.

Baselines and comparing systems

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

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 only.

3C shared task best submission

In the 3C Shared Task 2021, the winning team tested various machine learning and deep learning models and found out that BERT based models like SciBERT outperformed Random Forest. The best result was obtained for uncased SciBERT with a linear classification layer. We also experiment with our current model with macro F1 as dictated by the challenge in Kaggle[2] and achieved a score of 26.973 in the competition.

Cohan model

This model has reported state-of-the-art results on the ACL-ARC dataset. It incorporates a MTL 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 are passed on to an MLP for classification. We get the first representation from the attention layer of the pre-trained BiLSTM+Attention (with SciBERT) baseline. The input sequence is obtained by combining the citation context and title of the cited paper, separated by the [SEP][3] token. We use the pre-trained Cohan model trained on SciCite to get the three-class predicted labels on the target dataset. Then, we combine these

[2] https://www.kaggle.com/c/3c-shared-task-purpose-v2/overview/evaluation.

[3] For BERT based models, the separator token [SEP] is used when building a sequence from multiple sequences.

Table 4 Results on the ACL-ARC dataset

Model	Macro F1	Accuracy
BiLSTM + Attention (with SciBERT)	57.1	63.3
BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)	70.1	76.3
BiLSTM-Attn + Section Title scaff. + Cited Paper Title scaff. (with SciBERT)	62.3	73.4
BiLSTM-Attn + Cit. Worthiness scaff. + Cited Paper Title scaff. (with SciBERT)	67.5	74.8
BiLSTM + Attention (with SciBERT) + three scaffolds	**73.2**	**77.0**
Cohan et al. (2019)	67.9	76.2
Representation Model	38.2	54.7
Late Fusion Model	48.3	61.9

The rows in Bold signifies the best-performing model

predictions with the citation context and pass it to the BiLSTM+Attention (with SciBERT) baseline to obtain the second attention layer representation.

Late fusion model

This baseline model has a similar structure to that of the BiLSTM+Attention (with SciBERT) baseline. We use the pre-trained Cohan model, trained on SciCite, to get the citation intent, section title, and the citation worthiness predicted labels. We concatenate these labels with the output of the attention layer of this baseline and pass it to an MLP for prediction.

Results

We show the results on the ACL-ARC in Table 4. The ACL-ARC citation function dataset (Jurgens et al., 2018) originally has 1969 citation instances and a total of 3083 instances when combined with (Teufel et al., 2006). For the 3C Challenge dataset, we show the submission results on the Kaggle platform due to the unavailability of the test data labels to the participants. Hence, we mention the Public and Private F1 scores. According to the Kaggle competition rules[4], the Public and Private F1 are the macro-averaged F1 scores on the initial 50% of test data and the rest 50% of test data respectively (please note that there are 1000 public test instances and 1000 private test instances in the 3C dataset). The Private F1 scores are used for the final ranking and are released at the end of the competition. Our results for the 3C Challenge dataset are shown in Table 5.

We perform an ablation study for both the datasets to understand the impact of each scaffold on performance of the model on the main task. From our experiments, it is evident that each scaffold helps the model to learn the main task more effectively, hence helping it to perform better than the simple baseline that does not include any scaffolds.

In case of ACL-ARC, it is important to note that the "BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)" model is similar to the state-of-the-art (Cohan et al., 2019) model, the significant difference being the usage of SciBERT

[4] https://www.kaggle.com/c/3c-shared-task-purpose-v2/overview/evaluation.

Table 5 Results on the 3C challenge dataset

Model	Public F1	Private F1
BiLSTM + Attention (with SciBERT)	20.9	20.8
BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)	25	22.3
BiLSTM-Attn + Section Title scaff. + Cited Paper Title scaff. (with SciBERT)	23	25.6
BiLSTM-Attn + Cit. Worthiness scaff. + Cited Paper Title scaff. (with SciBERT)	28	**26.5**
BiLSTM + Attention (with SciBERT) + three scaffolds	**30.3**	26.1
3C Shared Task Best Submission[a]	33.9	27
Representation model	20.6	23.1
Late fusion model	22.4	22.4

[a]https://aclanthology.org/2021.sdp-1.17/.

The rows in Bold signifies the best-performing model

embeddings instead of Embeddings from Language Model (ELMO) (Peters et al., 2018) and Global Vectors for word representation (GLOVE) (Pennington et al., 2014). We can observe that usage of SciBERT has improved the performance upto some extent (Macro F1 score of 70.1 ($\delta = 2.2$) and a validation accuracy of 76.3 ($\delta = 0.1$)) but the addition of the Cited Paper Title scaffold helps the model to perform even better. We observe that our best model including all the three scaffolds is able to significantly surpass the previous state-of-the-art Cohan et al. (Macro F1 score of 73.2 ($\delta = 5.3$) and a validation accuracy of 77 ($\delta = 0.8$)). This suggests that along with the structural scaffolds, the Cited Paper Title scaffold helps the model to learn the main task more effectively. For the last two baselines, which are mainly based on using the external knowledge obtained by using the pre-trained Cohan model, we find a significant dip in the performance. This suggests that this external knowledge does not provide any useful signals beyond what the simple baseline already learns from the data.

