FINECite: A Novel Framework for Fine-grained Citation Context Extraction

Anonymous submission

Abstract

Citations are a cornerstone of scientific discourse and academia, situating novel contributions in the context of existing research. As links that connect and contextualize publications, they have been widely analyzed to understand the content or impact of a publication and its relation to the broader research field. The particular context of a citation, however, is often overlooked. This is critical as the citation link alone does not convey the nuanced information on *what is cited*, *how it is cited*, and *why*. The few existing attempts considering citation contexts are limited in multiple aspects, for instance, in that they artificially restrict context windows to match sentence boundaries. To address these concerns, we develop a novel three-facet framework, termed FINECITE, for fine-grained citation context analysis (CCA). This framework particularly considers the intricate structure of scientific literature and the requirements of downstream applications. We realize this framework by constructing a novel corpus containing 1,056 manually annotated fine-grained citation contexts. Next, we establish baseline models for two important CCA applications: citation context extraction and citation classification. Importantly, our experiments demonstrate the positive impact of our finer-grained context definition, which leads to an increase in performance on both tasks and improvement on previous approaches.

1 Introduction

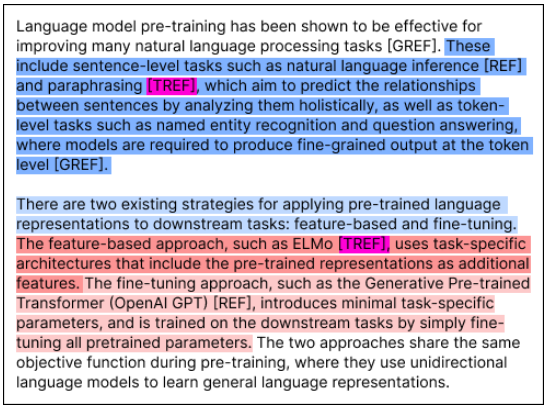
Citations play an elementary role in science, serving as a fundamental mechanism that directs researchers to relevant prior work and provides evidence supporting claims and theories mentioned in publications (Frost 1979). Given their importance to science, a substantial body of literature has been dedicated to the description and analysis of citations (Jurgens et al. 2018; Abu-Jbara, Ezra, and Radev 2013; Teufel, Siddharthan, and Tidhar 2006). In computational linguistics, citation links are used in many ways, including citation-based summarization (Cohan and Goharian 2015), recommendation (Bai et al. 2019), or the embedding of whole corpora (Cohan et al. 2020; Ostendorff et al. 2022).

In most of these cases, the particular context of a citation is not considered. This is critical as the citation link alone does not convey the nuanced information on what is cited, how it is cited, and why. To address this issue, considerable literature explores the role and purpose of citations, also known as *citation context analysis* (CCA) (Lauscher et al. 2022; Swales 1986). Most of the research focuses on assigning citations to a specific class considering paradigms like function (Lauscher et al. 2022; Jurgens et al. 2018; Teufel, Siddharthan, and Tidhar 2006), purpose (Pride and Knoth 2020; Abu-Jbara, Ezra, and Radev 2013), sentiment (Athar and Teufel 2012), or intent (Cohan et al. 2019). For the classification, a span of text is considered, largely limiting the citation context artificially. Cohan et al. (2019), for instance, only considered the citing sentence, arguing that multi-sentence context would add noise. However, *little* attention has been given to determining the exact span of text containing a citation's relevant attributes. Existing studies still fail to supply a comprehensive definition preserving the fine-grained semantics and structure of scientific texts.

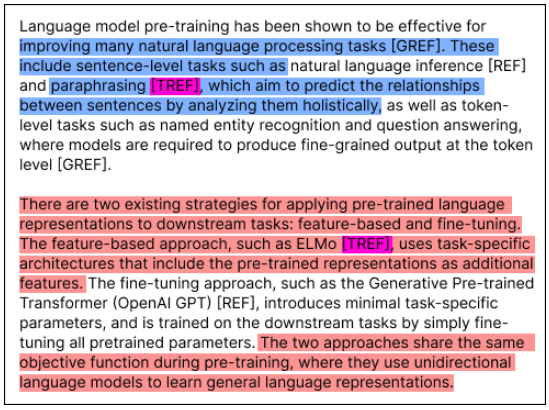
To bridge this gap, this paper introduces a new framework, called FINECITE, for fined-grained citation context analysis. More specifically, we introduce the notion of '*fine-grained citation context*,' provide our corpus painstakingly built for implementing the notion, and present baseline citation context analysis models trained and tested on the corpus. Our FINECITE framework fully captures the semantic structure of scientific literature and precisely represents the context information for downstream applications, such as citation context extraction and classification. In particular, the proposed FINECITE fully embodies three important facets—*information*, *perception*, and *background*—that have not yet been explored well enough in citation context research. In addition, our key idea to building the corpus is to reflect three structural characteristics—*sub-sentence segmentation*, *non-contiguity*, and *dynamic context window*—in citations. Our experimental results indicate that both structural characteristics and information facets have a critical role in improved understanding of the citation-link.

The following are the contributions of this paper:

- We propose a novel framework, termed FINECITE, addressing the lag of structural and semantical appropriate definition for fine-grained CCA.
- We create and publish the FINECITE corpus, consisting of 1,056 manually annotated fine-grained citation contexts following the proposed FINECITE framework.
- We establish a baseline for two important CCA tasks, revealing the advantage of the FINECITE framework compared to previous context extraction approaches.



