FINECITE: A Novel Approach on Fine-Grained Citation Context Analysis

Anonymous ACL submission

Abstract

Citation context analysis (CCA) is a field of research studying the role and purpose of citation in scientific discourse. While most of the efforts in CCA have been focused on elaborate characterization schemata to assign function or intent labels to individual citations, the citation context as the basis for such a classification has received rather limited attention. This relative neglect, however, has led to the precedence of vague definitions and restricting assumptions, limiting the citation context in its expressiveness. It is a common practice, for example, to restrict the context to the citing sentence. While this might be enough to cover mentions and background citations, more influential ones are often thoroughly discussed, extending beyond a one-sentence context window. To address this concern, we analyze the semantic structure of citation contexts in terms of their elemental dimensions and distribution throughout the citing text. To evaluate this approach, we construct and publish the FINECITE Corpus containing 1,056 manually annotated fine-grained citation contexts. Our experiments on established CCA benchmarks demonstrate the effectiveness of our finer-grained context definition, showing improvement compared to state-of-the-art approaches. We will release our code and dataset to the public upon acceptance.

1 Introduction

Scientific research is inherently collaborative, with each discovery building upon a foundation of prior studies. To acknowledge previous work and provide proper credit to original authors, it is standard practice to include citations that connect past findings to new contributions. Recognizing the importance of citations in scientific communication, researchers have extensively studied their role and purpose—a field known as citation context analysis (CCA) (Kunnath et al., 2022; Swales, 1986).

In computational linguistics, CCA is mainly concerned with the automatic classification of citations

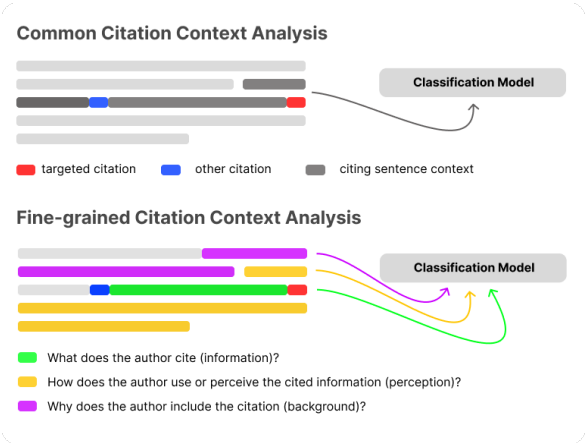


Figure 1: Citation contexts in the citation context analysis literature are often represented by the citing sentence. The related information, however, often extends to the surrounding sentences.

along various dimensions, such as citation function (e.g. Lauscher et al. 2022; Cohan et al. 2019; Jurgens et al. 2018; Teufel et al. 2006), sentiment (e.g. Lauscher et al. 2017; Abu-Jbara et al. 2013; Athar and Teufel 2012), or influence (e.g. Pride and Knoth 2020; Cohan et al. 2019). Given a passage of text surrounding the citation marker–the citation context–the task is to assign one or multiple classes defined by a citation classification scheme.

Although much of the research in CCA has focused on classification schemes, the citation context itself—the basis for these classifications—has received *relatively little* attention. This lack of focus has led to an absence of a clear, comprehensive definition and the prevalence of datasets with overly simplistic and coarse-grained contexts (e.g. Pride and Knoth 2020; Cohan et al. 2019). This gap poses two key challenges for the field of CCA. First, the limited understanding of citation contexts hinders the performance of classification models (Nambanoor Kunnath et al., 2022). Second, it creates barriers to applying CCA in modern text-based systems, such as retrieval-augmented generation

(RAG) (Lewis et al., 2020; Edge et al., 2024) and question-answering (Q&A) frameworks (Lauscher et al., 2022; Dasigi et al., 2021).

Defining citation context in a practical and meaningful way, however, poses several challenges. First, citation contexts are often interwoven with the broader argument, making it difficult to separate citation-related content from unrelated information. Second, understanding scientific texts and their contexts requires domain-specific expertise, typically limited to a small group of experts who may not have the time for labor-intensive context annotation.

To address these challenges, this paper aims to develop and evaluate a *citation context definition* that: (i) is structured around the semantic composition of the context rather than rigid sentence boundaries, (ii) reduces ambiguity in determining context boundaries, and (iii) remains practical and applicable without deep domain expertise.

