IITP@3C @ SDP @ NAACL 2021: Citation Classification (Task A) and Significance Detection (Task B)

Anonymous NAACL-HLT 2021 submission

Abstract

Besides providing additional contexts to research papers, citations act as trackers of the direction of research in a field and as an important measure in understanding the impact of a research publication. With the rapid growth in research publications, automated solutions for identifying the purpose and influence of citations are becoming very important. The 3C Citation Context Classification Task organized as part of the Second Workshop on Scholarly Document Processing @ NAACL 2021 is a shared task to address the aforementioned problems. In this work, we present our team, IITP@3C's submission to the shared tasks. For Task A, citation context purpose classification, 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. For Task B, citation context influence classification, we use a set of basic features to classify citations based on their perceived significance. We achieve comparable performance to the best performing system in Task A and superseded the majority baseline in Task B with very simple features.

1 Introduction

Citations are crucial in a research paper and the 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, and measure the impact of research articles, venues, researchers, etc. Correctly identifying the intent of the citations finds applications ranging from predicting scholarly impact, finding idea propagation, to text summarization, to establishing more informative citation indexers. Authors use citations to frame their contributions and connect to an intelligent lineage (Latour, 1987). However, not all citations are created equal, nor they play similar roles. Citations have different intents depending upon the citation context, the section under which they occur, etc. For example - a citation might indicate motivation or usage of a method from a previous work or a comparison of results of various works. And not even all citations are equally (Zhu et al., 2015) effective in finding relevant research. A majority of the papers cite a work contextually (Pride and Knoth, 2020) for providing additional background context. Such background contextual citations help in the broader understanding; however, they are not central to the citing paper's theme. While some papers use the methods or ideas from the previous works, build upon them, and further progress the research in the field. These papers are expected to acknowledge them by citing them duly. These citations, which heavily rely on a given work or build upon that work, are significant. In this paper, we describe our team, IITP@3C's entry for the 3C Citation Context Classification Task. This shared task consisted of two subtasks. The goal of Task A was to identify the purpose of the citations. The Task B intended to classify the citations based on their importance into either influential or incidental. For identifying the purpose of the citations (Task A), we employed a deep Multi-Task Learning (MTL) 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. We utilize the SciCite (Cohan et al., 2019) dataset apart from the training data available for this task. We achieve a Public F1 Macro Score of 30.258% and a Private Macro F1 score of 26.059%. For the task of assessing the importance of the citation (Task B), we pursue a feature-engineering approach to curate certain features from cited-citing paper pairs. In this task, we achieve a Public F1 Macro Score of 45.773% and

a Private Macro F1 score of 53.588%.

2 Related Works

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. In this work, we utilize the cited paper information as additional context and leverage the structural information related to the scientific discourse to improve the results.

Measuring academic influence has been a research topic since publications associate with academic prestige and incentives. Several metrics (Impact Factor, Eigen Factor, *h*-index, citation counts, alt metrics, etc.) came up to comprehend research impact efficiently. Still, each one is motivated on a different aspect and has found varied importance across disciplines. Zhu et al. (2015) did pioneering work on academic influence prediction leveraging on citation context. Valenzuela et al. (2015) explored citation classification into *influential* and *incidental* using machine learning techniques.

3 Task and Dataset Description

The 3C Citation Context Classification Shared Task organized as part of the Second Workshop on Scholarly Document Processing @ NAACL 2021 is a classification challenge, where each citation context is categorized based on its purpose and influence. It consists of 2 subtasks:

- **Task A:** Multiclass classification of citation contexts based on purpose with categories - BACKGROUND, USES, COMPARES CONTRASTS, MOTIVATION, EXTENSION, and FUTURE.
- **Task B:** Binary classification of citations into INCIDENTAL or INFLUENTIAL classes, i.e. a task for identifying the importance of a citation

The training and test datasets used for Task A and Task B are the same. The training data and test data consist of 3000 and 1000 instances, respectively. We also use the SciCite (Cohan et al., 2019) dataset for Task A, which includes 11,020 citations and provides a concise classification scheme with three intent categories: BACKGROUND, METHOD, and RESULT_COMPARISON. The dataset also contains data for the structural scaffolds.

