Information Processing and Management 61 (2024) 103583

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Citation prediction by leveraging transformers and natural language processing heuristics

Davide Buscaldi [a], Danilo Dessí [b], Enrico Motta [c], Marco Murgia [d], Francesco Osborne [c,e], Diego Reforgiato Recupero [d,*]

[a] Laboratoire d'Informatique de Paris Nord, Sorbonne Paris Nord University, Paris, France
[b] Knowledge Technologies for Social Sciences Department, GESIS Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany
[c] Knowledge Media Institute, The Open University, Walton Hall, Kents Hill, Milton Keynes, MK76AA, United Kingdom
[d] Department of Mathematics and Computer Science, Via Ospedale 72, Cagliari, 09121, Italy
[e] Department of Business and Law, University of Milano Bicocca, Via Bicocca degli Arcimboldi, 8, Milan, 20100, Italy

ARTICLE INFO

Keywords:
Citation prediction
Transformers architecture
Mask-filling
Named entity recognition
BERT

ABSTRACT

In scientific papers, it is common practice to cite other articles to substantiate claims, provide evidence for factual assertions, reference limitations, and research gaps, and fulfill various other purposes. When authors include a citation in a given sentence, there are two considerations they need to take into account: (i) where in the sentence to place the citation and (ii) which citation to choose to support the underlying claim. In this paper, we focus on the first task as it allows multiple potential approaches that rely on the researcher's individual style and the specific norms and conventions of the relevant scientific community. We propose two automatic methodologies that leverage transformers architecture for either solving a Mask-Filling problem or a Named Entity Recognition problem. On top of the results of the proposed methodologies, we apply ad-hoc Natural Language Processing heuristics to further improve their outcome. We also introduce s2orc-9K, an open dataset for fine-tuning models on this task. A formal evaluation demonstrates that the generative approach significantly outperforms five alternative methods when fine-tuned on the novel dataset. Furthermore, this model's results show no statistically significant deviation from the outputs of three senior researchers.

1. Introduction

In scientific papers, it is common practice to cite other articles to substantiate claims, provide evidence for factual assertions, reference limitations, and research gaps, and fulfill various other purposes. Consequently, acquiring appropriate referencing skills becomes essential for researchers and students in higher education.

When authors decide to include a citation in a given sentence, they are confronted with two key considerations:

1. where in the sentence to place the citation, and
2. which citation to choose to support the underlying claim.

The first task presents considerable complexity and may allow for multiple potential solutions, depending on the researcher's individual style and the specific norms and conventions of the targeted scientific community. For example, in the sentence "*CLIP*

* Corresponding author.
E-mail address: diego.reforgiato@unica.it (D. Reforgiato Recupero).

https://doi.org/10.1016/j.ipm.2023.103583
Received 11 August 2023; Received in revised form 11 October 2023; Accepted 12 November 2023
Available online 16 November 2023

is a state-of-the-art vision-language model that was originally developed by OpenAI", a citation can be either included right after the token *CLIP* or at the end of the sentence to properly justify the relevant claim. In the scientific domains, authors tend to position citations in the middle of the sentence as pointed out by researchers in Cerovšek and Mikoš (2014) and Yu and Yan (2023). For example, citations can be situated immediately after the name of a particular method or data, after an acronym, or at the end of the sentence. Conversely, the same study has also noticed that, in non-scientific domains, citations are usually positioned at the end of the sentence. Sometimes, the presence of particular expressions such as *recent works* or *previous studies* also necessitates the presence of a citation.

The second task, i.e., selecting the appropriate citation to support the underlying claim primarily depends upon the nature of the relevant claim. Although the two tasks are strictly connected, they do not necessarily depend on each other and, therefore, can be studied independently.

Here we focus on the first task, leaving the second one for future research endeavors. Our objective is to develop a "virtual assistant" that could be used by researchers to improve their productivity to identify pertinent literature regarding a part of a scientific work or a project proposal. This assistant would include the capability to identify parts of the text that are citation-worthy. Consequently, it is paramount to devise a methodology applicable during the writing process, so that can also be used on single sentences lacking surrounding context. T Specifically, we study how transformers can be used to decide whether a sentence needs a citation and on which token it should be placed. Transformers (Vaswani et al., 2017) are well-known deep-learning models that are widely used in several fields, including speech recognition, computer vision, and Natural Language Processing (NLP). They were first proposed as a machine translation sequence-to-sequence model, but further studies demonstrated that pre-trained models based on Transformers can perform at the cutting edge on a variety of tasks and on different domains (Lin, Wang, Liu, & Qiu, 2022). Transformers have consequently emerged as the preferred design in NLP (Jain, 2022). In recent years, methods using large-scale transformer-based models have become a new paradigm of Natural Language Generation (NLG), allowing the generation of more diverse and fluent text (Liu et al., 2023).

In this paper, we propose and evaluate two methodologies for determining the necessity and placement of citations within a sentence. These solutions can be adopted by a variety of tools for assisting students and researchers in their scientific writing endeavors. The first methodology conceptualizes the task as a *Mask-Filling* problem and leverages a generative model to solve it. The second strategy formulates it as a *Named Entity Recognition* task and aims to identify the tokens that should immediately precede a citation. We refined the outcomes of both methods by applying a set of NLP heuristics, further enhancing the accuracy of the predictions. In order to facilitate the fine-tuning of transformers models for this task, we constructed *2orc-9K*, a novel open dataset including 9,000 articles where citations have been replaced with placeholders.

We evaluated the proposed methodologies and the heuristics over a gold standard of manually annotated sentences. The generative approach fine-tuned on *2orc-9K* outperformed the other methods. Particularly noteworthy is that this approach yielded results comparable to those achieved by researchers. To further substantiate these results, we conducted a statistical analysis to assess the consistency between senior researchers and computational models. The results of our investigation reveal that our proposed solution exhibits a level of performance that is statistically indistinguishable from that of experienced researchers.

In summary, the contributions of this paper are the following:

- We formulate the problem of citation prediction and map it to both Mask-Filling and Named Entity Recognition problems.
- We propose two new transformer-based methodologies based on these formalizations.
- We developed a set of NLP heuristics that can significantly enhance the performance of both methodologies.
- We released *s2orc-9K*, an open dataset for fine-tuning models on this task, and a new gold standard composed of 133 sentences annotated by three senior researchers.
- We carried out a comparative evaluation of the two proposed methodologies against humans. We also present an in-depth analysis of the best method, investigating the impact of different parameters, training set sizes, and citation intents.
- We present a statistical analysis demonstrating that one of the proposed generative models yields results that are not statistically different from those achieved by three senior researchers ($p = 0.58$).
- We provide the full codebase for implementing our methodologies, the relevant datasets, and the gold standard.[1]

The remainder of the paper is organized as follows. Section 2 discusses related works and the differences with our methodologies. Section 3 contains the research objectives and, as such, formally defines the task we aim to solve. The datasets used for training and evaluating the methodologies are presented in Section 4. Section 5 discusses in detail the two new methodologies and the approaches we implemented for each of them. Section 6 reports the evaluation and the statistical analysis. Finally, Section 7 ends the paper with conclusions and future directions of research.

