Citation Intent Classification with Scicite Dataset

Federico Nocentini (7045399)
federico.nocentini@stud.unifi.it

Corso Vignoli (7064123)
corso.vignoli@stud.unifi.it

Abstract

Citation analysis in scientific papers is a vast area of research in machine learning. In fact, statistical approaches are mainly used for this purpose, which does not look into the internal context of the citations. For citation intent analysis, the datasets must have a citation context labeled with different citation intent classes. Most of the datasets either do not have labeled context sentences, or the sample is too small to be generalized. In this study, we tried to replicate the results of the Scicite project [1] (code and data are available at: https://github.com/allenai/scicite). Basically, our project consists of structural scaffolds, a multitask model to incorporate structural information of scientific papers into citations for effective classification of citation intents. We applied Global Vectors (GloVe) and ELMo word embedding methods and compared their F1 measures. It was found that GloVe + ELMo embedding performs significantly better, having an 83.1% Macro F1 score.

1. Introduction

Citations are the keys for understanding and analyzing scientific work. Their nature can be different, so it's crucial to identify the intent of citations (Figure 1) in order to improve automated analysis of academic literature and scientific impact measurement.
The authors approach the problem of citation intent classification by modeling the language expressed in the citation context. A citation context includes text spans in a citing paper describing a referenced work and has been shown to be the primary signal in intent classification. Existing models for this problem are feature-based, modeling the citation context with respect to a set of predefined features (such as linguistic patterns or cue phrases) and ignoring other signals that could improve prediction. In this work we argue that better representations can be obtained directly from data. To this end, we studied a neural multitask learning framework to incorporate knowledge into citations from the structure of scientific papers. In particular, we implemented two auxiliary tasks as **structural scaffolds** to improve citation intent prediction:

1. Predicting the section title in which the citation occurs
2. Predicting whether a sentence needs a citation

It is easy to collect large amounts of training data for scaffold tasks since the labels naturally occur in the process of writing a paper.

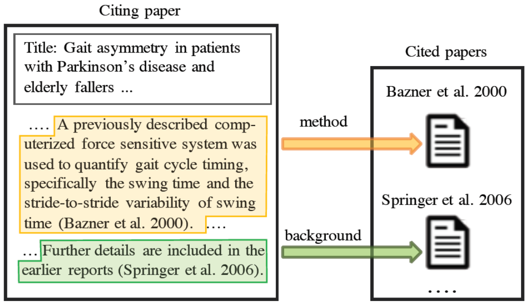
Citing paper
Title: Gait asymmetry in patients with Parkinson's disease and elderly fallers ...
.... A previously described computerized force sensitive system was used to quantify gait cycle timing, specifically the swing time and the stride-to-stride variability of swing time (Bazner et al. 2000). ....
... Further details are included in the earlier reports (Springer et al. 2006).
method
background
Cited papers
Bazner et al. 2000
Springer et al. 2006
....

Figure 1. Example of citations with different intents(BACKGROUND and METHOD).

2. Model

The authors propose a neural multitask learning framework for classification of citation intents. In particular, they introduce and use two structural scaffolds, auxiliary tasks related to the structure of scientific papers. The auxiliary tasks may not be of interest by themselves but are used to inform the main task. Our model uses a large auxiliary dataset to incorporate this structural information available in scientific documents into the citation intents. The overview of their model is illustrated in Figure 2.

Let C denote the citation and x denote the citation context relevant to C. We encode the tokens in the citation context of size n as $x = \{x_1, ..., x_n\}$, where $x_i \in R^{d_1}$ is a word vector of size d_1 which concatenates non-contextualized word representations GloVe [2] and contextualized embeddings ELMo [3], i.e.:

$$x_i = [x_i^{GloVe}; x_i^{ELMo}]$$

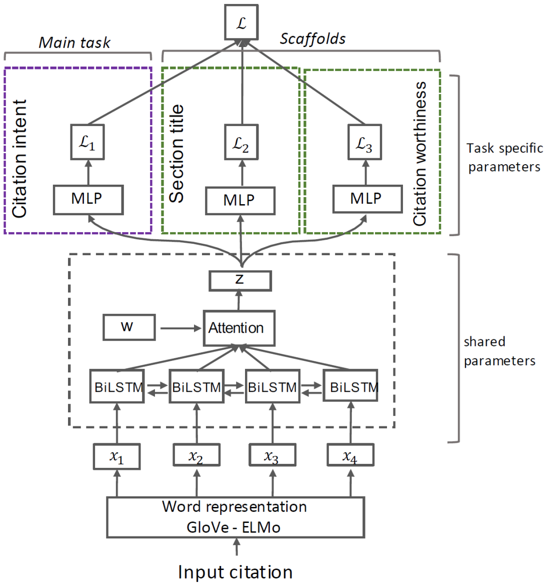

Figure 2. Proposed scaffold model for identifying citation intents. The main task is predicting the citation intent (top left) and two scaffolds are predicting the section title and predicting if a sentence needs a citation (citation worthiness).

