Computer Science 572: Final Project Proposal

Arushi Bhatia, Victor Liu, Qingjie Lu, Jessica Tiu

1 Introduction and Motivation

How a scientific publication is cited in another document is important in assessing the former's impact. Automated analysis of scientific literature must include metrics showing how, not just how many times, a certain publication has been cited - whether as a direct use of a method in that publication, serving as an acknowledgement of previously conducted, or other ways (Cohan, 2019). This type of classification is known as "Citation Intent Classification."

Understanding how a scientific publication has been used allows for more informed analysis of scientific literature. It is more helpful for researchers to be able to search specifically for publications that are cited as the theoretical basis for newer empirical studies, publications that are cited for the results and benchmarks they describe, or publications that are cited for the datasets they contribute, rather than searching for generic mentions of a publication. Improving models that categorize citations could increase the efficiency and quality of scientific research and literature reviews, and clarify the main findings of existing research (Cohan, 2019; Jurgens, 2018).

We choose to conduct original research exploring how to improve existing models for citation intent classification in scientific publications. We will specifically focus on how to obtain such improvements by integrating BERT and Transformer Neural Networks on top of models that use Bi-LSTM and Multilayer Perceptrons (MLP). Our goal is to determine whether adding improvements such as BERT and Transformer Neural Networks improve the F1 score that was achieved by the Cohan, 2019 paper, which was 67.9.

2 Literature Review

Several models have been built to classify citation intent. Previous techniques used neural models, primarily Bi-LSTM and MLP, and token representations such as GloVe and ELMo. *Structural Scaffolds For Citation Intent Classification in Scientific Publications* used the context, defined as the sequence of tokens surrounding the citation, to analyze the author's intention for using a citation. This was compounded by two additional auxiliary models, which were to predict whether a sentence needed a citation and what the heading for a section using a citation was. These two auxiliary tasks aided in the training of the model and were unique because they required no additional datasets to create and were created using existing scientific publications (Cohan, 2019). *Measuring the Evolution of a Scientific Field through Citation Frames* employed several linguistic resources for classification (e.g. presence of any of 23 connective phrases, verb tense), in addition to structural features (e.g. the citation's positions in a paper and the number of other citations in the same subsection, sentence, or clause). It then used this large combination of features to build a random forest classifier (Jurgens, 2018). *Deep Contextualized Word Representations* contributed a new kind of embedding to augment or replace those used in classification models (Sarzynska-Wawer, 2021). Each token was assigned a representation that took into account the entire sentence that contained it. Of these models and model modifications, the structural-scaffolds model currently holds the highest F1 score of 67.9 and achieves this without feature engineering or the feature-rich approach of Jurgens, 2018 (Cohan, 2019; Sarzynska-Wawer, 2021; Jurgens, 2018).

Both Jurgens, 2018 and Cohan, 2019 contributed datasets to aid future research in citation classification. The authors of Jurgens, 2018 created a dataset of annotated citations which was then used to classify the citations

of Natural Language Processing papers. In Cohan, 2019, the authors contributed the SciCite dataset to cover more scientific domains and also provide broader classification labels for citation intent.

3 Datasets and Evaluation Metrics

Our team will utilize the ACL-ARC dataset along with parts of the SciCite dataset created in the *Structural Scaffolds for Citation Intent Classification in Scientific Publications* paper to train our model. The ACL-ARC dataset alone contains 1941 citation intents, and the SciCite dataset contains 11K citation intents gathered through crowdsourcing (Cohan, 2019). We will also use pre-trained word embeddings such as GloVe and ELMo. We will use F1-scores to evaluate the efficacy of our model, and we aim to score higher than 67.9 F1 on the ACL-ARC citations benchmark (which was the metric reached by Cohan, 2019).

4 Tools and Resources

We plan to use PyTorch for our main framework. We aim to reproduce the model implemented in *Structural Scaffolds for Citation Intent Classification in Scientific Publications* using the Bi-LSTM and MLP from PyTorch (Cohan, 2019). Then, we will improve the model with transformers (also using PyTorch libraries) and some hyperparameter tuning. Pre-trained models like BERT (similar to the one in *SciBERT: A pretrained language model for scientific text*) and some of the regularizations and techniques implemented in this paper (Beltagy, 2019) will be incorporated to make further improvement.

With the budget of $200 in Google Cloud Credit, we can continuously train the model on a single Nvidia Tesla v100 with a N1 tier server for about 4 days. Given the size of the dataset, along with free GPU resources to test correctness on Colab we are confident that we will be able to fully train our model a couple of times.

5 Project Milestones

Milestone	**Deadline**
Data preparation, processing, and splitting into training, validation, and test sets	Oct. 29
Establishing baselines for model performance	Nov. 3
Model implementation and training	Nov. 22
Testing, performance evaluation, and creating/giving presentation	Nov. 29
Model optimizations (including hyperparameter tuning and other improvements)	Dec. 11
Final analysis (error analysis ,qualitative and quantitative analysis) and final report	Dec. 16

6 Team Task Breakdown

All team members will equally contribute to the model construction, experiments, presentation, and final report writing. Besides this, each member is responsible for leading several components:

Arushi: Project proposal, data preprocessing, model improvements and integrations
Jessica: Data compilation, data quality assessment
Qingjie: Structural scaffolds model reproduction, model training monitoring
Victor: Model performance evaluation, project presentation, final report writing

References

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Cohan, A., Ammar, W., Van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Jurgens, David & Kumar, Srijan & Hoover, Raine & McFarland, Dan & Jurafsky, Dan. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. Transactions of the Association for Computational Linguistics. 6. 391-406. 10.1162/tacl_a_00028.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.