For the 3C Challenge dataset, we observe a comparable performance with respect to the best performing system in the competition. We observe that out of all the baselines we use in our ablation studies, our best model including all the three scaffolds achieves the best Public F1 score, although it is marginally lagging behind behind the "BiLSTM-Attn + Cit. Worthiness scaff. + Cited Paper Title scaff. (with SciBERT)" model in case of the Private F1 scores. We also observe that the last two baselines perform slightly better than the BiLSTM + Attention (with SciBERT) baseline. Both the baselines perform bad as compared to the BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT) baseline on the Public test data but achieve slightly better results on the Private test data. This behavior is different as compared to the performance on the ACL-ARC dataset, which may be due to the multi-domain 3C data.

Based on our ablation studies, we can understand the importance of each scaffold s_i by calculating the difference in F1 scores (δ) between our best model (including all the three scaffolds) and the baseline including the scaffolds other than s_i. We observe that in the case of ACL-ARC, the scaffold significance order is : Citation worthiness ($\delta = 10.9$) > Section Title ($\delta = 5.7$) > Cited Paper Title ($\delta = 3.1$). But in the case of the 3C Challenge dataset, we observe that the order changes, which may be due to the fact that 3C includes data from multiple domains. Therefore, it may be difficult to generalize. On the public leaderboard, the significance order is: Citation worthiness ($\delta = 7.3$) > Cited Paper Title ($\delta = 5.3$) > Section Title ($\delta = 2.3$), while on the Private leaderboard, the order becomes : Cited Paper Title

(δ = 3.8) > Citation worthiness (δ = 0.5) > Section Title (δ = -0.4). This indicates that the Section Title scaffold is not helping the model to perform better on the 3C Challenge dataset, in fact it slightly has a negative impact on the performance on the Private test data.

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), 'BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)' baseline and our best-proposed model 'BiLSTM+Attention (with SciBERT)+three scaffolds.' Table 6 shows the predictions of different models on the examples from the two datasets.

In Table 6, the first two examples show the difference in predictions of the simple baseline, Cohan et al. (2019), Cohan et al. (2019) and our best performing model. In the first and second examples, the true labels are FUTURE and COMPARE respectively, our model classifies them correctly unlike the simple baseline, and Cohan et al. (2019). Note that our model includes the cited paper title scaffold and the SciBERT word representations, unlike the simple baseline and the Cohan et al. model, both of which lack either one or both of them. The word embeddings from SciBERT help the model to get better vector representations of the input sequence while the scaffold provides the model with additional context from the cited paper for better classification. We also compare between the simple baseline, 'BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)' baseline and our best model by referring to the last two examples in Table 6. In the third example, the true label is FUTURE. The simple baseline incorrectly predicts it as COMPARE, whereas the 'BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)' baseline and our model predict it correctly. This might be due to the lack of structural scaffolds in the simple baseline, unlike the other two. The true label is BACKGROUND for the fourth example. Both the simple and the 'BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)' baselines incorrectly predict it as USE, whereas our model correctly predicts it. This might be because the other two models got distracted by the phrase "use",hence classifying it in the USE category. Note that our model consists of additional information from the cited paper title compared to the other two models, which provides further context, hence helping the model to classify better.

We investigate the type of errors made by our proposed model on the two datasets. We found it surprising to note that on the ACL-ARC dataset, the model has more tendency to produce false-positive errors in the COMPARE category, although it being the second most dominating category (in terms of the number of instances in the dataset). Whereas for the 3C Challenge dataset, our model makes many false-positive errors in the BACKGROUND, METHOD, MOTIVATION and USES categories.

Figures 2 and 3 show the confusion matrices of our proposed model on the ACL-ARC and the 3C Challenge datasets respectively. Some errors in the ACL-ARC dataset are due to the model falsely classifying the instances of the BACKGROUND category as the COMPARE category.

