Predicting the presence of inline citations in academic text using binary classification

Elena Callegari
AhmedBahaa ElDIn
Peter Vajdecka
Atli Snær Ásmundsson
Desara Xhura

Abstract

Properly citing sources is a crucial component of any good-quality academic paper, and yet academics, particularly junior ones, often struggle with using citations correctly. With this study we wanted to determine what kind of accuracy we could reach in predicting whether a sentence should or should not contain an inline citation using a simple binary classification model and by training said model on a limited amount of data. Using regular expressions to pre-process data, we created a dataset containing 120k sentences extracted from 2000 existing scientific articles; 50k out of these 120k contained inline citations. We then ran two classification experiments (using RoBERTa and Sentence-BERT) on such dataset. With RoBERTa alone we achieved an precision of 0.87.

1 Introduction

Providing accurate, relevant citations is an essential aspect of good academic research. Not only do citations allow authors to better contextualize the results of their paper, they also lend credibility and authority to the claims made in the article. On the flip side, failing to give credit to existing research when credit is due shows lack of academic integrity and is strongly frowned upon by the academic community. Providing accurate citations however takes some effort to achieve, particularly in light of the increased number of academic texts being made available to the academic community in recent years thanks to the advent of academic preprints and electronic journals. Junior academics such as graduate and undergraduate students also appear to often be unsure about where inline citations should go, and what should or should not be cited (Vardi, 2012) (Carson et al., 1992) (Pennycook, 1996). Having automated techniques that help academics identify if a concept, theory or idea requires attribution would help researchers mitigate the risk of committing plagiarism. One way in which one can automatically determine whether a given paragraph requires (additional) inline citations is through automatic plagiarism detection systems. However, plagiarism detection tools are not always freely available: access to them is often restricted to university staff, with subscription plans that might be very costly. Moreover, processing a document to determine whether some sections of it have been plagiarized may require a considerable amount of time, particularly if the document exceeds a certain length. Finally, building a plagiarism checker is not a trivial process and not always accurate (Kohl Kerstin, 2012), as it requires scanning the full web for documents, requires access to research articles which might lay behind a paywall but also often might fail in finding similarity between concepts just because of the usage a different way of formulating the same sentences. Because of these downsides, we were interested in exploring how much mileage we could get out of a simple binary classification experiment trying to predict whether a given sentence should or should not include an inline citations. In particular, we reasoned that it should be possible to predict at least to some extent whether a sentence should contain an inline citation just by looking at the presence vs. absence of specific lexical cues. For example, verbs such as "claimed", nouns such as "authors" and phrases such "as seen in" tend to appear together or in the vicinity of inline citations. To that end, using regular expressions we created a dedicated dataset containing 120k sentences taken from existing scientific articles. Of these 120k sentences, approximately 42% contains inline citations. We then ran two classification experiments, one leveraging RoBERTa, one using Sentenc-BERT.

References play an essential role in academia and as such have been the target of several NLP studies. In the last years we see an increased tendency towards using references as way to build knowledge graphs (Viswanathan et al., 2021) and

speed up the search for relevant research articles. There is also a tendency towards using references as way to aid automated text summarization (Yasunaga et al., 2019). To our knowledge, however, this is the first study aiming to predict the need for citation in text without going through the need of processing huge amounts of research articles and performing plagiarism checks.

2 Preparing the Data

To get an idea of what types of inline citation styles are used in different disciplines, we randomly selected two articles for each of the following 18 research fields: Medicine, Biology, Chemistry, Engineering, Computer Science, Physics, Math, Psychology, Economics, Political Science, Business, Geology, Sociology, Geography, Environmental Science, Art, History, Philosophy. After analyzing these 36 articles we concluded that most of the articles adopted the IEEE, APA or the Chicago reference styles.

We then created a dataset of 2000 research articles; these were selected from the ArXiv and PubMed datasets (Cohan et al., 2018) that are freely available on Huggingface Datasets library under the name of "scientific papers" (Lhoest et al., 2021) (https://huggingface.co/datasets/scientific_papers). We further added a couple hundred articles from the Humanities to balance our dataset and avoid papers from the hard sciences being overrepresented.