This paper addresses those objectives with the following contributions:

- We identify the elemental building blocks of the citation context and provide a, to our knowledge, first, comprehensive fine-grained definition of the citation context.
- We construct and publicize the FINECITE *corpus* comprising 1,056 manually annotated fine-grained citation contexts.
- Our experiments on established benchmarks demonstrate that fine-grained context improves citation function classification by up to 25% compared state-of-the-art approaches.

The rest of the paper is organized as follows. The subsequent section reviews the relevant literature in the field of CCA. Section 3 provides the analysis and definition of the citation context. In Section 4, we describe the curation process of our novel corpus, termed FINECITE. In Section 5, we evaluate our context definition on established benchmarks and compare its performance to state-of-the-art approaches. Section 6 concludes the paper with a summary and suggests future research directions.

2 Related Work

CCA is the subject of a substantial body of research with (Garfield, 1972) often mentioned as one of the pioneering works. Reaching back to (Teufel et al., 2006), CCA research in computational linguistics is commonly conceptualized as an automatic mapping of varying spans of text surrounding the citation marker to a set of commonly occurring classes, like citation function (Lauscher et al., 2022; Jurgens et al., 2018; Teufel et al., 2006), purpose (Pride and Knoth, 2020; Abu-Jbara et al., 2013), sentiment (Athar and Teufel, 2012), or intent (Cohan et al., 2019). For a comprehensive survey on citation analysis, refer to (Kunnath et al., 2022).

Despite (i) the continued research in CCA, (ii) the introduction of new and larger datasets (Cohan et al., 2019; Jurgens et al., 2018), and (iii) updated modeling approaches (Lauscher et al., 2022; Cohan et al., 2019), some underlying paradigms of the CCA were not adapted to allow for richer citation representation. In some recent publications, even an opposite trend can be observed. (Cohan et al., 2019), for example, reduces the number of citation classes from the commonly used 6-class framework of (Jurgens et al., 2018) to merely three classes. This reduction is motivated to reduce the complexity of the task, although tragic, as the conceptualization of mutually exclusive classes and a structurally highly constrained citation contexts fall short of representing the rich information of the citation link. Table 1 compares the relevant research.

Structural Constraints of the Citation Context. The citation context is the span of text surrounding a citation marker constituting the author's description and argumentation on why a particular citation is included. As such the citation context is the basis for all CCA tasks. Previous research assumes that this context is adequately approximated by a fixed-sized window surrounding the citation marker. The size of the context window varies between one (Pride and Knoth, 2020; Cohan et al., 2019), or multiple sentences (Abu-Jbara et al., 2013; Athar and Teufel, 2012), a specific number of characters (Jurgens et al., 2018), or whole paragraphs (Teufel et al., 2006). While some older approaches (Abu-Jbara et al., 2013; Athar and Teufel, 2012) further refine the context to a dynamic-sized subset, only recent publications stress the importance of fully dynamic CCA (Lauscher et al., 2022; Nambanoor Kunnath et al., 2022). Lauscher et al. find, for example, that over one in six citation contexts extend beyond the citing sentence, providing additional information through the extended context.

Secondly, citation contexts are assumed to stretch continuously from the citation marker. Even

AUTHOR (YEAR)	TASK	NO. CLS.	SEMANTIC	DYNAM.	NON-CONT.	SUB-SENT.
Lauscher et al. (2022)	function cls.	7	✘	✔	✔	✘
Kunnath et al. (2022)	function cls.	6	✘	✔	✔	✘
Ferrod et al. (2021)	intent cls.	5	✔	(✔)	(✔)	✔
Pride and Knoth (2020)	purpose cls.	6	✘	✘	✘	✘
Cohan et al. (2019)	intent cls.	3	✘	✘	✘	✘
Jurgens et al. (2018)	function cls.	6	✘	✘	✘	✘
Abu-Jbara et al. (2013)	purpose cls.	6	✘	(✔)	✔	✘
Athar and Teufel (2012)	sentiment cls.	3	✘	(✔)	✔	✘
Abu-Jbara and Radev (2012)	context ext.	-	✔	(✔)	(✔)	✔
Teufel et al. (2006)	function cls.	11	✘	✘	✘	✘
FINECITE (this work)	context ext.	-	✔	✔	✔	✔