We found out that some errors could be prevented by providing some additional context apart from the cited paper title information (for example, providing contextual information around the citation text, abstract from the cited paper, etc.). Such errors are shown in Table 7. For the first example in this table, the model is probably distracted by the phrases

Table 6 A sample of predictions of the models on examples from the ACL-ARC and the 3C Challenge datasets

Example	True	Model	Predicted
One possible direction is to consider linguistically motivated approaches , such as the extraction of syntactic phrase tables as proposed by Yamada and Knight (2001)	FUTURE	BiLSTM + Attention(with SciBERT)	BACKGROUND
		Cohan et al.	BACKGROUND
		Our model	FUTURE
The advantage of tuning similarity to the application of interest has been shown previously by Weeds and Weir (2005)	COMPARE	BiLSTM + Attention(with SciBERT)	MOTIVATION
		Cohan et al.	BACKGROUND
		Our model	COMPARE
A possible future direction would be to compare the query string to retrieved results using a method similar to that of Tsuruoka and Tsujii (2003)	FUTURE	BiLSTM + Attention(with SciBERT)	COMPARE
		BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)	FUTURE
		Our model	FUTURE
Others use concepts such as expansion and contraction (Mattsson, 1987); extension and consolidation (#AUTHORTAG, 1982) and splitting and joining (Hertz, 1996)	BACKGROUND	BiLSTM + Attention(with SciBERT)	USE
		BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)	USE
		Our model	BACKGROUND

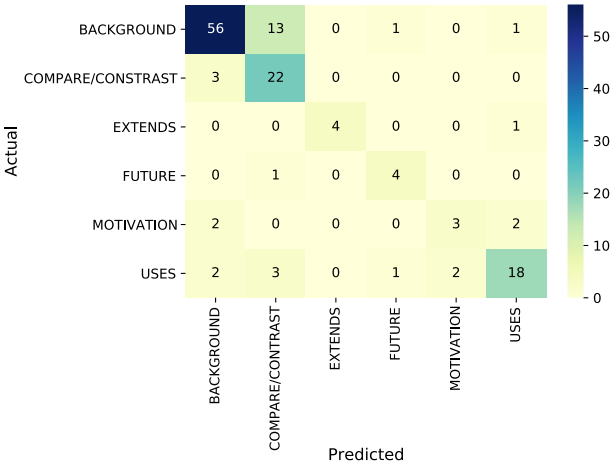

Fig. 2 Confusion matrix showing the classification errors of our best model on the ACL-ARC test data. (we create a held-out test set of 139 instances)

"We use" and "as described in Collins and Singer (1999)", leading to an inference that there is a usage of some 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 (∼5%), preventing the model from learning such subtle details. For the second and third examples, the cited paper title information is insufficient, so the model needs additional context for better classification. Similarly, in the last example, the text seems ambiguous without accessing some additional context apart from the cited paper title.

Conclusion and future work

In this work, we demonstrate, structural information related to a research paper with additional context (title information) of the cited article can be leveraged to classify the citation's intent effectively. We propose a novel deep MTL framework with three auxiliary

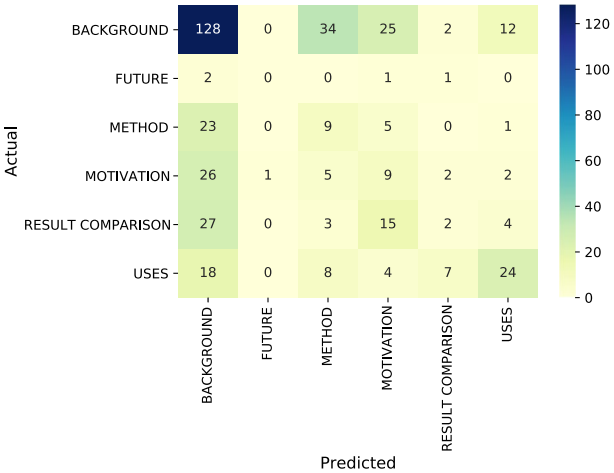

Fig. 3 Confusion matrix showing the classification errors of our best model on the 3C Challenge test data. (we leave out 400 instances from the training data for this prediction)

Table 7 A sample of model's classification errors on the ACL-ARC dataset

Example	Cited paper title	True label	Prediction
We experiment with four learners commonly employed in language learning: Decision List(DL): We use the DL learner as described in Collins and Singer (1999), motivated by its success in the related tasks of word sense disambiguation(Yarowsky, 1995) and NE classification (Collins and Singer, 1999).	*Unsupervised Models for Named Entity Classification*	MOTIVATION	USE
ASARES is presented in detail in (Claveau et al., 2003)	*Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming.*	USE	BACKGROUND
Zollmann and Venugopal (2006) substituted the non-terminal X in a hierarchical phrase-based model by extended syntactic categories.	*Syntax Augmented Machine Translation via Chart Parsing*	COMPARE	BACKGROUND
Note that although our current system uses MeSH headings assigned by human indexers, manually assigned terms can be replaced with automatic processing if needed (Aronson et al. 2004).	*Long-term efficacy of BCG vaccine in American Indians and Alaska Natives: A 60-year follow-up study*	FUTURE	COMPARE

tasks (two of them related to the structure of the scientific work and the third one based on the relationship between citation context and the cited paper). The proposed approach exhibits an increase of 5.3% F1 (F1 score of 73.2%) over the previous state-of-the-art technique (Cohan et al., 2019) on the ACL-ARC Citation Function dataset (Jurgens et al., 2018).