(a) Fixed-sized approach: one-sentence (dark) and two-sentence (dark + light)



(b) Fine-grained window: sub-sentence boundaries, dynamic size, and non-contiguity

Figure 1: Example citation contexts: (a) existing fixed-sized common context, (b) proposed fine-grained context

We will make our artifacts publicly available for reproducibility and expansion of our work upon acceptance.

The rest of the paper is organized in the following. The subsequent section reviews the relevant literature in the field of citation context extraction. Section 3 details terminology and research question. Section 4 proposes our framework for citation context analysis. Section 5 describes the curation process of the FINECITE corpus. In Section 6, we evaluate the framework for two important CCA tasks (citation extraction and classification). Lastly, Section 7 concludes the paper with a summary and suggests future research directions.

2 Related Work

While citations primarily contextualize a publication into the broader research literature (Frost 1979), they have been widely recognized for their properties for assessing (Garfield 1955; Hirsch 2005), summarizing Cohan and Goharian 2015, or organizing Cohan et al. 2020; Ostendorff et al. 2022 scholarly work. In many cases, however, the particular context of a citation is not considered, neglecting the nuanced information of the relation between two papers.

Existing work generally focuses on the classification of a citation-link by assigning function (Lauscher et al. 2022; Jurgens et al. 2018; Teufel, Siddharthan, and Tidhar 2006), purpose (Pride and Knoth 2020; Abu-Jbara, Ezra, and Radev 2013), sentiment (Athar and Teufel 2012), or intent labels (Cohan et al. 2019) using a varying span of text surrounding the citation marker. However, in most cases, the considered citation context window is artificially constrained by at least one of the three common fallacies of CCA: fixed-sized context windows, context contiguity, and sentence segmentation. Table 1 shows a structured comparison of the relevant research regarding fine-grained CCA.

Fixed-size context windows are prevalent. A prevalent practice in the literature is to consider the sentence with the citation marker as an adequate representation of the citation context (Pride and Knoth 2020; Cohan et al. 2019). Other fixed-sized approaches include a specific number of characters (Jurgens et al. 2018), or whole paragraphs (Teufel, Siddharthan, and Tidhar 2006). All of these approaches constrain the citation context as the relevant context information is usually not distributed evenly around the citation marker throughout a paper, let alone a whole corpus. Early approaches introducing dynamic context extraction to CCA investigated hybrid approaches, where a dynamic number of sentences is chosen from a fixed text window (Abu-Jbara, Ezra, and Radev 2013; Athar and Teufel 2012). Recently, some researchers further explored the benefits of fully dynamic citation contexts (Lauscher et al. 2022; Nambanoor Kunnath, Pride, and Knoth 2022). More particularly, Nambanoor Kunnath, Pride, and Knoth (2022) investigated context extraction based on embedding similarities between sentences surrounding the citation marker and metadata from the cited paper. Lauscher et al. (2022) instead annotated a large-scale, multi-sentence, multi-label, and multi-mention citation corpus based on 7 function classes and conducted several experiments on that corpus. Both works find significant utility in the dynamically sized citation contexts.

Occasional investigation of non-contiguity. Non-contiguity is a far less researched property of CCA. Even though a notable number of publications technically allow for the extraction of non-contiguous contexts (Lauscher et al. 2022; Abu-Jbara, Ezra, and Radev 2013; Athar and Teufel 2012)), only one work investigated the phenomenon in particular (Nambanoor Kunnath, Pride, and Knoth 2022). They find that non-contiguity slightly outperforms in direct comparison to a contiguous extraction approach. This seems obvious as context information is often not a contiguous span of texts stretching from the citation marker and can be separated by non-context passages (see Figure 1).

Sub-sentence segments are neglected. Sentence boundaries are predominantly accepted as the atomic unit of information in citation contexts (Cohan et al. 2019; Nambanoor Kunnath, Pride, and Knoth 2022; Lauscher et al. 2022). This, however, is not necessarily the case. Abu-Jbara and Radev (2012), for instance, shows that sentences with multiple citations consist of multiple sub-sentence context fragments. We observe this phenomenon also in contexts exceeding the citing sentence (see Figure 1).

Author/Year	Context	Conceptualized on	Dynamic	Non-Cont	Sub-Sent
Lauscher et al. (2022)	variable no. of sent.	function	✔	✔	✘
Kunnath et al. (2022)	variable no. of sent.	topic similarity	✔	✔	✘
Pride and Knoth (2020)	single sentence	purpose, influence	✘	✘	✘
Cohan et al. (2019)	single sentence	intent	✘	✘	✘
Jurgens et al. (2018)	300 characters	function	✘	✘	✘
Abu-Jbara et al. (2013)	variable in 4 sent.	purpose, polarity	(✔)	✔	✘
Athar and Teufel (2012)	variable in $\pm$ 4 sent.	sentiment	(✔)	✔	✘
Abu-Jbara and Radev (2012)	words in single sent.	affiliation	(✔)	(✔)	✔
Teufel et al. (2006)	paragraph	function	✘	✘	✘
FINECITE (this work)	variable no. of words	affiliation, information	✔	✔	✔

Table 1: Existing work in CCA compared to FINECITE (this work), based on context type (**Context**), dynamic-context size (**Dynamic**), non-contiguity (**Non-Con**), and sub-sentence segmentation (**Sub-Sent**).