2. Related works

In literature, various works have addressed the prediction or recommendation of citations for a given research paper. One of the first mentions of "citation prediction" was in the 2003 KDD cup (Gehrke, Ginsparg, & Kleinberg, 2003), a Data Mining and Knowledge Discovery competition organized by the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). The task consisted of the prediction of the number of citations that a paper would receive over time, therefore, it resembled a

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

bibliometric exercise more than a recommendation one. One of the first works approaching citation recommendation from a text-based perspective is the one from Tang and Zhang (2009). In their work, they carried out a topic-based recommendation, applying Restricted Boltzmann Machines to identify sub-topics in a paper, thus creating a context for it; then they suggested citations using a measure based on the Kullback–Leibler distance of context representation to the candidate paper. He, Pei, Kifer, Mitra, and Giles (2010) went further in this direction and proposed a recommendation system that was able to extract local contexts, where "local" refers to the context that surrounds citations. In their work, they proposed a tool to assist researchers, where the user puts a marker for a possible citation and obtains candidate papers that would fit the context specified by the user.

Some research works addressed the citation intent, that is what a citation is used for when it is included in a text. CiTO (Peroni & Shotton, 2012) is an ontology describing the nature of reference citations in scientific research articles and other scholarly works, providing more than 20 citation types. Lauscher, Glavaš, Ponzetto, and Eckert (2017) studied the applicability of Convolutional Neural Networks and embeddings to the semantic classification of citation intents. GraphCite (Berrebbi, Huynh, & Balalau, 2022) used a hybrid graph and textual-based embedding method to classify citation intents into 6 different types on the ACL-ARC dataset (Bird et al., 2008).

However, works addressing the task of determining if a sentence needs citations or the prediction of the position of a citation in a text are rare. This is a problem that only recently attracted some interest from the research community. Boyack, van Eck, Colavizza, and Waltman (2018) carried out a study over a large collection of research papers in different domains to identify when and where citations occur. They showed on average about 20% of sentences contain a reference, while less than 2% of sentences contain three or more mentions. They also found out that most citations occur within the first 5%–10% centiles of the studied documents. Cohan, Ammar, van Zuylen, and Cady (2019) used a Multi-Layer Perceptron model to determine the "citation worthiness" of a paragraph. They also proposed a dataset (SciCite) for citation intent prediction that is larger than ACL-ARC. Vajdecka, Callegari, Xhura, and Ásmundsson (2023) recently proposed a Large Language Model-based method to predict whether a citation should appear or not in a sentence, obtaining F1-scores between 75% and 89%. Finally, Gosangi, Arora, Gheisarieha, Mahata, and Zhang (2021) approach the task as three different scenarios: sentence classification, sentence classification including a representation of its context, and sequence modeling, using contextual embeddings (RoBERTA) and BiLSTMs. Their results show a significant increase in accuracy when the context is taken into account. In any case, these models do not predict the exact position of a citation in the text, but they limit to the problem of determining whether a sentence should contain a citation or not.

In this paper, we try to close this gap by providing two novel methodologies to determine whether a sentence from a scientific paper needs a citation and to identify the most suitable placement.

3. Research objectives

The task we target in this paper is to decide whether a sentence requires a citation and, in case it does, identify the most appropriate placement. We do not specify which reference should be used but limit our task to the identification of a placeholder where a citation should be positioned. We also do not consider different citation intents (Roman, Shahid, Khan, Koubaa, & Yu, 2021).

We can think of this task at coarse and fine-grained levels:

1. **Coarse-grained level**: determining whether a sentence necessitates the inclusion of a citation.
2. **Fine-grained level**: determining whether a sentence requires a citation and, if it does, identifying the position of a given sentence where a citation has to be placed.

The formulation at the coarse-grained level can be seen as a binary classification task that takes as an input a sentence s from a research paper and returns as an output a Boolean value from the set $\{0, 1\}$; 0 indicates that s does not need a citation, 1 indicates that s should include at least one citation.

The formulation at the fine-grained level extends the first one and aims to identify whether a sentence needs a citation, and, if it does, the token in the sentence that should precede a citation. Formally, given the sentence s and its tokens $T^s = \{t_0^s, \ldots, t_n^s\}$, the task is the identification of the token $t_i^s \in T^s$ which should be followed by a citation if this should appear in the sentence.

For instance, consider the sentence: "*Therefore, the mask tensor search process is easily trapped in a bad local optimal because of its low global exploration efficiency or needs a longer time to fully explore the loss landscape*". Our objective is to identify potential citation positions, such as after the token *efficiency*, after the token *landscape*, or at other suitable locations. As a result, the output for this sentence would resemble the following: "*Therefore, the mask tensor search process is easily trapped in a bad local optimal because of its low global exploration efficiency* [ ] *or needs a longer time to fully explore the loss landscape* [ ]", where [ ] denotes a placeholder for a citation. It is important to note that we refrain from making any assumptions regarding the actual citations that would fill these placeholders, leaving this problem for future research endeavors.

The formulation at the fine-grained level may be formalized either as a *Mask-Filling* task or as a *Named Entity Recognition* (NER) task.

The Mask-Filling task involves concealing certain words within a sentence by replacing them with masks and subsequently predicting the appropriate words that should occupy these masked positions (Devlin, Chang, Lee, & Toutanova, 2019).

When framing the problem as a Mask-Filling task, we adopt a generative approach. Given the input sentence s containing the citation, we apply the mask as follows: first to the last token of the sub-sentence containing the initial two tokens from s, then to the last token of the sub-sentence containing the first three tokens from s, and so forth. At each iteration, if the mask is filled with a token representing a placeholder for a potential citation, the corresponding placeholder is inserted in that position.

More precisely, we can iteratively present a sentence to a transformer model for text generation, starting with the first token and gradually increasing the input length, and record where the transformer predicts a citation as the next token. The reader is referred to https://huggingface.co/tasks/fill-mask for further details and examples.

Alternatively, we can approach the task of predicting citations by formalizing it as a NER problem with just two types of entities. NER is a natural language processing technique that involves identifying and classifying specific entities, such as names of persons, organizations, locations, dates, numerical values, and other relevant information, within a given text (Mollá, van Zaanen, & Smith, 2006). For example, consider the following sentence: "*I know that Sardinia is wonderful in summer.*" The NER basically operates through two well-defined steps: it determines whether and which pre-defined entities are present in the text. In this case, we assume that the entities are *Sardinia* and *summer*. The next step is to label the entities found: in our example, Sardinia is labeled as *Place* and summer as *Time*. All the other tokens will be labeled with a value indicating they do not belong to any predefined type of entity.

In the case addressed in this paper, we aim to classify tokens within unstructured text into two categories: *NORMAL* if the token is not supposed to precede a citation and *CITATION* if the token should precede a citation.

