Automating Citation Placement with Natural Language Processing and Transformers

Davide Buscaldi[1], Danilo Dessí[2,*], Enrico Motta[3], Marco Murgia[4], Francesco Osborne[3,5] and Diego Reforgiato Recupero[4]

[1]*Laboratoire d'Informatique de Paris Nord, Sorbonne Paris Nord University, Paris, France*
[2]*GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany*
[3]*Knowledge Media Institute, The Open University, Walton Hall, Kents Hill, Milton Keynes, MK76AA, United Kingdom*
[4]*Mathematics and Computer Science Department, University of Cagliari, Cagliari, Italy*
[5]*Department of Business and Law, University of Milano Bicocca, Milan, Italy*

Abstract

In scientific writing, references are crucial in supporting claims, spotlighting evidence, and highlighting research gaps. However, where to add a reference and which reference to cite are subjectively chosen by the papers' authors; thus the automation of the task is challenging and requires proper investigations. This paper focuses on the automatic placement of references, considering its diverse approaches depending on writing style and community norms, and investigates the use of transformers and Natural Language Processing heuristics to predict i) if a reference is needed in a scientific statement, and ii) where the reference should be placed within the statement. For this investigation, this paper investigates two techniques, namely Mask-filling (MF) and Named Entity Recognition (NER), and provides insights on how to solve this task.

Keywords

Citation Prediction, Named Entity Recognition, Generative Approach, Natural Language Processing

1. Introduction

Citing research papers is common practice to provide evidence, build upon existing knowledge, and substantiate findings. This activity relies on two main tasks: i) authors have to decide where to place a citation based on the written statements, and ii) the cited papers should align with the content of the factual information expressed in the sentences. The first task depends on the author's experience in adding references to support claims and does not need knowledge about the state-of-the-art literature. The second task requires domain and state-of-the-art expertise to refer to the best and most timely literature. Although these two tasks are related to each

SemTech4STLD '24: Second International Workshop on Semantic Technologies for Scientific, Technical and Legal Data, May 26th, 2024, Hersonissos, Greece.

*Corresponding author.

davide.buscaldi@lipn.univ-paris13.fr (D. Buscaldi); danilo.dessi@gesis.org (D. Dessí); enrico.motta@open.ac.uk (E. Motta); marco.murgia3@unica.it (M. Murgia); francesco.osborne@open.ac.uk (F. Osborne); diego.reforgiato@unica.it (D. R. Recupero)
0000-0003-1112-3789 (D. Buscaldi); 0000-0003-3843-3285 (D. Dessí); 0000-0003-0015-1952 (E. Motta); 0009-0003-8994-4031 (M. Murgia); 0000-0001-6557-3131 (F. Osborne); 0000-0001-8646-6183 (D. R. Recupero)

other they can be studied independently; this paper focuses on the first task. In more detail, this paper proposes:

- A model designed to solve a mask-filling problem using a generative approach.
- A Named Entity Recognition (NER) model designed to label tokens that should precede a citation.

We analysed both methods as well as their typical errors. We also introduced a few heuristics that are able to solve some of the common issues and improve their performance. The proposed models are investigated on a manually annotated gold standard. In summary, the contribution of this paper is two-fold:

- An analysis of a mask-filling and NER approach for the reference placement problem.
- An error analysis and heuristics to solve the identified errors.

We provide the code as well as the data used in our analysis through a GitHub repository[1].

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the proposed models, their results, and an error analysis. Finally, Section 4 concludes the paper and outlook our future research towards a fully-fledged knowledge-aware reference recommendation system

2. Related Work

The task of deciding which statements need a citation is relatively recent. To start with, there exist resources that provide information about why a reference is cited within a sentence. For example, CiTO [1] provides more than 20 citation types to describe the nature of cited references in scholarly works. This resource is already leveraged in studies to automatically predict the intent behind a citation. For example, in [2] the applicability of Convolutional Neural Network (CNN) as well as latent representation are examined to classify research intents. Another recent study in this direction is [3] where a hybrid approach combining graph and textual embedding features is proposed to classify citation intents into 6 different categories. Other works related to our examination fall mostly under the umbrella of suggesting a potential suggestion; for example, in [4] the authors first identify paper topics using Restricted Boltzmann Machines, then they adopted Kullback-Leibler distance to align the extracted topics with candidate references. In literature is also possible to find tools implementing recommendation systems that use the context given a reference placeholder to suggest the best references [5]. Recently, we saw a proliferation of LLM-based services for supporting academic writing [6], some of which can also suggest relevant citations (e.g., Textero.ai[2], Jenni.ai[3]). Some other solutions make use of conversational agents that can support researchers by retrieving and summarizing specific papers [7]. These systems often rely on scholarly knowledge graphs [8, 9, 10, 11] that describe

[1] https://github.com/Marcomurgia97/Citation-Prediction-by-Leveraging-Transformers-and-Natural-Language-Processing-Heuristics

[2] https://textero.ai/

[3] https://jenni.ai/

networks of papers according to a variety of metadata and can support advanced research analysis [12, 13] as well as hypothesis generation [14].