No self-contained definition. The citation context definitions exhibited by the previous work in CCA are mostly based on schemata built for classifying citations, like function classification (Lauscher et al. 2022; Jurgens et al. 2018; Teufel, Siddharthan, and Tidhar 2006), or purpose calssification (Pride and Knoth 2020; Abu-Jbara, Ezra, and Radev 2013). This imposes a conceptualized interpretation of what essentially is an information extraction task. This is problematic, as the resulting context is not a general representation of the context information but a task or schema-specific rendering of it.

3 Preliminaries

We first introduce some of the terminology used in CCA and present our research questions.

3.1 Terminology

The term '*citation context*' refers to the span of text containing relevant information about a given citation. The citation is marked by a citation marker, which in the case of computer science either comes in an author-date (e.g., Mustermann, 20XX) or sequential numbers in square brackets format (e.g., [1][2]).

To reduce redundant information, we adopt a unified schema to represent single or groups of references proposed by Abu-Jbara and Radev (2012). The four tokens ([REF][GREF][TREF][GTREF]) are used to replace a single reference marker (e.g., Musterfrau, 20XX $\rightarrow$ [REF]), group references (e.g., Musterfrau, 20XX, Mustermann, 20XX $\rightarrow$ [GREF]), or indicate when one is chosen as the target for analysis or annotation by adding a 'T' (e.g., [TREF], [GTREF]). This representation reduces the citation marker to its essential properties and makes it easier digestible for model training, as it can be represented in one token.

Other terms often used in CCA are '*citance*' (Nakov, Schwartz, and Hearst 2004) and 'explicit or implicit citation' (Abu-Jbara, Ezra, and Radev 2013). The citance and explicit citation are two names for the sentence in which the citation is located. Implicit citation refers to context sentences, which are not part of the citance. In this paper, we won't use the terms explicit or implicit citation, as they inherently represent sentence-segmented context extraction.

3.2 Research Questions

The term citation context remains poorly defined, as there is no clear description of what information should be represented by the citation context and which semantic structures of scientific texts must be considered. The two questions motivating this research are:

What are the intrinsic pieces of information that a citation context should contain? Previous research defines the citation scope as context facilitating a specific outcome or interpretation. Thus, the annotated context in those cases does not depict a universal representation of the citation-related information. Instead, it requires an unspecific, general representation of the outcome, where the majority of information needed for downstream tasks is represented.

Which aspects of the semantic structure of scientific literature must be respected for context extraction? As mentioned in Section 2, many of the commonly used presuppositions on context distributions do not represent the semantic and structural reality of scientific texts. Thus, it demands an investigation into structural factors like dynamicity, contiguity, and segment granularity and how each of those contributes to optimal context representation.

How do we structure the context in a way to provide conceptual simplicity while offering maximum utility for downstream use? Citation context have different types of information. For simple conceptualization and tailored use in downstream tasks, the information should be separated into its major information types.

These research questions accompany three tasks we will address in the following three sections: **T1)** The conceptualization of a framework respecting the prerequisite for information-rich context representation (Section 4), **T2)** The

application of the novel framework for dataset construction (Section 5), and **T3)** The evaluation of the novel framework on common CCA application (Section 6).

4 FINECITE: Context Framework

There are four areas of concern regarding the citation context conceptualization: The framework must (*i*) be conceptualized around the different scopes of information that are relevant for the citation context, (*ii*) respect the semantic structure of scientific literature, (*iii*) be capable of representing the context in a shape that is useful for downstream applications, and (*iv*) should be easy to understand and annotate.

With this in mind, we propose our novel framework, termed FINECITE, effectively covering three citation context facets. The three scopes represent the information of *what is cited*, *how it is cited*, and *why*, respectively. The separation is motivated by the intent to distinguish between information from the cited paper, the perception and use of the cited information, and the background indicating the role the citation plays in the larger concept of the paper. Further, it satisfies our ambitions to create a conceptually approachable definition. Each scope, namely *information scope*, *perception scope*, and *background scope*, is explained briefly in the following paragraphs. A more detailed description of the three citation scopes, including annotated sample paragraphs, is shown in Appendices A and B.

The Information Scope represents the details the citing author is directly extracting from the referenced paper. The information could be findings, practices, opinions contained in the cited paper, or objective facts about its content.

The Perception Scope relates to the author's subjective perception and use of the information from the cited document. Perception includes statements about capability, comparisons, criticism, or any other kind of judgment of the cited information. Use describes the form in which the extracted concepts or methods are applied in the citing paper.

The Background Scope represents information on why the cited paper is mentioned. This includes details about the broader argument, an explanation of an underlying technique, and further elaboration on the topic of the cited paper. As background is a concept that occurs in different shapes, this scope is the most difficult to conceptualize.

The required structural prerequisites for a fine-grained context extraction were explored in Section 2. We find that context dynamicity, non-contiguity, and sub-sentence segmentation are often overlooked but fundamental parts of a detailed context representation (see Figure 1). Neglecting them would constrain the expressiveness of the context.

5 FINECITE: Corpus

Using our novel framework, the FINECITE corpus is a manually annotated dataset for fine-grained context extraction. With the dataset creation, we aim to (*i*) assess whether the theoretical framework is applicable to scientific texts, (*ii*) investigate the assumption on structure and information distribution against common CCA paradigms, and (*iii*) create a resource for the evaluation of the framework.

5.1 Dataset Construction

The corpus is constructed in the following steps.