Table 1: Structural comparison of previous work in computational linguistics on CCA (SEMANTIC = semantic based conceptualization, DYNAM. = Dynamic context, NON-CONT. = Non-contiguous Context, SUB-SENT. = Sub-sentence context)

though a notable number of publications technically allow for the extraction of non-contiguous contexts (Lauscher et al., 2022; Abu-Jbara et al., 2013; Athar and Teufel, 2012)), only one study (Nambanoor Kunnath et al., 2022) particularly investigated the phenomenon. The authors directly compared a non-contiguous context window with a smaller contiguous version and found that the former slightly outperforms the latter.

Thirdly, the context is restraint through the presumption of sentence segmentation as the atomic unit of information in citation contexts (Cohan et al., 2019; Nambanoor Kunnath et al., 2022; Lauscher et al., 2022). This, however, is not necessarily the case. Abu-Jbara and Radev (2012), for instance, shows evidently that sentences with multiple citations consist of multiple sub-sentence context fragments. We also observe this phenomenon beyond the multi-citation settings.

Conceptual Restraints. Next to the restriction of the citation context, the conception as single label classification task was criticized (Lauscher et al., 2022). The most prevalent form of CCA is the classification of a citation on a schema of mutually exclusive labels (Pride and Knoth, 2020; Cohan et al., 2019; Jurgens et al., 2018). Lauscher et al. (2022) addressed this by creating a multi-labeled dataset. They find that nearly one in five citations have at least two labels, with some reaching up to four. Another solution to the restraints of single label classifications comes from Ferrod et al. (2021). Instead, they supplement their class assignments with snippets of contextual information, further enriching the citation link. They define the *citation object* as the concept taken into consideration by the citation and the *context* as background information, or constraints on the object. In light of the vast improvements in text-understanding during the last years (Brown et al., 2020; Vaswani et al., 2017) the enrichment of the CCA task with context infromation seems promising, as frameworks like retrieval-augmented generation (RAG) (Lewis et al., 2020; Edge et al., 2024), or question-answering (Q&A) systems (Lauscher et al., 2022; Dasigi et al., 2021) would benefit from it. The conceptualization presented in Ferrod et al. (2021), however, captures only a limited subset of the diverse citation contexts found in scientific literature.

3 Approach: Defining Semantic Dimensions and Structural Properties for Fine-grained Citation Contexts

Through examples, we explore the fundamental semantic dimensions of the citation context and examine its structural properties in a natural setting.

Semantic Dimensions. Previous research on argumentation recognition in scientific texts is categorized along four principal semantic classes (Teufel, 2014): (i) statements about the author's own work (citing paper), (ii) properties of existing solutions (cited paper), (iii) the relationships between existing solutions and the author's contribution, and (iv) general properties of the research space. When

applied to the problem of citation context definition, (ii) and (iii) encompass most of the relevant contextual information.

We define the first semantic dimension of citation contexts as the information the citing author references from the cited paper. Consider the following citation:

> "*...our paper extends the citation labeling scheme of <CITATION> and then reports similarities ...*"

This phrase, "*The citation labeling scheme of <CITATION>*," describes here *what* information the author is referring to. In the following, we denote this as the *Information Dimension* (INF).

The second dimension describes the direct relationship between the citing and the cited work. In the excerpt

> "*...our paper extends the citation labeling scheme of <CITATION> and then reports similarities ...*"

the passage "*our paper extends*" describes *how* the author uses the cited information in their work. While use constitutes a major fraction of occurring relations, other forms of perception, such as judgment or comparison, must also be considered. Thus, we refer to it as the *Perception Dimension* (PERC).

While these two dimensions cover the most critical aspects of a citation context,–*what* is cited and *how* is it perceived or used–they do not necessarily include the information encompassing *why* the author chose to include a citation.

Consider the following example:

> "*Unlike recent language representation models <CITATION>, BERT is designed to pretrain deep bidirectional representations from...*"

The sole reason for the citation relates to a property of the novel contribution, which falls under the semantic class (i) and has not yet been considered in the citation context. In other instances, such a motivating factor could be related to a property of the research space (iv) or other direct citations covered by (ii) and (iii). We categorize these passages, which explain *why* a citation was included, as the *Background Dimension* (BACK).