A future line of research could be to use the abstract of the cited paper as further contextual information for the task and investigate alternative approaches to solve overfitting on the 3C Challenge dataset. Another relevant line of work could be to explore the design of other auxiliary tasks that are relevant to the main task.

Declarations

Conflicts of interest The authors declare no conflict of interest with any party with this submission.

References

Beltagy, I., Cohan, A., Feigenblat, G., Freitag, D., Ghosal, T., Hall, K., Herrmannova, D., Knoth, P., Lo, K., Mayr, P., Patton, R., Shmueli-Scheuer, M., de Waard, A., Wang, K., & Wang, L. (2021). Overview of the second workshop on scholarly document processing. In *Proceedings of the Second Workshop on Scholarly Document Processing*, (pp. 159–165). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.sdp-1.22

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, (pp. 3613–3618). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Caruana, R. (1997). Multitask learning. *Machine learning, 28*(1), 41–75.

Chaker, J., Herrera-Viedma, E., & Cobo, M. (2021). The use of citation context to detect the evolution of research topics: A large-scale analysis. *Scientometrics*. https://doi.org/10.1007/s11192-020-03858-y.

Chandrasekaran, M. K., Feigenblat, G., Freitag, D., Ghosal, T., Hovy, E. H., Mayr, P., Shmueli-Scheuer, M., & de Waard, A. (2020). Overview of the first workshop on scholarly document processing (SDP). In Chandrasekaran, M. K., de Waard, A., Feigenblat, G., Freitag, D., Ghosal, T., Hovy, E. H., Knoth, P., Konopnicki, D., Mayr, P., Patton, R. M., Shmueli-Scheuer, M. (Eds.) *Proceedings of the First Workshop on Scholarly Document Processing, SDP@EMNLP 2020, Online, November 19, 2020*, (pp. 1–6). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sdp-1.1

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16,* 321.

Cohan, A., Ammar, W., Van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. http://arxiv.org/abs/1904.01608

Garfield, E., et al. (1965). Can citation indexing be automated. In *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings* (vol. 269, pp. 189–192). Washington

Garzone, M., & Mercer, R. E. (2000). Towards an automated citation classifier. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 337–346). Springer

Hernandez-Alvarez, M., Soriano, J. M. G., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering, 23*(4), 561–588.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics, 6,* 391–406.
Latour, B. (1987). *Science in action : How to follow Scientists and Engineers throug society*. Harvard University Press.
Li, Z., & Ho, Y.-S. (2008). Use of citation per publication as an indicator to evaluate contingent valuation research. *Scientometrics, 75,* 97–110. https://doi.org/10.1007/s11192-007-1838-1.
Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86–92.
Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology, 54*(5), 413–422.
Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Icml*.
Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018* (pp. 2227–2237).
Pride, D., & Knoth, P. (2020). An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 337–340).
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467.
Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics, 11*(1), 46–62.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.
Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, (pp. 103–110).
Upham, S., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics, 83*(1), 15–38.
Varanasi, K. K., Ghosal, T., & Kordoni, V. (2021). Additional context helps! leveraging cited paper information to improve citation classification. In: Glänzel, W., Heeffer, S., Chi, P.-S., Rousseau, R. (Eds.) *Proceedings of the 18th International Conference on Scientometrics and Informetrics, ISSI 2021, Leuven, Belgium, July 12-15, 2021*, (pp. 1187–1192). ISSI Society.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365–391.
Zhang, F., & Wu, S. (2021). Measuring academic entities' impact by content-based citation analysis in a heterogeneous academic network. *Scientometrics, 126,* 7197–7222.

Authors and Affiliations

Tirthankar Ghosal[1,4] · Kamal Kaushik Varanasi[2] · Valia Kordoni[3]

✉ Tirthankar Ghosal
ghosal@ufal.mff.cuni.cz

Kamal Kaushik Varanasi
1801ce31@iitp.ac.in

Valia Kordoni
evangelia.kordoni@anglistik.hu-berlin.de

[1] Faculty of Mathematics and Physics, Institute of Formal And Applied Linguistics, Charles University, Malostranské náměstí 25, 11800 Prague, Czech Republic

[2] Department of Civil Engineering, Indian Institute of Technology Patna, Bihta, Patna, Bihar 801106, India

[3] Department of English and American Studies, Humboldt University Berlin, Unter den Linden 6,

10099 Berlin, Germany

[4] Present Address: National Centre for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, US