Step 1: Procurement. Our dataset is built on the ACL Anthology Network Corpus (Radev, Muthukrishnan, and Qazvinian 2009), which contains over 80K papers from several ACL conferences and other venues in computational linguistics. To access structural information of the documents alongside the surrounding paragraph of each citation marker, we use the GROBID (Lopez, Patrice and Foppiano, Luca and others 2008-2024) parsed full paper dataset published by Rohatgi (2022). We skipped documents containing faulty meta-information, languages other than English, or for scientific literature untypical formats ($>$3 sections, $>$5 references), the latter mostly to avoid poorly parsed documents. After the cleansing, we sampled 72 of the remaining 70,000 Documents, totaling 1,056 paragraphs, each containing one citation marker highlighted as the target citation and a three-scope annotation of the citation context.

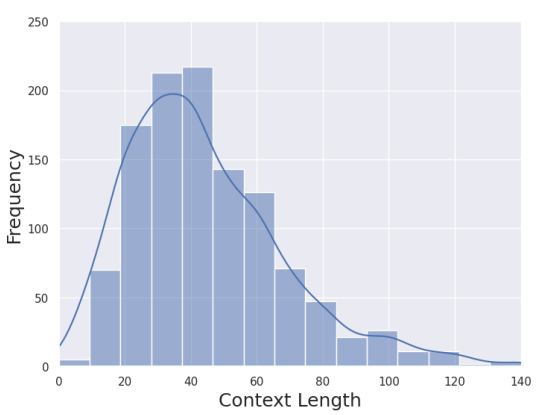
Step 2: Guideline creation. The annotation guidelines comprise best practices and rules on how to annotate a paragraph per the three context scopes described in Section 4. The instructions were created in an iterative process in which several annotators completed five to ten tasks separately. They subsequently discussed differences in annotation and updated the guidelines with a new rule accordingly. This process was repeated with a new batch of tasks until the inter-annotator agreement (IAA) was sufficiently high. More information on the IAA can be found later in this section.

For the sake of simplicity, we assume that citation markers are set correctly, the information attributed to the cited document is a matter of fact from that document, and group references have a sufficiently similar context, making it feasible to annotate them as one context. Furthermore, in case of ambiguity, we prioritize information over perception scope and perception over background scope. The complete Annotation Guidelines can be found in Appendix B.

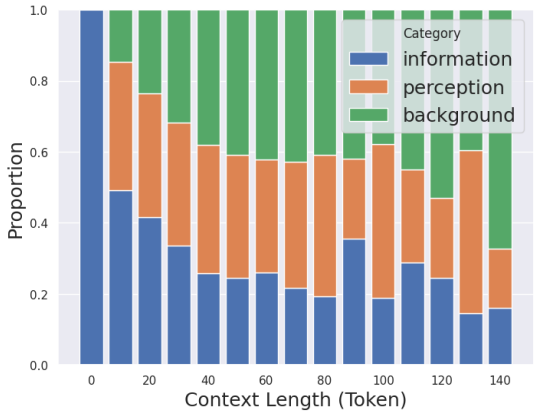
Step 3: Annotation. The annotation was performed on a single paragraph. The annotator was asked to read the paragraph and subsequently highlight the three context scopes of the targeted reference. The annotation took place using a tool specifically designed for this task. The target reference and other non-target references were color-coded to omit confusion, and metadata of the citing as well-cited documents, section titles, and surrounding paragraphs of the same section were provided in a side menu. We encouraged the annotators, however, only to consider the additional information in ambiguous cases, as the context mostly follows semantic structures that do not require further information. More details on the annotation are exhibited in Appendix C.

Five annotators with a background in computer science and medium to high experience in reading English scientific literature annotated the 1,056 instances. None of them, however, are native English speakers or specially trained in English linguistics. All annotators underwent an extensive training process and participated in at least part of the guideline-building process.

Step 4: Validation. The inter-annotator agreement (IAA) is a set of metrics to estimate the quality of the annotation guidelines and resulting dataset. We employ the



(a) Distribution of context length (words).



(b) Label distribution per context length (words).

Figure 2: Results of statistical analysis of the FINECITE dataset, showing the variation of context length and its interrelation with label distribution.

F-measure (Hripcsak and Rothschild 2005), commonly used for evaluating span annotations, and complement it with Cohens κ (Cohen 1960) for the agreement on label assignment above that expected by chance. A short description of how the F-measure is used for IAA is located in Appendix D.

We calculate five different F-scores to capture different aspects of the annotation process separately: Two aggregate measures ($F1$ and $F1_{macro}$) representing the whole annotation and three specific measures ($F1_{inf}$, $F1_{perc}$, and $F1_{back}$) for each citation scope.

The three specific F-scores measure agreement on one distinct scope. More specifically, $F1_{inf}$ relates to the information, $F1_{perc}$ to the perception, and $F1_{back}$ to the background scopes. respectively.

The aggregate metric, $F1_{macro}$, is a *macro F-score* of the three context scopes:

$$F1_{macro} = \frac{F1_{inf} + F1_{perc} + F1_{back}}{3} .$$

The $F1_{macro}$ measures the average class-specific agreement at one particular annotation task.

The second aggregate IAA is $F1_{total}$, for which we ignore the scope classifications and only compare the agreement on the whole annotated area of the two annotators, represented by $precision_{total}$ and $recall_{total}$.

$$F1_{total} = \frac{2 \times precision_{total} \times recall_{total}}{precision_{total} + recall_{total}} .$$

The $F1_{total}$ metric evaluates the class-unspecific agreement at one particular annotation task.