We then use a bidirectional long short-term memory BiLSTM [4] network with hidden size of d_2 to obtain a contextual representation of each token vector with respect to the entire sequence:

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

where $h \in R^{(n, 2d_2)}$ and $\overrightarrow{\text{LSTM}}(x, i)$ processes x from left to right and returns the LSTM hidden state at position i (and vice versa for the backward direction $\overleftarrow{\text{LSTM}}(x, i)$). We then use an attention mechanism to get a single vector representing the whole input sequence:

$$z = \sum_{i=1}^{n} \alpha_i h_i, \quad \alpha_i = softmax(w^T h_i),$$

where w is a parameter served as the query vector for dot-product attention. We have now obtained the citation representation as a vector z. Next, we describe our two proposed structural scaffolds for citation intent prediction.

2.1. Structural Scaffolds

To leverage the connection between the structure of scientific papers and the intent of citations, we studied a multitask framework with two structural scaffolds (auxiliary tasks) related to the structure of scientific documents. A key point for the proposed scaffolds is that they do not need any additional manual annotation as labels for these tasks occur naturally in scientific writing. The structural scaffolds in our model are the following:

- **Citation Worthiness** : it indicates whether a sentence needs a citation.
- **Section title** : it predicts the section title in which a citation appears.

We use a multitask formulation approach to inductive transfer learning that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias.
We use a Multi Layer Perceptron (MLP) for each task and then a softmax layer to obtain prediction probabilites. In particular, given the vector z we pass it to n MLPs and obtain n output vectors $y^{(i)}$:

$$y^{(i)} = softmax(MLP^{(i)}(z))$$

We are only interested in the output $y^{(1)}$ and the rest of outputs ($y^{(2)}$, ... , $y^{(n)}$) are regarding the scaffold tasks and only used in training to inform the model of knowledge in the structure of the scientific documents. For each task, we output the class with the highest probability in y.

2.2. Training

Let D_1 be the labeled dataset for the main task $Task^{(1)}$, and D_i denote the labeled datasets corresponding to the scaffold task $Task^{(i)}$ where $i \in (2, ..., n)$. Similarly, let L_1 and L_i be the main loss and the loss of the auxiliary task i, respectively. The final loss of the model is:

$$L = \sum_{(x,y)\in D_1} L_1(x, y) + \sum_{i=2}^{n} \lambda_i \sum_{(x,y)\in D_1} L_i(x, y) \tag{1}$$

where λ_i is a hyper-parameter specifying the sensitivity of the parameters of the model to each specific task. Here we have two scaffold tasks and hence $n = 3$. We train this model jointly across tasks and in an end-to-end fashion. In each training epoch, we construct mini-batches with the same number of instances from each of the n tasks.We compute the total loss for each mini-batch as described in Equation 1, where $L_i = 0$ for all instances of other tasks $j \neq i$. We compute the gradient of the loss for each mini-batch and tune model parameters using the Adam optimizer with learning rate of 10^{-4}.

3. Data

We worked on SciCite, a new dataset of citation intents that addresses multiple scientific domains developed by the authors of the paper. Below there is a description of this dataset:

3.1. SciCite dataset

SciCite is a new dataset of citation intents, significantly larger compared with existing datasets. Through examination of citation intents, the authors found out many of the categories defined in previous work such as motivation, extension or future work, considered as background information providing more context for the current research topic. More interesting intent categories are a direct use of a method or comparison of results. Therefore, the dataset provides a concise annotation scheme that is useful for navigating research topics and machine reading of scientific papers. They consider three intent categories:

- **Background** : The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field
- **Method** : Making use of a method, tool, approach or dataset
- **Result Comparison** : Comparison of the paper's results/findings with the results/findings of other work

In Table 1, we illustrate the distribution of the classes in the dataset and the number of papers used to create it.

Dataset	Categories(Distr.)	Source	#papers	#instances
Scicite	Background (0.58)	Computer	6627	11020
	Method (0.29)	Science &		
	Result Comparison (0.13)	Medicine		

Table 1. Scicite dataset.

3.2. Data for scaffold tasks

For the first scaffold (Citation Worthiness), they sample sentences from papers and consider the sentences with citations as positive labels. They also remove the citation markers from those sentences such as numbered citations or name-year combinations to not make the second task artificially easy by only detecting citation markers. For the second scaffold (Citation Section Title) they sample citations from the ACL-ARC corpus and Semantic Scholar corpus [5] and extract the citation context as well as their corresponding sections.

They manually define regular expression patterns mappings to normalized section titles: “introduction”, “related work”, “method”, “experiments”, “conclusion”. Section titles which did not map to any of the aforementioned titles were excluded from the dataset. Overall, the size of the data for scaffold tasks on the SciCite datset is about 91K and 73K for Section Title Scaffold and Citation Worthiness Scaffolds, respectively.