However, the task of determining whether a statement in a research paper needs a citation is less explored and only recently has raised the interest of the community working on Artificial Intelligence (AI) systems for the scholarly domain. For example, the problem is explored in [15] where a multi-layer perceptron model is proposed to measure the citation worthiness of scientific statements. Authors in [16] tackled the task with various scenarios, including sentence classification with and without context representation, and sequence modeling using contextual embeddings and BiLSTMs. While their results showed significant accuracy improvements with context, their focus remained on deciding whether a citation is necessary, not its precise placement within the text. Most recently, large language models have also been explored for this task. For instance, Vajdecka et al. [17] operated a Large Language Model (LLM) to predict whether a citation is needed in a sentence, achieving F1-scores between 75% and 89%. However, their approach does not pinpoint the optimal placement of the citation within the sentence. Additionally, LLMs can exhibit inherent instability even with clear instructions, making them unsuitable for tasks demanding consistent performance.

3. Reference Placement Analysis

This section outlines the task investigated in our analysis, the proposed models, and the resulting evaluation.

3.1. Task Definition & Approaches

The tasks investigated in this paper are the *Citation Required Task* and *Citation Placement Task*. *Citation Required Task* has the main goal of determining whether a sentence needs a citation. It can be seen as a binary classification task that given a scientific statement as an input returns 1 if at least one citation is needed, 0 otherwise. *Citation Placement Task* has the main objective to extend the *Citation Required Task* by correctly predicting the tokens that should be followed by citations. The reader notices that our work does not include the classification of the citation intents. These two tasks are addressed using mask-filling and NER approaches.

Mask-Filling Approach. The first solution is based on a mask-filling approach where certain words on the original input are masked and the model has the assignment of predicting appropriate words that can be used to replace the masked ones. To solve the above tasks with this strategy, we adopted a generative approach. In more detail, given a sentence s, its tokens $T^s = \{t_0^s, \dots, t_n^s\}$, and a mask $MASK$, the mask was moved over the tokens to feed the model. More precisely, the model was iteratively presented with m tokens at each iteration where $2 < m < n$. For example, for a sentence $s1$ with tokens $T^{s1} = \{t_0^{s1}, \dots, t_3^{s1}\}$, the model is fed with $t_0^{s1}, t_1^{s1}, MASK$ in the first iteration, $t_0^{s1}, t_1^{s1}, t_2^{s1}, MASK$ in the second iteration, and $t_0^{s1}, t_1^{s1}, t_2^{s1}, t_3^{s1}, MASK$ in the third iteration. When the model predicts a reference placeholder, this is recorded by our system and the next iteration is investigated.

NER Approach. The second approach investigated in this paper is a NER approach that usually classifies entities, such as names of persons, organizations, locations, dates, numerical values, and other relevant information, within a given text. In the context of citation placement prediction, the NER model is set to classify each token of an input sentence into *REGULAR* and *CITATION* tokens. The former indicates tokens that should not be followed by a citation. The latter indicates tokens which should be followed by a citation.

3.2. Transformers

For our exploration, we focused on the following models:

- **GPT-2**. The generative methodology uses the text-generation capabilities inherent in transformer models. Among these models, the Generative Pre-trained Transformer (GPT) family by OpenAI stands out as one of the most widely employed. In our approach, we specifically utilized the GPT-2 transformer [18]. GPT-2 is a generative language model pre-trained on approximately 40 GB of unlabeled data. Its function is to predict the subsequent token given an input phrase or word. Our choice fell on the variant of GPT-2 equipped with 117 million parameters and 12 layers. This selection was driven by GPT-2's accessibility, integration within popular deep-learning libraries, and the possibility of running in low-setting machines.
- **BERT** [19]. BERT, short for Bidirectional Encoder Representations from Transformers, is renowned for its encoder-only architecture, featuring stacked layers that refine text representations. Our use of the BERT-base, comprising 12 layers and 110 million parameters, underscores its usefulness in text classification tasks. BERT's bidirectional processing captures contextual information from surrounding words, enabling a nuanced understanding of language.

3.3. Experimental Setting & Analysis

This section describes the experimental setting, the evaluation of the models, and our pre-processing steps to further enhance the overall results.