With Cohens Kappa (κ), we measure agreement on the label assignment for mutually annotated areas. We follow the common definition of

$$\kappa = \frac{p_o - p_e}{1 - p_e} ,$$

where p_o is the proportion of agreement and p_e is the expected proportion of agreement expected by chance.

At the end of the guideline-building process, we reached a $F1_{macro}$ of 0.49 and a $F1_{total}$ of 0.71, showing that even though our annotation scheme captured the context accurately, the annotation of the distinct scopes is more complex. This is confirmed in the specific IAA measure. While the $F1_{inf}$ at 0.69 indicates high agreement, the $F1_{perc}$ and $F1_{back}$ are at 0.48 and 0.30, respectively. The κ is at 0.54.

During the annotation of the final dataset, we measured the IAA on 10% of the tasks to continuously control annotation quality. Here the $F1_{macro}$ is at 0.48, the $F1_{total}$ at 0.75, the $F1_{inf}$ at 0.65, the $F1_{perc}$ at 0.42, and the $F1_{back}$ at 0.34, following the same trend as before. The κ is at 0.55.

Both results show that despite the complexity of the task and the rather unskilled annotators, our annotation schema can sufficiently capture a wide variety of citation contexts, however the increasing complexity of answering the question of how (perception) and why (background) a citation marker was placed results in a lower agreement on class-assignment. However, the overall high $F1_{total}$ indicates that despite the complexities of assigning perception and background scope, a mutual understanding of what belongs to the citation scope and what does not exists. The κ of 0.55 is classified as moderate agreement (Landis and Koch 1977).

5.2 Corpus Statistics

The FINECITE corpus contains 1,056 manually annotated citation contexts for different paragraphs from 72 scientific papers. Overall, the information scope accounts for 27% of the annotated words, the perception scope for 35%, and the background scope for 38% of the annotated words. Figure 2 provides a detailed visualization of the context distribution.

To assess our assumptions on structural characteristics, we calculated the $F1_{macro}$ and $F1_{total}$ between the FINECITE framework and several common context restrictions like sentence segmentation, contiguity, and fixed-size context windows. For the sentence segmentation, we first assigned the majority label and then a priority label fol-

Category	F1 Macro (majority / priority)	F1 Total
sentence-level	0.63 / 0.57	0.95
contiguity (token)	0.88	0.88
fixed-sized:		
- 1 sent	0.44 / 0.38	0.72
- 2 sent (avg.)	0.53 / 0.47	0.73
- 4 sent	0.61 / 0.55	0.69

Table 2: Average similarity between fine-grained context representation and common approaches in the literature. For $F1_{macro}$, majority and priority tokens are assigned to sentence segments. For more details, refer to Section 5.2.

lowing the previously described scope order (information > perception > background). The results are shown in Table 2.

The result clearly demonstrates the importance of the three structural assumptions (sub-sentence segmentation, non-contiguity, and dynamic context) in the FINECITE Dataset. While the effect of contiguity is considerable in comparison, sentence segments, and fixed-size context windows would heavily restrict the captured information. The $F1_{total}$, which depicts a case where the context scopes are not considered, indicates that sentence segmentation might be a feasible tactic in such a case.

6 FINECITE: Baselines

This section evaluates the FINECITE framework on two important CCA Tasks. The evaluation of the framework should provide (*i*) an analysis of the performance of the framework on important CCA applications, (*ii*) a comparison to previous approaches, and (*iii*) a test of the assumption on structure and information representation regarding model performance. We do this by first assessing whether common models can learn the extraction framework using a limited amount of training data and, second, investigating the quality of the extracted context by comparing it to other extraction approaches on two citation classification tasks.

6.1 Citation Context Extraction

The extraction task aims to identify how capable commonly used models are of learning the proposed FINECITE framework on a limited amount of data.

Data segmentation. We applied two different segmentation approaches to the FINECITE data: sentence and sub-word-token segmentation. For the sentence segments, we adjusted the gold context to fit the segment boundaries first by finding the majority label of the sequence and second by assigning the priority label according to the priorities mentioned in Section 5.1.

Extraction model. For the extraction model, we chose the common approach of an embedding model with a linear classification head (Lauscher et al. 2022)). We applied three different sequence embedding models, SCIBERT (Beltagy, Lo, and Cohan 2019), a domain-specific BERT Model (Devlin et al. 2019), and two with bidirectional attention retrofitted large decoder-only language models (BehnamGhader et al. 2024)—LLM2Vec Mistral 7B (Jiang et al. 2023) and the LLM2Vec Llama 3 8B (AI@Meta 2024). The final hidden layers of the embedding models were respectively passed to a linear classifier to predict the class of the given segments. Subsequently, the cross-entropy loss was calculated. For the two LLM2Vec models, we applied QLoRA (Dettmers et al. 2023) for parameter-efficient fine-tuning.

Experiment setup. We used the pre-trained weights of SCIBERT, LLM2Vec Mistral 7B, and LLM2Vec Llama 3 8B from huggingface transformers (Wolf et al. 2020). As an optimizer, we used AdamW (Loshchilov and Hutter 2019) with a linear warm-up-ratio of 10% steps, a peak learning rate of 5e-5, and after that, linearly decaying overall training steps. All models were fine-tuned using early stopping with patients of two epochs, a batch size of 2, and a dropout of 0.1. We used the same LoRA configuration as used in the LLM2Vec fine-tuning (BehnamGhader et al. 2024). All training was conducted on NVIDIA A100 GPU.