Consider for example the following sentence: "*The aim of diminishing sequential computation creates the basis of the Extended Neural GPU, ByteNet, and ConvS2S, which use convolutional neural networks as basic building blocks*". The goal of NER applied to our considered task is to tag the tokens which are supposed to immediately precede a citation (e.g., GPU, ByteNet, ConvS2S) with the *CITATION* label.

4. The datasets

To produce an accurate citation prediction model, it is advisable to employ a model that has been pre-trained on scientific article data. For this purpose, there are two possible strategies: (1) utilizing an existing transformer that was pre-trained on research papers or (2) pre-train a transformer on an available dataset. The first option could be more convenient and less expensive, but it is also less flexible. In this paper, we have explored both solutions.

For the pre-trained transformer, we opted for *arxiv-nlp*,[2] a GPT-2 transformer that was pre-trained on *arXiv-80MB*, a dataset composed of 80 MB of plain text extracted from arXiv[3] papers about Computation and Language. We adopted *arxiv-nlp* as one of the generative models that we used to tackle citation prediction as a Mask-Filling problem. Since this dataset was not specifically curated for citation prediction, it contained a variety of citation styles, such as square brackets with numbers or brief text, parentheses with citations expressed using the first author's name and publication year, and other formats. This is evident in the behavior of the resulting transformer, which generates citations in many different styles.

In order to solve this issue, we created and made publicly available *s2orc-9K*, a novel dataset for citation prediction in which all citations have been normalized according to a single format. By providing a consistent citation format throughout the dataset, we aim to enhance a transformer's ability to comprehend and learn the underlying concept of citation more effectively.

We generated the *s2orc-9K* dataset from the *s2orc* dataset (Lo, Wang, Neumann, Kinney, & Weld, 2020), a corpus containing about 80 million scientific articles from Semantic Scholar.[4] For each paper, *s2orc* includes several metadata, such as the title of the paper, the relevant venue, the year of publication, the URL on Semantic Scholar, the topic, the abstract, and the paper's sections. In order to produce *s2orc-9K*, we selected 9,000 papers from *s2orc* covering the Computer Science domain according to the *topic* field in the metadata. Subsequently, we standardized all citations by replacing them with the placeholder [ ]. The resulting dataset comprises approximately 43.5 MB of text files.

The *s2orc-9K* dataset has been used both to train an alternative transformer for the generative methodology and the NER approach.

5. The proposed methodologies

In this section, we will describe in detail the two methodologies that we propose for addressing citation prediction, as defined in Section 3. We will refer to them as the generative methodology (formalizing the task as a Mask-Filling problem) and the NER-based methodology.

5.1. Generative methodology

The idea behind the generative methodology is to leverage the text-generation capabilities of transformer models. Currently, the most widely used transformers for this type of task is the GPT family of OpenAI.[5] We adopted the GPT2 transformer (Radford et al., 2019). GPT-2 is a generative language model pre-trained on a set of about 40 GB of unlabeled data. Given an input phrase or word, GPT2 predicts the next token. Specifically, we adopted the version of GPT-2 with 117 million parameters and 12 layers.[6] We chose GPT-2 because it is open and it is available in the most used deep learning libraries such as Transformers.[7]

[2] https://huggingface.co/lysandre/arxiv-nlp
[3] https://arxiv.org/
[4] https://www.semanticscholar.org/
[5] https://platform.openai.com/docs/models
[6] https://huggingface.co/transformers/v2.2.0/pretrained_models.html
[7] https://huggingface.co/docs/transformers/index

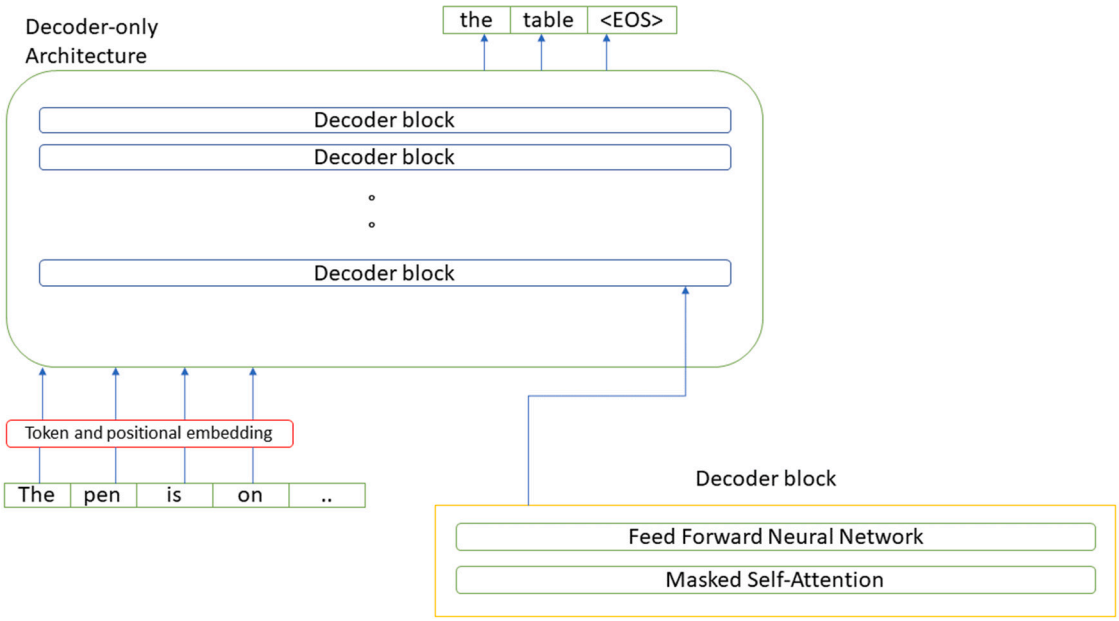

Fig. 1. Decoder-only architecture of GPT-2.

GPT-2 operates on a decoder-only architecture, setting it apart from other transformer models like BERT (Devlin et al., 2019). Indeed, the architecture of transformers generally consists of two main modules: encoder and decoder (Vaswani et al., 2017). A decoder-only architecture is characterized by the absence of the encoder module and has different decoder blocks on top of each other. As shown in Fig. 1, each layer of the transformer consists of a masked self-attention layer and a feed-forward neural network layer.

GPT-2 generates text, token by token, one at a time, so it is possible to check each generated token and identify whether or not it is a citation. Basically, our methodology for citation prediction starts by giving as context the first token and checking whether the predicted token is a citation. If it is, we associate a citation to the first token. Then, we repeat the generation process by feeding the model a context of the first two tokens and so on. Whenever the model predicts a citation, we associate it with the preceding token.

When generating text with a transformer model, each candidate token in the vocabulary is associated with a certain probability, and the one with the highest is generated as the next token. As an example, if we feed GPT2 with the sentence *Marco every day*, it will suggest the token *goes* as it is the token with the highest probability among the words in the vocabulary following the mentioned piece of text. This is known in the literature as *greedy search*[8] as, at each generating step, the transformer chooses the token with the highest probability.