3.3.1. Datasets & Model Implementation

To set the analysis of the models described above it is deemed to exploit models that have been trained on scholarly data. For this reason, two strategies are possible: i) use an already trained model, and ii) fine-tune a pretrained model on newly prepared data. For this, we referred to two datasets. The first dataset on which our analysis is based is *arXiv-80* a dataset consisting of plain text containing citations from *arXiv*. A limitation of this dataset is that it contains citations in different formats, a characteristic that might mislead models for the above-mentioned tasks. Therefore, we also created a new dataset *s2orc-9k*, a dataset containing 9,000 papers about Computer Science from *s2orc* where citations have a unique format. These datasets were used to create and experiment with the following models:

- **NER-s2orc**, a BERT-based model that has been fine-tuned on the NER task.
- **ArXiv-NLP GPT-2**[4], a pretrained model on the *arXiv-80* dataset with papers from the computational linguistic field.

[4] https://huggingface.co/lysandre/arxiv-nlp

Table 1
Results of the models on the Citation Required Task

Model	Precision	Recall	F1
NER-s2orc	0.950	0.275	0.426
ArXiv-NLP GPT-2	0.818	0.652	0.725
GM-s2orc	0.782	0.782	0.782

Table 2
Results of the models on the Citation Placement Task

Model	Precision	Recall	F1
NER-s2orc	0.545	0.173	0.263
ArXiv-NLP GPT-2	0.363	0.289	0.323
GM-s2orc	0.449	0.449	0.449

- **GM-s2orc**, the ArXiv-NLP GPT-2 model fine-tuned on the new dataset s2orc-9K.

Finally, we created a gold standard of 170 sentences from papers about Computer Science from the year 2023 that were manually labeled by three senior researchers with more than 100 papers in Computer Science. The annotators were asked to label tokens which must be followed by a citation. The computed inter-annotator agreement was 79.7%. This dataset serves for the evaluation of the *Citation Required Task* as well as *Citation Placement Task*.

3.3.2. Results

We evaluated the models with precision, recall, and f1-score. While for the *Citation Required Task* we used standard metrics, for the *Citation Placement Task* we defined true positives as the correctly predicted tokens immediately preceding a citation, true negatives as the correctly predicted tokens that do not precede a citation, false negatives tokens that precede a citation but erroneously predicted and, finally, false positives as the erroneously predicted tokens that do not actually precede a citation. The results of the three experimented models are reported in Table 1 and Table 2 for the *Citation Required Task* and *Citation Placement Task* respectively. Both tables show that the generative approach is more suitable for the proposed task of obtaining a relatively high f1-score. In addition, both tables show that injecting well-formatted features during the fine-tuning of the generative approach further improved the model performance.

3.3.3. Post-processing Analysis

In addition, to the evaluation above we performed a manual inspection of the predicted citation placement from the various models. This revealed some error patterns that can be easily solved by some natural language processing (NLP) heuristics that implement common practice in governing paper citations. Here, the reader can find the kind of error and a description of the heuristic to solve it:

- *Citation placeholder between an acronym and its extended terms.* The heuristic to solve this error takes the citation placeholder predicted and moves it after the acronym.
- *Citation placeholder after Table, Fig., Figure, etc.* The heuristic simply removes the citation placeholder appearing after these keywords.
- *Citation placeholder after a verb.* The heuristic labels the tokens of the input text with part-of-speech tags, and removes the predicted citation placeholders appearing after

Table 3
Results of the models enhanced with the heuristics on the Citation Required Task

Model	Precision	Recall	F1
NER-s2orc	0.909	0.434	0.588
ArXiv-NLP GPT-2	0.872	0.695	0.774
GM-s2orc	0.802	0.826	0.814

Table 4
Results of the models enhanced with the heuristics on the Citation Placement Task

Model	Precision	Recall	F1
NER-s2orc	0.515	0.246	0.333
ArXiv-NLP GPT-2	0.509	0.405	0.451
GM-s2orc	0.492	0.507	0.500

tokens with tags VB, VBD, VBG, VBN, VBZ, and VBP. The SpaCy[5] library was used for this task.

- *Citation placeholder predicted within a noun phrase.* The heuristic labels the tokens of the input text with part-of-speech tags, and detects the noun phrase tokens; if a citation placeholder appears in a noun phrase, then it is moved to the last token of that noun phrase.
- *Citation placeholder predicted in consecutive tokens.* The heuristic only keeps the last predicted placeholder, discarding the ones appearing in the previous tokens.
- *Missing citation placeholders after common phrases.* Certain phrases such as previous work, prior studies, etc., are often followed by a citation. The heuristic adds a citation as a placeholder after these phrases if they appear in the text. The list of these phrases is available in the repository of this paper.