4. Experiments

4.1. Implementation

The goal of our work is try to replicate the authors results. For our purpose, our datasets have been used as follows : the data was split into three standard stratified sets of train, validation, and test with 85% of data used for training and remaining 15% divided equally for validation and test.

We implement our proposed scaffold framework using Pytorch library. For word representations, we use 300-dimensional GloVe vectors trained on a corpus of 6B tokens from Wikipedia and Gigaword. For contextual representations, we use ELMo vectors with output dimension size of 256 which have been trained on a dataset of 0.8B tokens of news crawl data from WMT 2011. We use a single-layer BiLSTM with a hidden dimension size of 50 for each direction. For each of scaffold tasks, we use a single-layer MLP with 20 hidden nodes , ReLU [6] activation and a Dropout rate [7] of 0.2 between the hidden and input layers.

The hyper-parameters used for the Loss function are :

- No scaffold $\rightarrow\lambda_2 = 0$ $\lambda_3 = 0$
- Section Title scaffold $\rightarrow\lambda_2 = 0$ $\lambda_3 = 0.07$
- Citation Worthiness scaffold $\rightarrow\lambda_2 = 0.09$ $\lambda_3 = 0$
- Both scaffold $\rightarrow\lambda_2 = 0.09$ $\lambda_3 = 0.06$

Batch size is 32 for SciCite dataset. Our best model takes approximately 10 minutes per epoch to train (training time without ELMo is significantly faster). We stop training the model when the development macro F1 score does not improve for five consecutive epochs.

In order to evaluate the accuracy of our models, we used macro F1 score. The Macro F1-score is defined as the mean of class-wise/label-wise F1-scores:

$$MacroF1 = \frac{1}{N}\sum_{i=0}^{N} F1score_i$$

where i is the class index and N the number of classes and

$$F1score_i = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$precision = \frac{T_p}{T_p + F_p}$$

$$recall = \frac{T_p}{T_p + T_n}$$

We used a workstation with the GeForce RTX 3090 graphics card with 24 GB of memory to run our code and train our models.

4.2. Results and Analysis

In order to evaluate the worthiness of our framework, we compared our results with the ones obtained using a baseline. This baseline uses a similar architecture to our proposed neural multitask learning framework, except that it only optimizes the network for the main loss regarding the citation intent classification (L_1) and does not include the structural scaffolds. There are two variants of this model: with and without using the contextualized word vector representations (ELMo). This baseline is useful for evaluating the effect of adding scaffolds in controlled experiments.
In Table 2 , we present our best results.

Model	Our Macro F1	Author Macro F1
BiLSTM + Attn	76.7	77.2
BiLSTM + Attn + ELMo	82.1	82.6
BiLSTM + Attn + Section Title Scaffold	77.6	77.8
BiLSTM + Attn + Citation Worthiness Scaffold	78.0	78.1
BiLSTM + Attn + Both Scaffolds	78.8	79.1
BiLSTM + Attn + Both Scaffolds + ELMo	**83.1**	**84.0**

Table 2. Results on the Scicite citations TestSet.

Table 2 shows the main results on SciCite dataset, where we see similar results compared to those obtained by the authors of the original paper. Each scaffold task improves model performance. Adding both scaffolds results in even major improvements. The best results are obtained by using ELMo representation in addition to both scaffolds. Starting with the 'BiLSTM + Attn' baseline with a macro F1 score of 76.7, adding the first scaffold task in 'BiLSTM + Attn + Section Title Scaffold', this improves the F1 score to 77.6. Adding the second scaffold in 'BiLSTM + Attn + Citation Worthiness Scaffold', this also leads to similar values: 78.0 . When both scaffolds are used simultaneously in 'BiLSTM + Attn + Both Scaffolds', the F1 score further improves to 78.8, suggesting that the two tasks provide complementary signal that is useful for citation intent prediction.

5. Conclusions and Future Developments

In this work, we show that structural properties, related to scientific discourse, can be effectively used to inform citation intent classification. We use a multitask learning model with two auxiliary tasks (predicting section titles and citation worthiness) as two scaffolds related to the main task of citation intent prediction. We demonstrate that carefully chosen auxiliary tasks that are inherently relevant to a main task can be leveraged to improve the performance on the main task.
An interesting line of future work is to explore the design of such tasks or explore the properties or similarities between the auxiliary and the main tasks.

References

[1] A. Cohan, W. Ammar, M. Van Zuylen, and F. Cady, "Structural scaffolds for citation intent classification in scientific publications," *arXiv preprint arXiv:1904.01608*, 2019.

[2] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[3] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," 2018.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] "Semantic Scholar." "https://semanticscholar.org/".

[6] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.