4 Methodology

4.1 Task A

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.

- **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.
- **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.

Our model architecture is shown in Figure 1. Let C be the tokenized citation context of size

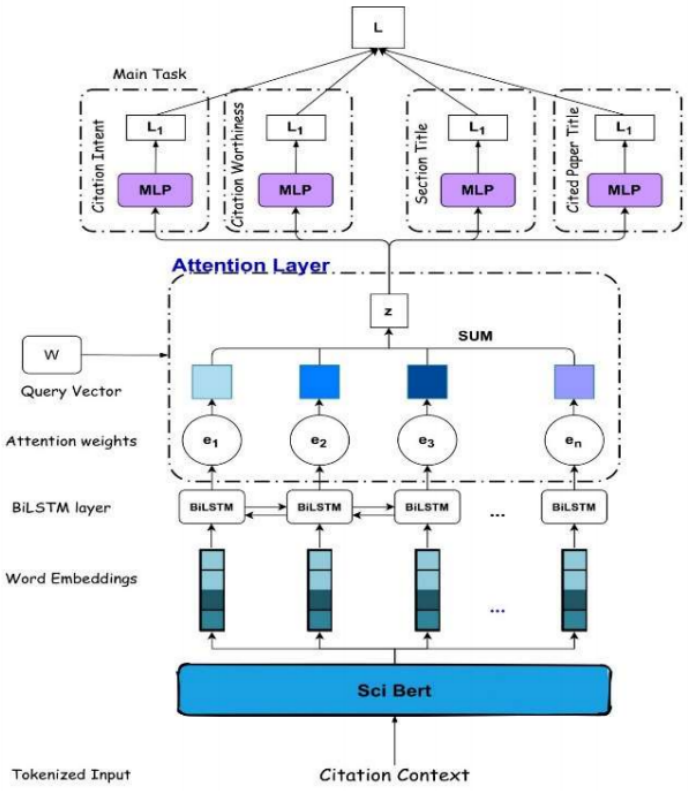

Figure 1: The architecture of our proposed model for Task A. The main task MLP is for prediction of citation intents (top left) followed by three MLPs for section title, citation worthiness, and cited paper title scaffolds.

n. We pass it onto the SciBERT (Beltagy et al., 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 $x_i \in R^{d_1}$. Then we use a Bidirectional long short-term memory (Hochreiter and Schmidhuber, 1997) (BiLSTM) network with a hidden size d_2 to get an output vector h of size $(n, 2d_2)$.

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

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

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

Here, α_i represents the attention weights.

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

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 m MLPs related to the m tasks with $Task_1$ as the main task and $Task_i$ as the m-1 scaffold tasks, where $i \in [2, m]$, to get an output vector $y = \{y_1, y_2, y_3, ...y_m\}$.

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

We compute the loss function as :

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

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 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.

The training is done in two stages:

- **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 3C train dataset:** We only use the Cited paper title scaffold while turning off the other two scaffolds (freezing the task-specific parameters of the other two scaffolds).

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 and 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. The batch size is 12 for SciCite and 8 for 3C Challenge dataset.

4.2 Task B

To identify significant citations, we pursue a feature-engineering approach to curate several features from cited-citing paper pairs. The objective is to classify the citations received by a given paper into INCIDENTAL or INFLUENTIAL. We use the following features for our approach:

1. ***tf-idf* features:** We calculate the cosine similarity between the tf-idf representations of the 1. Titles of cited and citing papers and 2. Citance and the title of the cited paper. Citances are sentences containing the citations in the citing paper. Cited paper titles may contain the contribution/purpose statements of a given paper. Hence similarity with citances may construe that the cited paper may have been used significantly in the current paper.