An alternative strategy is known in the literature as *beam search*[8] and considers the conditional probability of the generated tokens. The beam search takes into account a hyperparameter K, known as *beam size*. At time step 1, we identify the K tokens with the highest predicted probabilities. Each selected token will be the start of K candidate output sequences. At the next time step, based on the K candidate output sequences selected at the previous time step, we continue to select K candidate output sequences with the highest predicted probabilities. In the end, the sequence of tokens with the highest conditional probability is chosen.

Figs. 2(a) and 2(b) show an example of greedy search and beam search, respectively. In Fig. 2(a), the input token is *Marco* and we assume that the two tokens with the highest probability to appear after are *is* and *goes*, each with a different probability. The method chooses the token with the highest one, *is*. Then, for the sequence *Marco is*, assuming the two tokens with the highest probability are *tired* and *angry*, the algorithm again chooses *tired*, which has the highest probability.

Conversely, the beam search strategy depicted in Fig. 2(b) computes the highest conditional probability $p(t2|t1)$. In this case, this is 0.2*0.5, which is higher than that of the sequence chosen by the greedy search, 0.4*0.2 (Fig. 2(a)).

The drawbacks of the beam search are related to the higher computational cost needed to compute the conditional probabilities of the sequences and that depends on the beam size. The higher the beam size, the more information the algorithm will have to compute and store.

Inspired by the same idea of the beam search, a contribution that we included in our methodology is the possibility of considering a range of high-probability tokens rather than just the one with the max probability. To this purpose, we take into account the top w_{gen} words sorted in decreasing order for their probabilities. We then detect a citation whenever the relevant token is in this set.

[8] https://d2l.ai/chapter_recurrent-modern/beam-search.html

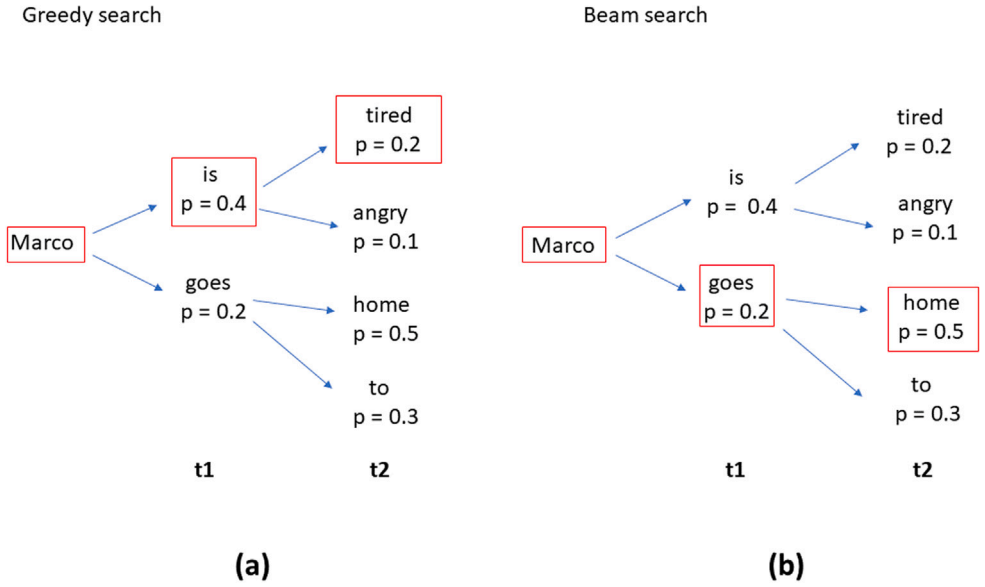

Fig. 2. Example of the greedy and beam search algorithm.

Table 1
Patterns indicating the presence of a citation.

Pattern of citations
[1]
[Author]
(Author, 1991)
(Author1, Author2 and Author 3, 1991)
(Author1, Author2 et al. 1991)
(Author1, Author2 & Author3, 1991)

The rationale for this method rests on the assumption that if a potential citation emerges among the top-ranked probable next words, it indicates a suggestion from the generative approach regarding the potential requirement for a citation. By extending the scope of exploration beyond the token with the highest probability, we enhance the likelihood of capturing relevant citations that might otherwise be overlooked. Clearly, the higher w_{gen}, the more likely that a citation will be predicted as the next token. This parameter also provides a convenient means to regulate the sensitivity of our methodology.

To properly recognize the predicted token as a citation, there are further elements to be considered. In fact, generated citations may manifest in various formats, depending on the transformer utilized for text generation. To address this variability, we adopt a vocabulary of citation formats, enabling the recognition of citations expressed in diverse styles. Table 1 illustrates the patterns currently integrated into our methodology.

We handle this process by using a Finite State Automaton (FSA). For example, if the current predicted token is [, we check the next predicted token. If it is either a number or a proper noun and then the following prediction is the symbol], then a citation is predicted. If, after the initial [symbol, we identify a common noun, then we do not consider it a citation.

As discussed in Section 4, the transformers used in our experiment were trained on two datasets: a set of non-preprocessed arXiv papers within the computational linguistics domain (*arXiv-80MB*) and a smaller dataset of about 43 MB papers (*s2orc-9K*) wherein all citations were replaced with a standardized placeholder. Since *arXiv-80MB* includes citations in various formats (as those indicated in Table 1), the corresponding transformer will generate a diverse array of citation styles. Consequently, there is a possibility that certain citations may adopt too many different styles, leading to potential challenges in their recognition. Conversely, the transformer pre-trained on the novel *s2orc-9K* dataset, introduced in this paper, will produce a single placeholder ([ ]). This facilitates significantly the identification of a citation.

5.2. NER-based methodology

The second methodology we propose to tackle the citation prediction problem is a NER approach that aims to identify the token preceding a citation. To this purpose, we adopted the BERT transformer[9] (Devlin et al., 2019). The pre-training of BERT was

[9] https://huggingface.co/bert-base-cased

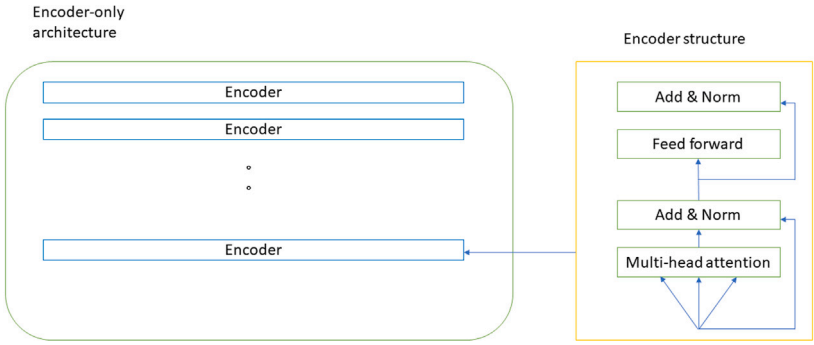

Fig. 3. Encoder-only architecture of BERT.