We evaluate the citation context extraction performance with the metrics described in Section 5.1: $F1_{macro}$, $F1_{total}$, $F1_{inf}$, $F1_{perc}$, and $F1_{back}$. Regarding sentence segmentation, we provide separate results for the majority and priority-token approach. We further distinguish between the extraction, taking the three citation scopes into account, and the total extraction, only distinguishing between context or not-context.

Result. Table 3 exhibits the results. The different extraction models show similar performances but slightly outperform the human performance during guideline creation and annotation. The LLM2Vec Mistral 7B, with a $F1_{macro}$ of 0.51, slightly outperforms the other models on fine-grained context extraction, while the SCIBERT model exhibits peak performance on general citation context representation with a $F1_{total}$ of 0.74. The three specific metrics $F1_{inf}$, $F1_{perc}$, and $F1_{back}$ follow a similar pattern as in the annotation task, the $F1_{back}$, however, is a bit higher in model extraction.

6.2 Citation Context Classification

With the citation classification tasks, we competitively evaluate the capability of the FINECITE framework to compress context information.

Data. As the FINECITE dataset does not have citation class labels, we instead use two alternative datasets. The first one is the ACL-ARC Dataset (Jurgens et al. 2018), consisting of 1933 citances labeled with one of six citation functions, all from scientific publications from the ACL community. We use the version of ARC-ACL published in (Nambanoor Kunnath, Pride, and Knoth 2022). The second Dataset we use is the SDP-ACT Dataset (N. Kunnath et al. 2021), which is a larger, mixed domain dataset with 4000 entries labeled in roughly the same manner as the ACL-ARC Dataset. As both datasets showed stark differences in model performance regarding the train-test-split, we applied 5-fold cross-validation to retrieve our results. Also, it should be noted that both datasets are annotated on the citing sentence alone, thus making the dataset prone to one-sentence-sized prediction.

Model	Macro F1	Total F1	F1 Inf	F1 Perc	F1 Backg
Human (guideline creation)	0.49	0.71	**0.69**	**0.46**	0.30
Human (annotation)	0.48	**0.75**	0.65	0.42	0.34
SciBERT & linear classifier	0.49	**0.74**	0.66	**0.46**	0.36
Mistral 7B & linear classifier	**0.51**	0.73	**0.69**	**0.46**	**0.37**
Llama 3 8B & linear classifier	0.50	0.73	**0.69**	0.45	0.36

Table 3: *Results of the citation context extraction task*: The first two rows show inter-annotator agreement during the guideline creation and annotation process. The three baseline extraction models slightly outperform the human annotations, demonstrating the capability to apply the FINECite framework automatically.

Approach	Context	ACL-ARC		SDP-ACT	
		macro	micro	macro	micro
Cohan et al. 2019	citance	0.583	0.678	0.227	0.433
Kunnath et al. 2020	dynamic number of sentences	0.581	0.689	0.269	0.428
Lauscher et al. 2020	dynamic number of sentences	0.582	0.698	0.254	0.447
FINECite (our work)	dynamic number of words	0.599	0.695	0.264	0.425
	dynamic number of phrases	**0.609**	**0.708**	**0.275**	**0.459**
	dynamic number of sentences	0.591	0.681	0.260	0.440

Table 4: *Results of the citation intent classification task:* Three baseline and three FINECite context are evaluated on two common citation classification datasets (ACL-ARC, SDP-ACT). Our FINECite work outperforms the three baseline extraction approaches, showing the advantage of fine-grained CCA.

Data processing. We extract several different contexts from the two datasets. The first is a citance context similar to the one used in Cohan et al. (2019). For the second and third baseline contexts, we reproduce the dynamic extraction approaches of Nambanoor Kunnath, Pride, and Knoth (2022) and Lauscher et al. (2022). For the FINECite context extraction, we apply three segmentation methods imposed on the predicted labels: word, phrase, and sentence segmentation. For the sentence and phrase tokenization, we use the NLTK package (Bird and Loper 2004), and in both cases, the majority label was assigned to the new segment. All FINECite contexts were extracted with the fine-tuned LLM2Vec Mistral 7B model (BehnamGhader et al. 2024).

Classification model. We considered the best-performing citation classification model from the 3C classification task 2021 (N. Kunnath et al. 2021), or a SciBERT model (Beltagy, Lo, and Cohan 2019) with a linear classification head and weighted loss to deal with class imbalance (Maheshwari, Singh, and Varma 2021).

Experiment setup. We used the pre-trained weights of SciBERT from huggingface transformers (Wolf et al. 2020). The best performance was achieved using AdamW (Loshchilov and Hutter 2019), a learning rate of 2e-5, a dropout of 0.1, and a batch size of 8. We used early stopping, with a linear warm-up over 10% of the training. All training was conducted on NVIDIA A100 GPU.

Result. We evaluate the model performance on the mean micro and macro F-scores over all folds. Table 4 exhibits the results. Our model shows superior performance regarding the three baseline models. The best-performing version is the sub-sentence phrase-segmented context with a $F1_{macro}$ of 0.61 and a $F1_{micro}$ of 0.71 on the ACL-ARC and a $F1_{macro}$ of 0.28 and a $F1_{macro}$ of 0.56 on the SDP-ACT dataset. The word, segmentation, has a slightly lower performance, however, is better than sentence segmentation.