2. **WMD features:** We measure the Word Mover's Distance (Kusner et al., 2015) between the 1. Title of citing & cited paper, 2. Citance and the title of citing paper, and 3. Citance and the title of cited paper. The intuition is to understand the similarity among the titles of citing/cited papers and the citance in the semantic space.

3. **VADER (Gilbert and Hutto, 2014) Polarity Index - Positive, Negative, Neutral, Compound:** We measure the VADER polarity index to quantify the intensity of the positive/negative emotion of the citance text.

4. **Keyword Overlap:** We compare the number of common keywords between 1. Title of citing & cited paper and 2. Citance and the title of cited paper.

5. **Length of Citance and Title of cited paper:** We compute the length (in words) of the Citance and the Title of the cited paper. Intuition is that if the citing paper has spent many words on the cited article, it may have significantly cited the corresponding article.

6. **Self Citation:** We check if the authors of the citing and cited paper are the same. This might be the case of self-citation or can also signal the extension of the work.

We train various machine learning algorithms (like SVM, KNN, Decision Tree, Random Forest and XGBoost) on the features generated and compare the results. We found out that the Random Forest classifier performs better than all the other classifiers on the validation data. So we use the Random Forest classifier as our best model for our submission in Task B.

5 Results

Table 1 and 2 show the public and private leaderboard scores for each of our submissions for Task A and Task B respectively. For Task A, we analyze the impact of different scaffolds on the performance of the model on the main task. From these experiments, it is evident that each scaffold helps the model to learn the main task more effectively. For Task B, we analyze different machine learning algorithms (like SVM, Random Forest, Decision Tree, KNN, etc) and choose the model that performs the best on our validation data. We have also analyzed the impact of including each feature on the performance of the model. In Task A, we ranked **6th**

Model	Public F1	Private F1
BiLSTM + Attention (with SciBERT)	0.20910	0.20790
BiLSTM-Attn + Section Title scaff. + Cited Paper Title scaff. (with SciBERT)	0.23008	0.25617
BiLSTM-Attn + Section Title scaff. + Cit. Worthiness scaff. (with SciBERT)	0.24958	0.22335
BiLSTM-Attn + Cit. Worthiness scaff. + Cited Paper Title scaff. (with SciBERT)	0.27965	0.26533
BiLSTM + Attention (with SciBERT) + three scaffolds (Our best model)	**0.30258**	**0.26059**
Majority Baseline	0.11938	0.11546

Table 1: Public and private leaderboard macro f1-scores for citation context classification based on purpose (Task A)

Model	Public F1	Private F1
Our Best Model (with all features)	0.45773	0.53588
Our Best Model (with all features except self citation feature)	0.47056	0.4878
Our Best Model (with all features except tf-idf features)	0.4957	0.51165
Our Best Model (with all features except WMD features)	0.49925	0.53758
Our Best Model (with all features except VADER features)	0.49502	0.53351
Our Best Model (with all features except length features)	0.49101	0.53281
Our Best Model (with all features except keyword overlap features)	0.47288	0.52898
3C Task B Best Submission	**0.60699**	**0.60025**
Majority Baseline	0.30362	0.32523

Table 2: Public and private leaderboard macro f1-scores for citation context classification based on influence (Task B)

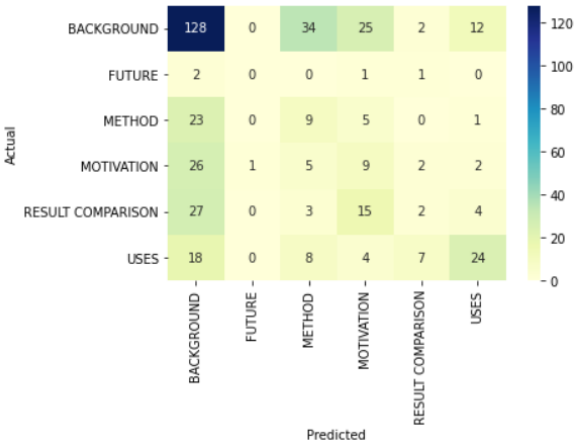

Figure 2: Confusion matrix showing the classification errors of our best model on the validation data (size: 400) for Task A.