A definition of the citation context at this level of abstraction has a significant advantage in that an annotator does not require an in-depth understanding of the propositional content. General skills in scientific text understanding are mostly sufficient.

Structural Properties. To allow for unconstrained citation contexts, we should consider three additional structural properties presented below. The first property is the *dynamic length of citation contexts*. This significance has been shown in recent works (Lauscher et al., 2022; Nambanoor Kunnath et al., 2022), which criticized the existing precedence of fixed-size context windows. Approaches neglecting this property tend to converge, through noise minimization, toward a single-sentence context window (Cohan et al., 2019), as most citation contexts do not extend beyond the citing sentence. More influential citations discussed more extensively are often not covered sufficiently in such a case.

Secondly, *citation contexts are non-contiguous*. In this example

> "*These include sentence-level tasks such as natural language inference <CITATION> and paraphrasing <CITATION>, which aim to predict the relationships between sentences by analyzing them holistically...*"

The first citation disrupts the context of the second, dividing it into two non-contiguous segments. This type of sub-sentence non-contiguity is frequently observed in multi-citation sentences (Abu-Jbara and Radev, 2012), but also occurs at the sentence level (Nambanoor Kunnath et al., 2022).

The third property is the *sub-sentence granularity of citation contexts*. Although sub-sentence segmentation proved to be the least critical of the three properties (see Section 4.2 for further discussion), it remains essential for annotation along the semantic dimensions. As demonstrated in all examples in this section, fine-grained citation context segments often exhibit a sub-sentence structure.

4 FINECITE: A Novel Corpus for Fined-grained Citation Context

Now, we present and elaborate on a new corpus, FINECITE, with fine-grained citation contexts manually annotated using the provided context definition.

With the dataset creation, we aim to (*i*) assess whether the theoretical framework practically applies to scientific texts, (*ii*) investigate the assumption on the semantic dimensions and structure of citation contexts, and (*iii*) create a resource for the evaluation of the framework on established CCA Benchmarks.

4.1 Dataset Construction

We construct the corpus in the following steps.

Step 1: Procurement. Our dataset was built on a sample of ACL Anthology Network Corpus (Radev et al., 2009), containing over 80K papers from several ACL conferences and other venues in computational linguistics. We used the GROBID (GROBID, 2024) library to parse the full-paper pdf dataset provided by Rohatgi (2022). Documents containing faulty meta-information, languages other than English, and miscellaneous documents with >3 sections and >5 references were skipped. We sampled 1,056 paragraphs from the remaining documents, each containing one citation marker highlighted as the target citation.

Step 2: Guideline creation. The annotation guidelines comprise best practices and rules for annotating a paragraph per the three context dimensions. The instructions were created in an iterative process in which several annotators completed five to ten tasks separately and subsequently updated the guidelines based on differences in the annotation. 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 in Step 4: Validation.

The complete Annotation Guidelines can be found in Appendix E.

Step 3: Annotation. The annotation was performed on a single paragraph. The annotator was asked to read the paragraph and annotate the citation context based on the guidelines. The annotation was performed using an interface that provided context and additional meta information to the annotator. A description of the book is exhibited in Appendix A

All annotators had general knowledge of scientific text understanding and were trained during the guideline creation. None of the annotators had an in-depth knowledge of computational linguistics.

Step 4: Validation. To validate the annotation quality, we computed as inter-annotator agreement (IAA) F-measure (Hripcsak and Rothschild, 2005), commonly used for evaluating span annotations. We then complemented it with Cohens κ (Cohen, 1960) for the agreement on label assignment above that is expected by chance. To capture different aspects of the annotation process separately, we provide IAA separately for each scope ($F1_{inf}$, $F1_{perc}$, and $F1_{back}$), as well as two aggregate measures ($F1_{macro}$ and $F1_{total}$), the latter covering dimension independent IAA. An exhaustive description of metrics and their variations is available in Appendix B.

During the annotation of the dataset, we measured the IAA on 10% of the tasks to control annotation quality. The final $F1_{macro}$ was 0.48, the $F1_{total}$ 0.75, and the scope specific metrics 0.65 ($F1_{inf}$), 0.42 ($F1_{perc}$), and 0.34 ($F1_{back}$) respectively. The κ on the validation samples was 0.55.