These 2000 articles were subsequently processed to discard articles with a citation pattern other than the IEEE, APA or the Chicago reference styles. The pre-processing task of detecting inline citations was handled through a simple python script. Utilising regular expressions, different kinds of citation styles were mapped to corresponding regex capture patterns. We started by writing regexes that would match the three citation styles that we identified as the most frequently used during the initial analysis stage: IEEE, APA and Chicago. Later on we also decided to include the alpha BibTeX style, as that appears to be pretty frequently used in ArXiV papers.

The Python script does the following: first, every given citation pattern is extracted from the article's plain text. Then, the style with the highest capture count is set as the article's default style. This means that even though the extraction process finds text that matches some style, that is not the article's intended style, the script should still be able to identify the right style (given that the false positives don't exceed the true positives). Finally, the inline citations matching the default style are substituted with an -ADD-CITATION- token; this step is important as it allows us to generalize across different referencing styles. If for some reason no citation style was detected, the token replacement fails, and thus the article is not included.

Using this Python script, we created a second dataset. This second dataset consisted of 120k sentences, of which approximately 50k contained citations. This second dataset was the dataset we used for both classification experiments.

3 First approach: RoBERTa

We present the results of two classification experiments trying to predict whether a given sentence should or should not contain an inline citation. For the first approach, we utilized RoBERTa. Language Models such as BERT, RoBERTa (Devlin et al., 2018; Liu et al., 2019) are considered a breakthrough in the field of Natural Language Processing which is trained in contextual representation fashion. These models are trained in a way called Masked Language Modeling where you hide part of the text and try to predict the [MASK] token using the contextual representation of the neighbouring words. This allows the model to understand the context well enough to be able to replace these [MASK] tokens and have quite rich semantic and contextual information from its output representations. The first approach builds on RoBERTa (Liu et al., 2019), which we use as our main encoder. We tried two different options. First, we tried to take only [CLS] token which holds valuable information about the overall representation of the whole sentence. Then, we took the representation of each token in the sentence. After we extracted the encoded information from RoBERTa, we used a feed-forward neural network as our classification heads. We use a 2-layered feed-forward neural network with dimensions 1024 and 512 respectively (as RoBERTa's output dimension is 1024 for each token). Before feeding the data into our classification head, we use the average global pooling technique proposed by Sentence-BERT (Reimers and Gurevych, 2019) to cram our token representation vectors into a single sentence-level representation vector for the classification part.

4 Second Approach:SENTENCE-BERT with triplet loss

For the second approach, we encoded texts by fine-tuning SENTENCE-BERT using triplet-loss objective function (Reimers and Gurevych, 2019). We also experimented with SENTENCE-BERT without any fine-tuning at all. In both cases, the classifier used was *Logistic Regression* with *L2 regularization*. Combination of fine-tuned SENTENCE-BERT and *Logistic Regression classifier* reached the best results in the classification task.

To fine-tune SENTENCE-BERT using triplet loss, one has to create triplets. We create a triplet as (a, p, n), where a is an anchor sentence, p a hard positive sentence similar (but not identical) to a and n a hard negative sentence different to sentence a with regard to cosine distance. Out data consists of two classes (cited or not cited). Let all the texts in one of the classes form an ordered set of anchors $A = \{a_1, a_2, \ldots, a_t\}$. Then for each anchor $a \in A$, we randomly select another anchor $p \in A$, such that $a \neq p$, which creates an ordered set $P = \{p_1, p_2, \ldots, p_t\}$, denoted as positives. Now, for each anchor a we randomly select text n from another class such that $n \notin A$, which generates an ordered set $N = \{n_1, n_2, \ldots, n_t\}$, denoted as negatives, where t is number of all text pieces in the selected class. Finally, for each anchor a_i we generate triplet (a_i, p_i, n_i).