Fig. 4. Example of Named Entity Recognition for the citation prediction problem. Tokens in blue correspond to the predicted NORMAL type entity meaning that no citations are suggested after them. Conversely, a citation is suggested right after the tokens in orange, corresponding to the predicted CITATION type entity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Example of NER representation.

Token	Label
advanced	NORMAL
NLP	NORMAL
tasks	CITATION

performed on BookCorpus[10] and Wikipedia (English).[11] Compared to GPT-2, which we used for the generative methodology, BERT is characterized by an encoder-only architecture, with each encoding block on top of the other, as depicted in Fig. 3. The model we have used (BERT-base) consists of 12 encoding layers and 110 million parameters. The presence of encoders is what makes BERT suitable for text classification tasks.

We fine-tuned the BERT transformer for the NER task defined in Section 3 using the *s2orc-9K* dataset. To accomplish this, we initially pre-processed the dataset by annotating each token preceding a citation with the CITATION label, while assigning a NORMAL label to all other tokens. Additionally, the citation placeholder was removed. For example, let us consider the annotated sentence "advanced NLP tasks []". In this case, the token [] is eliminated from the sentence, and the preceding word, "tasks" in this instance, is labeled as CITATION as depicted in Table 2.

Fig. 4 showcases an illustrative outcome obtained by employing the fine-tuned model on the input sentence: "The early work CMN achieves this objective using a memory network to store matrix representations of videos". As per the model's predictions, a citation is suggested after the token *CMN* and another citation is recommended after the token *videos*.

5.3. Natural Language Processing heuristics

We defined a set of NLP heuristics to further enhance the performance of the proposed methodologies. These heuristics are systematically applied to the outcomes obtained from both the generative and NER methodologies and are not involved in any phase of the training steps.

The heuristics applied in this study are designed to adhere to standard rules governing the usage of citations. For instance, one heuristic aims to address cases where a predicted citation is situated between a series of uppercased tokens and the corresponding

[10] https://huggingface.co/datasets/bookcorpus
[11] https://huggingface.co/datasets/wikipedia

acronym that refers to them. In such instances, the heuristic removes the predicted citation and places it directly after the acronym, as it is commonly acknowledged that this is the appropriate citation placement. To illustrate this point, let us consider the following sentence: "*One classic technique is Knowledge Distillation (KD), which transfers knowledge from a larger teacher to a smaller student model.*" If any of the methodologies suggest a citation immediately after the term "Knowledge Distillation" and before its corresponding acronym "(KD)", it would be deemed an error and then moved right after the acronym.

We have identified a list of patterns that dictate whether a citation should or should not be present. Subsequently, we devised two sets of heuristics: *Removal* and *Including*. The *Removal* heuristics serve the purpose of eliminating citations generated by our methods when they appear in uncommon or nonsensical positions, such as before an acronym. Conversely, the *Including* heuristics are employed to ensure the inclusion of a citation in positions that conform to identified patterns, such as after an acronym.

In the following, we list all the heuristics implemented in our methodologies:

- The first heuristic handles the aforementioned case where a citation is suggested between an acronym and its defining terms (e.g., *Single-Shot Network Pruning [ ] (SNIP)*. In this case, the suggested citation is removed, and, if not present already, a citation is added at the end of the acronym.
- Words such as *Fig., Table, Section, Tab*, etc. do not typically precede a citation. Therefore, the citation is removed if our method suggests a citation after one of them.
- Usually, a citation is not inserted immediately after a verb. Therefore, we remove citations in this position.
- In a sentence, there can be several *chunks*, parts of a sentence with a noun plus the words describing it (see *noun chunks*[12]). Therefore, if a citation is suggested within a noun chunk, it is removed, and a citation is inserted immediately after the chunk. For example, given the following chunk: *advanced NLP tasks*, if the system predicts a citation after *advanced*, the citation will be moved after the token *tasks*.
- It may happen that consecutive citations are generated on consecutive tokens not separated by commas. In that case, we keep only the last citation of the generated sequence. For example, the sentence *Marco [ ] goes [ ] home [ ]*, becomes *Marco goes home [ ]*.
- The presence of certain expressions such as *previous work, prior studies*, etc., usually indicate the occurrence of a citation where the works refer to. We insert a citation at the end of each sentence containing the mentioned expressions.

6. Evaluation

This section reports the evaluation of the two methodologies introduced in Section 5, along with the datasets for fine-tuning detailed in Section 4 and the heuristics introduced in Section 5.3. In the following, we first describe the generation of the gold standard and discuss the experimental design. We then report the performance of the different techniques in terms of precision, recall, and F1. Next, we present an analysis of the impact of the w_{gen} parameter, the training set size, and citation intents (Roman et al., 2021) on the performance of the best method. We conclude the section by conducting a rigorous statistical analysis to assess the consistency between senior researchers and the computational models.

6.1. Gold standard generation

To construct a gold standard for the citation prediction task, as defined in Section 3, we randomly selected 170 sentences from arXiv papers published in 2023. The relevant papers pertain to the fields of *Artificial Intelligence*[13] and *Computation and Language*.[14]

In order to maintain the integrity of the sentences and avoid influencing human experts during the annotation process, we exclusively selected sentences where the removal of the citation does not alter their meaning or grammatical structure. For instance, if a sentence reads as follows: "Authors in [ ] proposed a method that outperforms the competitors", the sentence becomes incomplete once the citation is removed.

For the annotation process, we enlisted three senior researchers possessing extensive expertise in Artificial Intelligence, Natural Language Processing, and Machine Learning, with an average experience of approximately 15 years in their respective fields. For each sentence, they were asked to indicate whether a citation was warranted and, if so, to select the token immediately preceding the citation. To facilitate this annotation process, the annotators employed Doccano (Nakayama, Kubo, Kamura, Taniguchi, & Liang, 2018), an open-source text annotation tool widely used for this purpose in academic research. The inter-annotator agreement among the three senior researchers was 79.7%.

We adopted a majority voting strategy to create the gold standard. We only considered sentences that received consistent annotations from at least two annotators. Hence, we excluded 37 sentences that received three distinct annotations. Such cases involve scenarios where individual annotators marked three different tokens or instances where one annotator considered the citation unnecessary while the other two identified different tokens. Consequently, the resulting gold standard includes 133 sentences: 69 sentences with one citation and 64 sentences without citation.

[12] https://spacy.io/usage/linguistic-features
[13] https://arxiv.org/list/cs.AI/recent
[14] https://arxiv.org/list/cs.CL/recent

Table 3
Precision, recall, and F1-score of all methods for the coarse-grained formulation of the citation prediction task (binary classification).