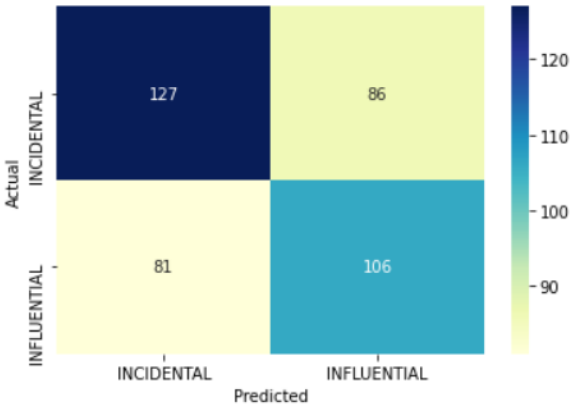

Figure 3: Confusion matrix showing the classification errors of our best model on the validation data (size: 400) for Task B.

(difference of **0.009** in the final Private F1 score between our model and the best performing system) and we ranked **17th** in Task B (difference of **0.0644** in the final Private F1 score between our model and the best performing system).

6 Analysis

Figure 2 and 3 show the confusion matrix of our proposed model on the validation data (400 instances) for Task A and Task B respectively. We investigate the type of errors made by our proposed model on the validation data for both tasks. We found out that for Task A, our model makes many false-positive errors in the BACKGROUND category. To overcome this problem of overfitting, we decided to use some oversampling techniques like

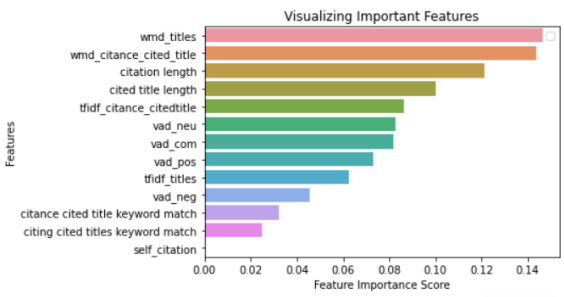

Figure 4: Feature Significance Graph for Task B.

SMOTE, but we still did not get any significant improvements. 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 case of Task B, Figure 4 shows the significance of each feature in helping the model to learn better. We can see that the Word Mover's Distance (measures semantic similarity) features are the most contributing, unlike the VADER (measures sentiment intensity) and the tf-idf features. This trend might be true because in general, research articles have a style of writing that involves significantly less subjective content and follows a more objective discourse. Hence, measurement of semantic similarity becomes the most important feature. We are yet to try many crucial features, but the current list of very simple features that we tried performed better than the majority baseline.

7 Conclusion and Future Work

In this work, we demonstrate our team IITP@3C submissions to the 3C Citation Context Classification Task that included two subtasks. For Task A, we show that the structural information related to a research paper and additional context (title information) of the cited article can be leveraged to classify the citation intent effectively. A future line of research would be to use the abstract of the cited paper as further contextual information for the task and investigate alternative approaches to solve the problem of overfitting on the given data. And we plan to explore additional full-text based features on the Task B data in the future.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Mark Garzone and Robert E Mercer. 2000. Towards an automated citation classifier. In *Conference of the canadian society for computational studies of intelligence*, pages 337–346. Springer.

CHE Gilbert and Erric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, volume 81, page 82.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Bruno Latour. 1987. *Science in action : How to follow scientists and engineers through society*. Harvard University Press.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 337–340.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January, 2015*, volume WS-15-13 of *AAAI Workshops*. AAAI Press.

Xiaodan Zhu, Peter D. Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *J. Assoc. Inf. Sci. Technol.*, 66(2):408–427.