These heuristics were applied to the results of the models above providing further improvements in solving the *Citation Required Task* and *Citation Placement Task* as shown in Table 3 and Table 4. Interestingly, it revealed a downgrading of the performance of the NER approach in terms of precision that nevertheless did not undermine the overall model, resulting in a higher f1-score of 0.162 for the *Citation Required Task*, and of 0.07 for the *Citation Placement Task*. For the generative approach, the heuristics increased the models' performance in terms of precision and recall. This demonstrates that including insights from citation placement governance enhances the overall approach, leading to improved performance.

4. Conclusion and Outlook

In this paper we have investigated two tasks for recognizing whether a scientific sentence needs a citation, and if yes, after which token the citation should be placed. For this, we have explored the NER and generative approaches obtaining promising results. We have also provided some heuristics to better shape the output of the used models and incorporate common practice in citation placement into our approach. However, our work needs to further develop in a few directions. More precisely, we plan to complete our approach by investigating and creating components to recommend potential papers that replace the citation placeholder. To achieve our goal, we intend to utilize knowledge graphs [8] that describe and interlink scientific literature and relevant concepts, such as the CS-KG [9, 20] and AIDA-KG [10]. Additionally, we

[5] https://spacy.io/

aim to assess the capability of large language models, potentially enhanced with knowledge injection methods [21], in identifying the most relevant papers for citation. This will also involve investigating citation intents [15] and exploring various scientific fields that may exhibit unique citation behaviors.

References

[1] D. Shotton, Cito, the citation typing ontology, in: Journal of biomedical semantics, volume 1, Springer, 2010, pp. 1–18.

[2] A. Lauscher, G. Glavaš, S. P. Ponzetto, K. Eckert, Investigating convolutional networks and domain-specific embeddings for semantic classification of citations, in: Proceedings of the 6th international workshop on mining scientific publications, 2017, pp. 24–28.

[3] D. Berrebbi, N. Huynh, O. Balalau, Graphcite: Citation intent classification in scientific publications via graph embeddings, in: Companion Proceedings of the Web Conference 2022, 2022, pp. 779–783.

[4] J. Tang, J. Zhang, A discriminative approach to topic-based citation recommendation, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2009, pp. 572–579.

[5] Q. He, J. Pei, D. Kifer, P. Mitra, L. Giles, Context-aware citation recommendation, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 421–430.

[6] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial intelligence for literature reviews: Opportunities and challenges, arXiv preprint arXiv:2402.08565 (2024).

[7] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, Ieee Access 11 (2023) 22468–22489.

[8] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial Intelligence Review (2023) 1–32.

[9] D. Dessì, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, CS-KG: A large-scale knowledge graph of research entities and claims in computer science, in: The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings, volume 13489 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 678–696.

[10] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Aida: A knowledge graph about research dynamics in academia and industry, Quantitative Science Studies 2 (2021) 1356–1398.

[11] M. Stocker, A. Oelen, M. Y. Jaradeh, M. Haris, O. A. Oghli, G. Heidari, H. Hussein, A.-L. Lorenz, S. Kabenamualu, K. E. Farfar, M. Prinz, O. Karras, J. D'Souza, L. Vogt, S. Auer, Fair scientific information with the open research knowledge graph 1 (2023) 19–21. doi:10.3233/FC-221513.

[12] S. Angioni, A. Salatino, F. Osborne, A. Birukou, D. R. Recupero, E. Motta, Leveraging knowledge graph technologies to assess journals and conferences at springer nature, in: International Semantic Web Conference, Springer, 2022, pp. 735–752.

[13] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, T. Vergoulis, New trends in scientific knowledge graphs and research impact assessment, 2021.

[14] A. Borrego, D. Dessi, I. Hernández, F. Osborne, D. R. Recupero, D. Ruiz, D. Buscaldi, E. Motta, Completing scientific facts in knowledge graphs of research concepts, IEEE Access 10 (2022) 125867–125880.

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

[16] R. Gosangi, R. Arora, M. Gheisarieha, D. Mahata, H. Zhang, On the use of context for predicting citation worthiness of sentences in scholarly articles, arXiv preprint arXiv:2104.08962 (2021).

[17] P. Vajdecka, E. Callegari, D. Xhura, A. S. Asmundsson, Predicting the presence of inline citations in academic text using binary classification, in: The 24rd Nordic Conference on Computational Linguistics, 2023.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[20] D. Dessí, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, Scicero: A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain, Knowledge-Based Systems 258 (2022) 109945.

[21] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Engineering Applications of Artificial Intelligence 133 (2024) 108166.