	Precision	Recall	F1-score
Scientist 1	100%	60.8%	75.6%
Scientist 2	100%	44.9%	62.0%
Scientist 3	76.2%	88.4%	81.8%
NER-s2orc	95%	27.5%	42.6%
NER-s2orc-H	90.9%	43.4%	58.8%
GM-arXiv	81.8%	65.2%	72.5%
GM-arXiv-H	87.2%	69.5%	77.4%
GM-s2orc	78.2%	78.2%	78.2%
GM-s2orc-H	80.2%	82.6%	81.4%

6.2. Experiments

This section outlines the tested methods and the metrics we adopted for their evaluation. We compared six alternative approaches that combine the previously discussed methodologies with the two datasets introduced in Section 4:

- **NER-s2orc**, the NER-based methodology described in Section 5.2 fine-tuned on *s2orc-9K*.
- **NER-s2orc-H**, NER-s2orc making use of the heuristics described in Section 5.3.
- **GM-arXiv**, the generative methodology described in Section 5.1 pre-trained on *arxiv-nlp*.
- **GM-arXiv-H**, GM-arXiv making use of the heuristics.
- **GM-s2orc**, the generative methodology pre-trained on *s2orc-9K*.
- **GM-s2orc-H**, GM-s2orc making use of the heuristics.

The last four methods were executed with window w_{gen} of size 5, as this value produced the best outcome in preliminary experiments. All methods were set up to predict a maximum of one citation per sentence.

Moreover, to establish a comprehensive comparison with typical human researchers, we extended our investigation by engaging three additional computer scientists. These researchers were assigned the same annotation tasks as the computational methods. As a result, we derived three other baselines, hereafter referred to as **Scientist 1**, **Scientist 2**, and **Scientist 3**.

We evaluated the methods and the human experts on the two tasks defined in Section 3. The first is a binary classification task, where the objective is to determine whether a given sentence necessitates a citation or not. The second task requires to specify the precise position where the citation should be placed.

The performances of the methods and human participants were assessed using Precision, Recall, and F1-score. For the coarse-grained formulation of our task (binary classification), we have adopted the standard definitions of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In particular, a TP/FP is when the method correctly/wrongly predicted the need for a citation for a sentence. A TN/FN is when the method correctly/wrongly predicted no need for a citation for a sentence.

For the fine-grained formulation of our task (citation prediction), we have reformulated them as follows:

- TP: number of tokens immediately preceding a citation which have been correctly predicted;
- TN: number of tokens correctly predicted as tokens which do not precede a citation;
- FN: number of tokens wrongly predicted as tokens which do not immediately precede a citation;
- FP: number of tokens immediately preceding a citation which have been wrongly predicted;

For example, let us consider the following sentence: “Recent years have witnessed spectacular developments in video action recognition [ ]”. Assume that a method predicts the following: “Recent years have witnessed spectacular developments [ ] in video action recognition”, therefore with the token “developments” preceding the citation. The number of TP, TN, FP, and FN will be, respectively, 0, 8 (because there are 8 tokens not preceding a citation correctly predicted), 1, and 1. Given the high number of TN, we will employ the F1-score in the experimental evaluation as it focuses on the other three quantities.

6.3. Results

Table 3 reports the performance of the six methods and the three researchers on the binary classification task.

The three researchers achieve an average F1-score of 73.1%, indicating the demanding nature of the benchmark, which proves challenging even for experienced computer scientists. This is mostly due to the fact that different researchers may have different opinions on which sentence deserves a citation.

Remarkably, GM-s2orc-H outperforms all the other methods and even surpasses human annotators, achieving an impressive F1-score of 81.4%. Notably, it excels in terms of recall (82.6%) when compared to human performance (averaging 64.7%), but it achieves a lower precision (80.2% versus 92.0%). Following GM-s2orc-H, the two most effective methods are GM-s2orc and GM-arXiv-H, which yield comparable performance.

The generative methods fine-tuned on the novel *s2orc-9K* dataset, introduced in this paper, demonstrate superior performance compared to the method employing *arXiv-nlp*. The improvement is particularly remarkable considering that *s2orc-9K* is about half

Table 4
Precision, recall, and F1-score of all methods for the fine-grained formulation of the citation prediction task (citation location).

	Precision	Recall	F1-score
Scientist 1	64.2%	39.1%	48.6%
Scientist 2	67.7%	30.4%	42.0%
Scientist 3	57.5%	66.6%	61.7%
NER-s2orc	54.5%	17.3%	26.3%
NER-s2orc-H	51.5%	24.6%	33.3%
GM-arXiv	36.3%	28.9%	32.3%
GM-arXiv-H	50.9%	40.5%	45.1%
GM-s2orc	44.9%	44.9%	44.9%
GM-s2orc-H	49.2%	50.7%	50.0%

Table 5
Precision, recall, and F1-score of GM-s2orc-H for different values of w_{gen} for the coarse-grained classification of the citation prediction task (binary formulation).

	Precision	Recall	F1-score
GM-s2orc-H-wgen1	92.0%	33.3%	48.9%
GM-s2orc-H-wgen3	80.3%	59.4%	68.3%
GM-s2orc-H-wgen5	80.2%	82.6%	81.4%
GM-s2orc-H-wgen7	73.2%	91.3%	81.2%
GM-s2orc-H-wgen9	68.6%	98.5%	80.9%

Table 6
Precision, recall, and F1-score of GM-s2orc-H for different values of w_{gen} for the fine-grained formulation of the citation prediction task (citation location).

	Precision	Recall	F1-score
GM-s2orc-H-wgen1	52.0%	18.8%	27.6%
GM-s2orc-H-wgen3	49.0%	36.3%	41.6%
GM-s2orc-H-wgen5	49.2%	50.7%	50.0%
GM-s2orc-H-wgen7	40.5%	50.7%	45.1%
GM-s2orc-H-wgen9	39.3%	56.5%	46.4%

the size of *arXiv-nlp* (43.0 MB vs 80 M). As discussed in Section 4, the main difference between the two datasets is that in *s2orc-9K* all references were substituted with a placeholder ([]). In contrast, references within *arXiv-nlp* manifest in a variety of styles, reflecting the conventional heterogeneity observed in preprint papers on arXiv. We thus postulate that the superior performance is attributable to the model's exposure to a consistent representation of a citation, enabling it to more effectively discern where a citation should be positioned. The heuristics introduced in Section 5.3 also proved to be highly effective, yielding an average improvement of 8.1% of the F1-score for the three methods that incorporated them.

Table 4 reports the performance of methods and researchers on the second task, i.e., indicating the specific placement of a citation. GM-s2orc-H again outperforms all other methods, obtaining a 50.0% F1-score. It also achieves comparable results to the three researchers, which yielded an average F1-score of 50.7%.

The result further confirmed the utility of both the *s2orc-9K* dataset and the heuristics. The generative methods fine-tuned on the novel *s2orc-9K* dataset (GM-s2orc, GM-s2orc-H) exhibit superior performance compared to their counterparts using *arXiv-nlp* (GM-arXiv, GM-arXiv-H), with an average F1 improvement of 8.75%. The three methods incorporating the heuristics (NER-s2orc-H, GM-arXiv-H, GM-s2orc-H) achieve an average F1 improvement of 8.3% compared to their respective versions without the heuristics (NER-s2orc, GM-arXiv, GM-s2orc).

6.4. Further analysis on GM-s2orc-H

In this section, we further analyze GM-s2orc-H, which has been identified as the most effective method. Specifically, we investigate the impact of the w_{gen} window, the size of the training dataset, and the citation intents on its performance.