7 Conclusion

In this paper, we presented fundamental work to spark new research in the area of citation context analysis. We provided a conceptual approach for understanding current issues in citation context analysis and proposed a novel framework, termed FINECite, that covers three information-driven context scopes and addresses the *what*, *how*, and *why* of a citation respectively. We further employed that framework to create the FINECite Dataset, which is the first dataset addressing the three commonly existing issues of fixed-size context windows, contiguity, and sentence-segmented citation contexts. Based on this dataset, we provided a baseline model for the citation extraction of previously unseen texts, evaluated the extracted context on citation function classification, and showed that the model effectively performed with our corpus.

Future work will include a) the extension of the FINECite corpus in size and scientific domain, b) the exploration of structured citation extraction approaches using novel domain-specific LLMs (Wadden et al. 2024), and c) the creation of a comprehensive benchmark for the evaluation of citation representation frameworks.

References

Abu-Jbara, A.; Ezra, J.; and Radev, D. 2013. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In Vanderwende, L.; Daumé III, H.; and Kirchhoff, K., eds., *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 596–606. Atlanta, Georgia: Association for Computational Linguistics.

Abu-Jbara, A.; and Radev, D. 2012. Reference Scope Identification in Citing Sentences. In Fosler-Lussier, E.; Riloff, E.; and Bangalore, S., eds., *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 80–90. Montréal, Canada: Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Athar, A.; and Teufel, S. 2012. Detection of Implicit Citations for Sentiment Detection. In Van Den Bosch, A.; and Shatkay, H., eds., *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, 18–26. Jeju Island, Korea: Association for Computational Linguistics.

Bai, X.; Wang, M.; Lee, I.; Yang, Z.; Kong, X.; and Xia, F. 2019. Scientific Paper Recommendation: A Survey. *IEEE Access*, 7: 9324–9339.

BehnamGhader, P.; Adlakha, V.; Mosbach, M.; Bahdanau, D.; Chapados, N.; and Reddy, S. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. arXiv:2404.05961.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Bird, S.; and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

Cohan, A.; Ammar, W.; van Zuylen, M.; and Cady, F. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 3586–3596. Minneapolis, Minnesota: Association for Computational Linguistics.

Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; and Weld, D. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270–2282. Online: Association for Computational Linguistics.

Cohan, A.; and Goharian, N. 2015. Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 390–400. Lisbon, Portugal: Association for Computational Linguistics.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20: 37 – 46.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 10088–10115. Curran Associates, Inc.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Frost, C. O. 1979. The Use of Citations in Literary Research: A Preliminary Classification of Citation Functions. *The Library Quarterly: Information, Community, Policy*, 49(4): 399–414.

Garfield, E. 1955. Citation Indexes for Science. *Science*, 122(3159): 108–111.

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America vol. 102,46*.

Hripcsak, G.; and Rothschild, A. S. 2005. Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12 3: 296–8.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jurgens, D.; Kumar, S.; Hoover, R.; McFarland, D.; and Jurafsky, D. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6: 391–406.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Lauscher, A.; Ko, B.; Kuehl, B.; Johnson, S.; Cohan, A.; Jurgens, D.; and Lo, K. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1875–1889. Seattle, United States: Association for Computational Linguistics.

Lopez, Patrice and Foppiano, Luca and others. 2008-2024. Grobid: A Machine Learning Software for Extracting Information from Scholarly Documents. https://github.com/kermitt2/grobid.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Maheshwari, H.; Singh, B.; and Varma, V. 2021. SciBERT Sentence Representation for Citation Context Classification. In Beltagy, I.; Cohan, A.; Feigenblat, G.; Freitag, D.; Ghosal, T.; Hall, K.; Herrmannova, D.; Knoth, P.; Lo, K.; Mayr, P.; Patton, R. M.; Shmueli-Scheuer, M.; de Waard, A.; Wang, K.; and Wang, L. L., eds., *Proceedings of the Second Workshop on Scholarly Document Processing*, 130–133. Online: Association for Computational Linguistics.

N. Kunnath, S.; Pride, D.; Herrmannova, D.; and Knoth, P. 2021. Overview of the 2021 SDP 3C Citation Context Classification Shared Task. In Beltagy, I.; Cohan, A.; Feigenblat, G.; Freitag, D.; Ghosal, T.; Hall, K.; Herrmannova, D.; Knoth, P.; Lo, K.; Mayr, P.; Patton, R. M.; Shmueli-Scheuer, M.; de Waard, A.; Wang, K.; and Wang, L. L., eds., *Proceedings of the Second Workshop on Scholarly Document Processing*, 150–158. Online: Association for Computational Linguistics.

Nakov, P.; Schwartz, A.; and Hearst, M. 2004. Citances: Citation sentences for semantic analysis of bioscience text.

Nambanoor Kunnath, S.; Pride, D.; and Knoth, P. 2022. Dynamic Context Extraction for Citation Classification. In He, Y.; Ji, H.; Li, S.; Liu, Y.; and Chang, C.-H., eds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 539–549. Online only: Association for Computational Linguistics.

Ostendorff, M.; Rethmeier, N.; Augenstein, I.; Gipp, B.; and Rehm, G. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11670–11688. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Pride, D.; and Knoth, P. 2020. An Authoritative Approach to Citation Classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, 337–340. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375856.

Radev, D. R.; Muthukrishnan, P.; and Qazvinian, V. 2009. The ACL Anthology Network Corpus. In Kan, M.-Y.; and Teufel, S., eds., *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, 54–61. Suntec City, Singapore: Association for Computational Linguistics.

Rohatgi, S. 2022. ACL Anthology Corpus with Full Text. https://github.com/shauryr/ACL-anthology-corpus.