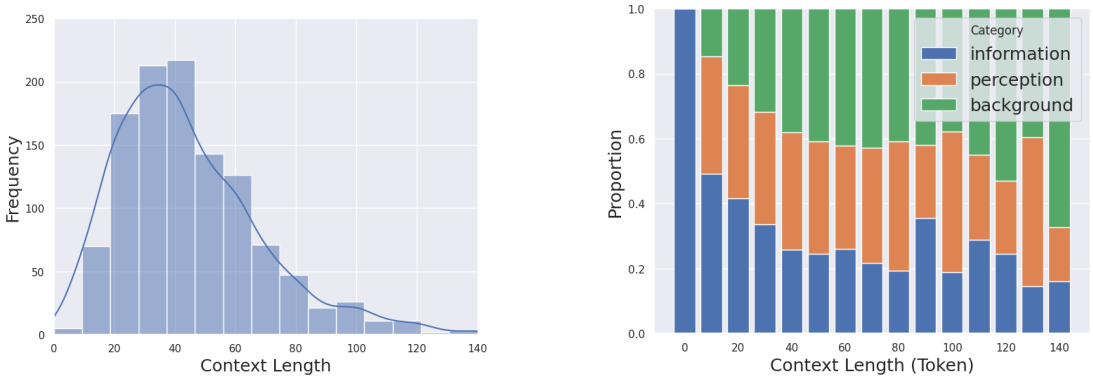
The result shows that despite the high task complexity, annotators with no in-depth domain knowledge can annotate the citation context with IAA values typical for scientific literature (Lauscher et al., 2022; Ferrod et al., 2021; Lauscher et al., 2018). The relative lower agreement on PERC and BACK indicates that information on *how* and *why* a citation is included is often more ambiguous than the INF dimension. That said, the overall high $F1_{total}$ suggests that despite the complexities of assigning the PERC and BACK dimensions, our context definition does provide a mutual understanding of what belongs to the context and what does not. The κ of 0.55 shows that the label assignment lies above expected agreement values by chance.

4.2 Corpus Statistics

Our FINECITE corpus contains 1,056 manually annotated citation contexts from 72 scientific papers for different paragraphs. Overall, INF accounts for 27% of the annotated words, the PERC for 35%, and the BACK for 38% of the annotated words. The average context is 45 words long and approximately normally distributed, with a long tail towards the upper end. The main contribution to the longer contexts is the BACK dimension. While BACK is around eight words long (30%) for context with less than 40 words, for context with more than 100 words, BACK expands to 54 words on average (43%). Paired with the low agreement value, this emphasizes the ambiguity of the BACK dimension, and a further delimitation would likely increase annotation performance. Figure 2 provides a detailed visualization of the context distribution.

To assess the assumptions on the structural characteristics of the citation context, we calculated the error between the FINECITE labels and context restrictions common in CCA datasets. We imposed restrictions like sentence segmentation, contiguity, and fixed-size context windows onto our dataset. The metrics provided are the residual to complete overlap on $F1_{total}$, $F1_{inf}$, $F1_{perc}$, and $F1_{back}$. The results are exhibited in Figure 3.

As expected, restricting the context to a fixed number of sentences results in a relatively high error, further exacerbated when considering 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.

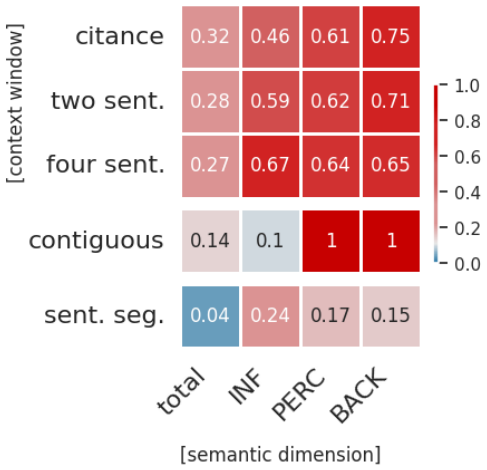


Figure 3: Error occurring when assuming some common structural restrictions

fine-grained semantic domains. Contiguity exhibits a minor error compared to fixed context windows, indicating that non-contiguity occurs less, and non-contiguous segments are rather small in size. The contiguity error of one for PERC and BACK can be explained by considering that these dimensions mostly don't occur directly next to the citation marker. Surprisingly, the total error induced through sentence segmentation is relatively tiny. While sub-sentence segmentation is necessary for fine-grained analysis, assuming sentence segmentation seems sufficient to capture the total citation context. However, one reason for the low error could be the relatively broad definition of the BACK dimension, and further delimitation would change this property.