The parameter w_{gen} dictates the number of words considered at each iteration for citation detection, as described in Section 5.1. We evaluated GM-s2orc-H with varying values of w_{gen}, using the same evaluation settings of the previous section. The outcomes are presented in Table 5 for the coarse-grained task and Table 6 for the fine-grained task. An increase in the value of w_{gen} enhances Recall while diminishing Precision. This suggests that varying the w_{gen} allows for achieving diverse trade-offs between Recall and Precision, contingent upon the specific requirements of a given application. For both the coarse-grained and fine-grained tasks, a w_{gen} of 5 appears to produce the best F1 score.

The size of the training set is expected to have a significant effect on the performance of the method. We thus evaluated various iterations of GM-s2orc-H trained on different portions of *s2orc-9K* (5 MB, 11 MB, 21 MB, and the complete 43 MB). Table 7 and Table 8 report the outcomes for the coarse-grained and the fine-grained tasks, respectively. As anticipated, the F1 score exhibits a

Table 7
Precision, recall, and F1-score of GM-s2orc-H with different training set sizes for the coarse-grained formulation of the citation prediction task (binary classification).

	Precision	Recall	F1-score
GM-s2orc-H-5mb	73.9%	73.9.8%	73.9%
GM-s2orc-H-11mb	75.7%	72.4%	74.0%
GM-s2orc-H-21mb	71.6%	76.8%	74.1%
GM-s2orc-H-43mb	80.2%	82.6%	81.4%

Table 8
Precision, recall, and F1-score of GM-s2orc-H with different training set sizes for the fine-grained formulation of the citation prediction task (citation location).

	Precision	Recall	F1-score
GM-s2orc-H-5mb	37.6%	37.6%	37.6%
GM-s2orc-H-11mb	45.4%	43.4%	44.4%
GM-s2orc-H-21mb	43.2%	46.3%	44.7%
GM-s2orc-H-43mb	49.2%	50.7%	50.0%

Table 9
Consistency of the three main methods with the senior researchers. Bold entries highlight results that are *not* statistically significant, suggesting alignment with the researchers.

	Task	p
GM-s2orc-H	coarse-grained	**0.51**
GM-arXiv-H	coarse-grained	0.002
NER-s2orc-H	coarse-grained	0.0001
GM-s2orc-H	fine-grained	**0.58**
GM-arXiv-H	fine-grained	0.02
NER-s2orc-H	fine-grained	0.05

monotonic increase with the expansion of the training set for both tasks. This observation suggests that leveraging a more extensive dataset could further enhance the performance.

The academic literature has also examined the reasons behind specific citations, as they can provide further insights (Peroni & Shotton, 2012). This has resulted in several systems and schemes that aim to categorize citations (Roman et al., 2021), such as Background, Motivation, Future Work, and more. A comprehensive study of the effects of citation intent on citation prediction is well beyond the scope of this paper. However, we carried out a preliminary study on the capability of GM-s2orc-H to identify citations with different intents. To this purpose, we run MultiCite[15] (Lauscher et al., 2021), a citation intent classification tool, on our predefined gold standard. It identified 52 citations as 'Background', 13 as 'Uses', 3 as 'Motivation', and 1 as 'Difference'.

We focused on the first two categories, as there were too few instances of the other categories to draw reliable conclusions. In the coarse-grained evaluation, GM-s2orc-H accurately identified 80.7% (42 out of 52) of the citations labeled as 'Background' and 84.6% (11 out of 13) labeled as 'Uses'. When considering the fine-grained task, GM-s2orc-H correctly pinpointed 48.0% (25 out of 52) of the 'Background' citations and 69.2% (9 out of 13) of the 'Uses' citations. This implies that GM-s2orc-H might perform more effectively when the citation's purpose is to utilize previous work, rather than merely mentioning it (e.g., in a literature review). This might occur because citations are introduced more explicitly in the first case. Nonetheless, a more exhaustive exploration of this aspect is planned for future research.

6.5. Consistency analysis

In order to further verify the previous results, we conducted a statistical analysis to assess the consistency between the three senior researchers described in Section 6.1 and the best computational methods. Unlike the previous section, the purpose is not to measure which method performs better with respect to a gold standard. Instead, we want to determine whether the computational methods *behaved consistently with the senior researchers*. Specifically, we evaluated the consistency between the senior researchers and three main methods: GM-s2orc-H, GM-arXiv-H, and NER-s2orc-H.

The main results are summarized in Table 9. For the first task, we used the non-parametric Cochran Q-test (Cochran, 1950) to evaluate the consistency of performance among a set of annotators (either humans or computational methods). The Cochran Q-test is used to assess whether a group of annotators consistently provided binary assessments (i.e., either 1 or 0) for a given set of items. It is well-suited for our case, considering the binary nature of this task. For all tests, a p-value ≤ 0.05 was considered significant, following standard practice. Therefore, when the resulting p-value is below this threshold, it denotes that the test

[15] MultiCite - (https://github.com/allenai/multicite).

detected a statistically significant difference among the annotators. Conversely, if the p-value exceeds this threshold, it indicates that the annotators performed consistently.

As expected, the three senior researchers exhibit similar behaviors ($p = 0.26$). Notably, GM-s2orc-H results are also not statistically different from those of the three senior users: the group including GM-s2orc-H and the three researchers yields $p = 0.51$. To validate this finding, we also conducted individual comparisons between GM-s2orc-H and each of the three researchers. For this purpose, the Cochran Q-test was once again applied to each pair. In all instances, the observed differences between GM-s2orc-H and the senior researcher were found to be not statistically significant ($p = 0.32$, $p = 0.72$, $p = 0.75$). Conversely, both GM-arXiv-H and NER-s2orc-H are statistically different from the three senior researchers, yielding $p = 0.002$ and $p = 0.0001$, respectively.

For the second task, we employed the chi-square test for an RxC contingency table (Sheskin, 2003), as the data lacked binary features. In this case, we utilized the number of agreements about and disagreements between two annotators regarding the positions of the citations. As before, a p-value ≤ 0.05 denotes a statistically significant difference. Otherwise, it indicates a consistency among the tested elements. The three senior researchers exhibited a high level of consistency ($p = 0.28$) also for this task. Remarkably, GM-s2orc-H exhibits no statistically significant difference from the researchers ($p = 0.58$). Conversely, GM-arXiv-H and NER-s2orc-H are not consistent with the three senior researchers.

In conclusion, the results of the statistical analysis validate and strengthen the experimental findings presented in Section 6.4. Specifically, the generative method trained on *s2orc-9K* (GM-s2orc-H) exhibits performance comparable to that of experienced researchers in two key tasks: determining whether a sentence requires a citation and identifying the most appropriate position.

7. Conclusions

Statements in scientific articles may describe new research methodology, datasets, and findings, or they make reference to prior research and conceptual knowledge that served as the foundation for the current discoveries. For the latter, researchers refer to other papers citing them in their claims.