Swales, J. 1986. Citation Analysis and Discourse Analysis. *Applied Linguistics*, 7(1): 39–56.

Teufel, S.; Siddharthan, A.; and Tidhar, D. 2006. Automatic classification of citation function. In Jurafsky, D.; and Gaussier, E., eds., *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103–110. Sydney, Australia: Association for Computational Linguistics.

Wadden, D.; Shi, K.; Morrison, J.; Naik, A.; Singh, S.; Barzilay, N.; Lo, K.; Hope, T.; Soldaini, L.; Shen, S. Z.; Downey, D.; Hajishirzi, H.; and Cohan, A. 2024. SciRIFF: A Resource to Enhance Language Model Instruction-Following over Scientific Literature. arXiv:2406.07835.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.

A Context Scope Description

Table 5 gives a structured overview of the three citation scopes. A more detailed depiction of the different facets of the context scope as well as some example annotations are shown in Appendix B.

B Annotation Guidelines

The annotation guidelines are added to the multimedia appendix and will be released to the public alongside the dataset on paper acceptance.

C Annotation Interface

Figure 3 shows the annotation tool with an annotated example and different features, facilitating an efficient context annotation.

D Inter Annotator Agreement

The F-measure for IAA is calculated by

$$F1 = \frac{2 \times precision \times recall}{precision + recall},$$

where *precision* refers to the proportion of agreement on the annotation of annotator 1 and *racall* refers to the proportion of agreement on the annotation of annotator 2.

E Hyperparamether Tuning

The following hyperparameters were explored for both baseline tasks respectively. **Citation Context Extraction.**

- learning rate: 1e-5, 2e-05, 5e-05, 8e-05, 1e-04
- batch size: 1, 2, 4, 8, 16
- dropout: 0.1, 0.15, 0.2, 0.25

Citation Classification

- learning rate: 1e-5, 2e-5, 5e-5
- batch size: 2, 4, 8, 16, 32
- dropout: 0.0, 0.1, 0.2

F Reproducibility Guidelines

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced **Yes**
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results **Yes**
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper **Yes**

Does this paper make theoretical contributions? **No**

Does this paper rely on one or more datasets? **Yes**
If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets **Yes**
- All novel datasets introduced in this paper are included in a data appendix. **Yes**
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. **Yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. **Yes**
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. **NA**

Does this paper include computational experiments? **Yes**
If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. **Yes**
- All source code required for conducting and analyzing the experiments is included in a code appendix. **Yes**
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Yes**
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from **Yes**
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. **NA**
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. **Yes**
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. **Yes**
- This paper states the number of algorithm runs used to compute each reported result. **Yes**
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. **Parital**
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). **No**
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. **Yes**
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. **Yes**

Scope	Category	Description
Information Scope	Content	Information from the paper
	Temp, Loc, Pers	Meta Information
Perception Scope	Use	Usage of cited content
	Judgement	Judgment of cited content
Background Scope	Background	Background information
	Further Information	Further information

Table 5: The three annotation scopes our labeling scheme is based on, with a short description of each scope and its biggest categories.

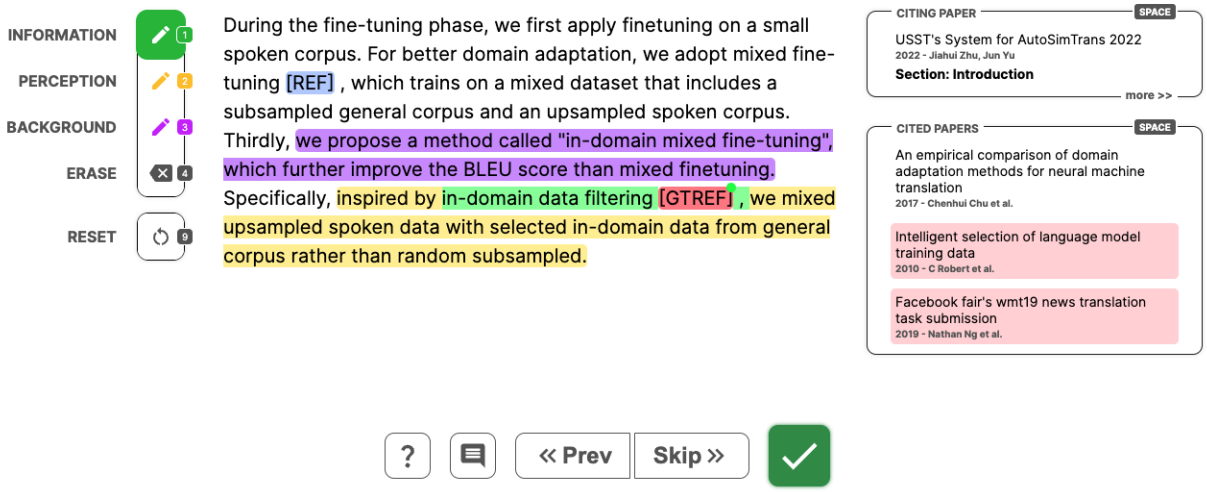


Figure 3: The Annotation Interface: Located on the left is the annotation toolbar, with the color-coded marker for each context scope, an erase tool, and the reset button. The center is the working area where the annotation task is displayed, and annotated. On the right side meta information regarding the citing and cited paper is provided, and alternatively, a comment section can be accessed to leave questions or notes. The navigation bar on the bottom gives (from left to right) access to the annotation guidelines, the comment section, and three buttons for returning to the previous task, skipping, or submitting the current task.