Overall, the results affirm the significance of the three structural assumptions—*sub-sentence segmentation*, *non-contiguity*, and *dynamic context*—for a fine-grained citation context extraction.

5 Evaluation

In this section, we present the evaluation results of our context definition through (i) citation context extraction on a held-out test set of the FINECITE corpus and (ii) citation classification on standard CCA benchmarks.

5.1 Citation Context Extraction

Ensuring that common extraction models can reliably learn to extract citation contexts is crucial for effectively applying our context definition. Given the inherent complexity of this task, rigorous verification is essential to assess their reliability and performance.

Data preparation. We used the same test samples employed during the evaluation of the annotation process as the test set, with the remaining samples reserved for training. We created two versions of the datasets, (i) one dataset with uniform token labels (0, INF, PERC, BACK) and (ii) another dataset with commonly used IOB (Inside–Outside–Beginning) labels (0, B-INF, B-PERC, B-BACK, I-INF, I-PERC, I-BACK).

Extraction model. The extraction model used SCIBERT (Beltagy et al., 2019) token embeddings, which were then passed to a classification head. We evaluate three different classification heads: a linear, a Bi-LSTM (Hochreiter and Schmidhuber, 1997), or a conditional random field (CRF) (Lafferty et al., 2001). We included Bi-LSTM and

Model	$F1_{macro}$	$F1_{total}$
HUMAN (annotation)	0.48	0.75
SCIBERT & Linear	0.557	0.77
SCIBERT & Bi-LSTM	**0.56**	0.759
SCIBERT & CRF	0.521	**0.787**

Table 2: Extraction results on the FINECITE dataset

CRF, as we noticed that the linear classifier tends to become overconfident with specific tokens, assigning isolated labels far from the other context segments. BiLSTM and CRF incurred additional regional dependency, mitigating this issue. To address the dataset imbalance, where most tokens do not belong to the citation context, we applied weighted loss for the Linear and Bi-LSTM classifiers. The best results were achieved using IOB labels for linear and Bi-LSTM classifiers, whereas the CRF classifier outperformed the others regarding uniform labels.

Experiment setup. We used the pretrained weights of SCIBERT from huggingface transformers (Wolf et al., 2020). We used AdamW (Loshchilov and Hutter, 2019) with a linear warm-up ratio of 5%, a peak learning rate of 5e-5, and linear decaying for all training steps. All models were fine-tuned using early stopping with patience of three epochs, a batch size of 4, and a dropout of 0.1. The training was conducted on NVIDIA A100 GPU. We evaluated the citation context extraction performance with the metrics described in Section 4.1.

Result. Table 2 exhibits the results of $F1_{macro}$ and $F1_{total}$. See Appendix D for more detailed results. We observe that all three extraction approaches outperform the human performance during the dataset annotation. The Bi-LSTM classifier exhibits a score of 0.56, the strongest performance on the $F1_{macro}$ metric, while The CRF classifier works best in terms of the $F1_{total}$ score of 0.787. These scores outperform IAA during human annotation by 0.08 and 0.037, respectively. The performance of the linear classifier is with a $F1_{macro}$ of 0.557 and a $F1_{total}$ of 0.77, only slightly lower than the other approaches. The experiment demonstrates that standard extraction models can sufficiently extract the citation context despite the task complexity. The anecdotal evidence suggesting that the CRF and Bi-LSTM classifiers produce more concise citation contexts is not clearly reflected in the performance metrics.

5.2 Citation Context Classification

Ablation. Since the extraction task only assesses our context definition in terms of internal robustness, we further evaluated whether the fine-grained context enhances performance in citation classification.

Data. We evaluate four commonly used benchmarks in CCA as follows.