In this paper, we presented two types of methodologies to tackle the task of determining whether a sentence from a scientific paper needs a citation and identifying the most suitable placement. Specifically, we reframed this challenging task as either a Mask-Filling problem or as a Named Entity Recognition problem. In the first case, we proposed an approach leveraging the text-generative capability of the GPT-2 Transformer. In the second case, we developed a methodology based on fine-tuning a BERT Transformer on a relevant NER task. On top of the results of the proposed methods, we applied a set of heuristics for further adjusting the citation placement.

We evaluated the suggested methodologies and heuristics using a gold standard of manually annotated sentences. Notably, the generative approach, fine-tuned on the *2orc-9K* dataset, outperformed the other methods. Particularly significant is that this approach yielded results comparable to those achieved by human researchers. To reinforce the validity of these findings, we conducted a thorough statistical analysis to assess the consistency between senior researchers and the computational models. The outcomes of our analysis provide compelling evidence that the proposed solution exhibits a level of performance that is statistically indistinguishable from that of senior researchers.

In future works, we aim to expand our proposed methodology to also address the task of predicting the specific citation to be inserted. Variations of this problem could be formulated based on the reference options available, whether from a predefined list or sourced online. We also plan to take advantage of the surrounding context of a sentence as well as the citation intent (e.g., *Background*, *Uses*) to further enhance the performance. Finally, we intend to investigate the application of large-scale knowledge graphs within the scholarly domain, such as CS-KG (Dessì, Osborne, Recupero, Buscaldi, & Motta, 2022) and ORKG (Jaradeh, Oelen, Prinz, Stocker, & Auer, 2019), to facilitate the selection of the most suitable reference.

CRediT authorship contribution statement

Davide Buscaldi: Conceived and designed the analysis, Wrote the paper. **Danilo Dessí:** Collected the data, Contributed data or analysis tools, Wrote the paper. **Enrico Motta:** Conceived and designed the analysis, Supervised the work. **Marco Murgia:** Contributed data or analysis tools, Performed the analysis, Wrote the paper. **Francesco Osborne:** Performed the analysis, Wrote the paper, Supervised the work. **Diego Reforgiato Recupero:** Collected the data, Contributed data or analysis tools, Wrote the paper, Supervised the work.

Data availability

I have attached the data in the repository linked in the paper.

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No. 3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

References

Berrebbi, D., Huynh, N., & Balalau, O. (2022). *GraphCite: Citation intent classification in scientific publications via graph embeddings* (pp. 779–783). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3487553.3524657.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., et al. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the sixth international conference on language resources and evaluation*. Marrakech, Morocco.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73. http://dx.doi.org/10.1016/j.joi.2017.11.005, URL https://www.sciencedirect.com/science/article/pii/S1751157717303516.

Cerovšek, T., & Mikoš, M. (2014). A comparative study of cross-domain research output and citations: Research impact cubes and binary citation frequencies. *Journal of Informetrics*, *8*(1), 147–161. http://dx.doi.org/10.1016/j.joi.2013.11.004, URL https://www.sciencedirect.com/science/article/pii/S1751157713001053.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, *37*(3/4), 256–266.

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 3586–3596). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1361, URL https://aclanthology.org/N19-1361.

Dessì, D., Osborne, F., Recupero, D. R., Buscaldi, D., & Motta, E. (2022). CS-KG: A large-scale knowledge graph of research entities and claims in computer science. In U. Sattler, A. Hogan, C. M. Keet, V. Presutti, J. ao Paulo A. Almeida, H. Takeda, P. Monnin, G. Pirrò, & C. d'Amato (Eds.), *Lecture Notes in Computer Science*: *vol.13489*, *The semantic web - iswc 2022 - 21st international semantic web conference* (pp. 678–696). Springer, http://dx.doi.org/10.1007/978-3-031-19433-7_39.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1423.

Gehrke, J., Ginsparg, P., & Kleinberg, J. (2003). Overview of the 2003 KDD cup. *Acm Sigkdd Explorations Newsletter*, *5*(2), 149–151.

Gosangi, R., Arora, R., Gheisarieha, M., Mahata, D., & Zhang, H. (2021). On the use of context for predicting citation worthiness of sentences in scholarly articles. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 4539–4545). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.359, URL https://aclanthology.org/2021.naacl-main.359.

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th international conference on world wide web* (pp. 421–430).

Jain, S. M. (2022). *Introduction to transformers for NLP*. Berkeley, CA: Apress Berkeley.

Jaradeh, M. Y., Oelen, A., Prinz, M., Stocker, M., & Auer, S. (2019). Open research knowledge graph: A system walkthrough. In *Digital libraries for open knowledge: 23rd international conference on theory and practice of digital libraries* (pp. 348–351). Springer.

Lauscher, A., Glavaš, G., Ponzetto, S. P., & Eckert, K. (2017). Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *WOSP 2017*, *Proceedings of the 6th international workshop on mining scientific publications* (pp. 24–28). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3127526.3127531.

Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Jurgens, D., Cohan, A., et al. (2021). MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. arXiv:2107.00414.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132. http://dx.doi.org/10.1016/j.aiopen.2022.10.001, URL https://www.sciencedirect.com/science/article/pii/S2666651022000146.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., et al. (2023). Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. arXiv:2304.01852.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4969–4983). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.447, Online URL https://www.aclweb.org/anthology/2020.acl-main.447.

Mollá, D., van Zaanen, M., & Smith, D. (2006). Named entity recognition for question answering. In *Proceedings of the australasian language technology workshop 2006* (pp. 51–58). Sydney, Australia: URL https://aclanthology.org/U06-1009.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text annotation tool for human. URL https://github.com/doccano/doccano, Software available from https://github.com/doccano/doccano.

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, *17*, 33–43.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Roman, M., Shahid, A., Khan, S., Koubaa, A., & Yu, L. (2021). Citation intent classification using word embedding. *Ieee Access*, *9*, 9982–9995.

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. Chapman and hall/CRC.

Tang, J., & Zhang, J. (2009). A discriminative approach to topic-based citation recommendation. In *Proceedings of the 13th pacific-asia conference on advances in knowledge discovery and data mining* (pp. 572–579). Berlin, Heidelberg: Springer-Verlag, http://dx.doi.org/10.1007/978-3-642-01307-2_55.

Vajdecka, P., Callegari, E., Xhura, D., & Ásmundsson, A. (2023). Predicting the presence of inline citations in academic text using binary classification. In *Proceedings of the 24th nordic conference on computational linguistics* (pp. 717–722). Tórshavn, Faroe Islands: University of Tartu Library, URL https://aclanthology.org/2023.nodalida-1.72.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in neural information processing systems. Vol. 30*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yu, D., & Yan, Z. (2023). Main path analysis considering citation structure and content: Case studies in different domains. *Journal of Informetrics*, *17*(1), Article 101381. http://dx.doi.org/10.1016/j.joi.2023.101381, URL https://www.sciencedirect.com/science/article/pii/S1751157723000068.