Each sentence of this triplet is encoded by the same *Siamese network* (Reimers and Gurevych, 2019) denoted as function f. Thus, we minimize the distance between vectors $f(a)$ and $f(p)$ and maximize the distance between vectors $f(a)$ and $f(n)$ by the triplet-loss function as follows:

$$\max\left(d\left(f(a) - f(p)\right) - d\left(f(a) - f(n)\right) + \epsilon, 0\right)$$

, where d means *cosine distance* and ϵ indicates the margin, which means that $f(p)$ is at least ϵ closer to $f(a)$ than $f(n)$.

5 Experimental Settings

For the first approach (ROBERTA) we first experimented with a small subset of the data set and in each of the tests we gradually increased the amount of layers and the data considered. Finally we trained our network with a total of 15000 steps with a batch size of 256 and a learning rate of 0.002.

For the second approach (SENTENCE-BERT), we fine-tuned for 3 epochs with batch size 30 and margin 5. We warmed up with 10% of training data and used the AdamW optimizer with learning rate 2×10^{-5}. The fine-tuned model is *all-mpnet-base-v2*[1].

5.1 Evaluation Results

We report the results for our two experiments in Table 1. Fine-tuned SENTENCE-BERT approach achieved significantly better results when optimized by triplet loss. By fine-tuning Sentence-BERT, we increased precision from 0.55 to 0.77, recall by 0.05, F1 score from 0.62 to 0.77, and finally accuracy by 0.07. The ROBERTA approach, on the other hand, achieved an unexpectedly high precision of 0.87.

6 Conclusion

The goal of this paper was to determine how effective binary classification models can be at predicting whether sentences in academic articles should or should not contain inline citations. To that end, we used regular expressions to identify inline citations in published research papers, and then created a dataset composed of 120k sentences, where approximately 50k contained inline citations. We then ran two classification experiments using such dataset, one using S-BERT and one using RoBERTa.

With the fine-tuned SENTENCE-BERT using the triplet loss we were able to reach an 82% accuracy. With RoBERTa, we reached an precision of 87%. While neither of these values is sensationally high, they are still good enough to conclude that distinct patterns featuring in sentences containing inline citations must exist, and that these patterns *can* be picked up by machine-learning algorithms. Our experiments also indicate that balancing the data set as well as increasing the amount of data plus training layers leads to an increased prediction accuracy.

We presented the problem as a binary classification task on a sentence level, i.e. we only considered the target sentence and did not consider the context in which the given sentence appeared, for example by also looking at the sentence appearing before and the sentence appearing after the target sentence. The sentences contained in our dataset however are all sequential: they come in the same sequence as they were found in the original paper. This means that information on the context

[1] https://www.sbert.net/docs/pretrained_models.html

Citation prediction				
Approach	**Precision**	**Recall**	**F1 score**	**Accuracy**
RoBERTa	**0.87**	0.57	0.69	0.74
Sentence-BERT not fine-tuned + Logistic Regression	0.55	0.72	0.62	0.76
Sentence-BERT fine-tuned + Logistic Regression	0.77	**0.77**	**0.77**	**0.83**

Table 1: Prediction results

in which a given target sentence appears is already available in our dataset. This paves the path for further experimentations that take contextual sentential information into account, such as training a transformer that tries to predict in which position a reference should come or even open doors towards experimenting with techniques that could be used to identify the scope of a reference and use the full scope for the classification part as well.

References

Joan G. Carson, Nancy D. Chase, Sandra U. Gibson, and Marian F. Hargrove. 1992. Literacy demands of the undergraduate curriculum. *Literacy Research and Instruction*, 31(4):25–50.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eleonora Kohl Kerstin. 2012. Fostering academic competence or putting students under general suspicion? voluntary plagiarism check of academic papers by means of a web-based plagiarism detection system. *Research in Learning Technology*, 19.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alastair Pennycook. 1996. Borrowing others' words: Text, ownership, memory, and plagiarism. *TESOL quarterly*, 30(2):201–230.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Iris Vardi. 2012. Developing students' referencing skills: a matter of plagiarism, punishment and morality or of learning to write critically? *Higher Education Research Development*, 31(6):921–930.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. Citationie: Leveraging the citation graph for scientific information extraction.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. 2019. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*.