- **ACL-ARC** (Jurgens et al., 2018) comprises 1,933 labeled citances following a six-label classification schema. The source papers originate exclusively from the computational linguistics domain.
- **ACT2** (N. Kunnath et al., 2021) is a larger, mixed-domain collection of 4,000 citations labeled with the same schema as the ACL-ARC dataset.
- **SCICITE** (Cohan et al., 2019) also covers multiple domains and contains approximately 11,000 samples, annotated with a simplified three-class schema.
- **MULTICITE** (Lauscher et al., 2022) is a multi-sentence, multi-label dataset annotated with seven citation function classes based on the scheme used in ACL-ARC. With 12,653 annotated citations it is the biggest dataset.

Although ACL-ARC and ACT2 are primarily modeled using the citance alone, they offer an extended context that we can utilize. In contrast, SCICITE solely provides the citing sentence, which heavily restricts extracting a fine-grained context. To reduce the model's tendency to memorize author names, we conceal the targeted citation and other citations behind `<TARGET_CITATION/>` and `<CITATION/>` tags, respectively. Each dataset is divided into approximately 85% training and 15% testing sets. For the FINECITE approaches, add a fine-grained context label using the extraction models presented in Section 5.1.

Classification model. We considered four baselines for the classification task: (i) the scaffolding approach presented in Cohan et al. (2019). (ii) the best-performing citation classification model from the 3C classification task 2021 (N. Kunnath et al., 2021), a SCIBERT model with a linear classification head (Maheshwari et al., 2021), (iii) GPT-4o (Achiam et al., 2023), and (iv) SCITULU 70B (Wadden et al., 2024), an LLM fine-tuned on instruction-following over scientific literature, both in zero-shot setting. The FINECITE model uses

	Approach	ACL-ARC		ACT2		SciCite		MultiCite		Mean
		macro	st. dev.	macro	st. dev.	macro	st. dev.	macro	st. dev.	
Baseline	Scaffolds	0.377	0.067	0.205	0.026	0.821	0.010	0.409	0.036	0.453
	SciBERT	0.517	0.018	0.242	0.012	0.841	0.005	0.584	0.006	0.546
	GPT 4o	0.401	-	0.117	-	0.766	-	0.434	-	0.43
	SciTulu 70B	0.37	-	0.114	-	0.783	-	0.353	-	0.405
FINECite (this work)	SciBERT$_{Linear}$	0.551	0.032	**0.302**	0.02	0.84	0.002	0.603	0.021	0.574
	SciBERT$_{BiLSTM}$	**0.584**	0.014	0.282	0.014	**0.845**	0.003	0.601	0.005	**0.578**
	SciBERT$_{CRF}$	0.563	0.007	0.274	0.024	0.841	0.002	**0.606**	0.010	0.571

Table 3: Results of the citation classification task on the four benchmarks ACL-ARC, ACT2, SciCite, and MultiCite. The standard deviation (st. dev.) is calculated over five consecutive seeds.

SciBERT(Beltagy et al., 2019) embeddings and a linear classification head similar to (ii). Instead of using CLS pooling, we employ mean pooling over each context dimension. The three concatenated context dimension embeddings are then passed through a linear pre-classification head, which reduces their size to the standard embedding size. Additionally, we experimented with mean pooling over the entire context and a dimension-balanced mean pooling approach, both without the pre-classification head.

Experiment setup. We utilized the pre-trained SciBERT weights as mentioned above. The best performance was achieved using AdamW (Loshchilov and Hutter, 2019), early stopping, and a linear warm-up of 5%. The training was conducted on NVIDIA A100 GPU. The optimal learning rate and batch size for each dataset are provided in Appendix C. Performance was evaluated on the macro F-score over five consecutive seeds.

Result. Table 3 exhibits the $F1_{macro}$ and standard deviation for each dataset. More detailed results are found in Appendix D.

Among the baseline approaches, SciBERT achieves the highest average $F1_{macro}$ score (0.546), followed by Scaffolds (0.453), and GPT-4O (0.43), and SciTulu 70B (0.405). These results suggest that the complexity of the task favors a smaller fine-tuned bidirectional model over larger autoregressive models. We further observe that the Scaffolds approach exhibits a relatively high standard deviation on certain tasks, as it struggles to predict minority labels correctly.

The FINECite models introduced in this work outperform the baselines across all datasets. Among them, SciBERT$_{BLSTM}$ achieves the best overall performance, with an average $F1_{macro}$ score (0.578), surpassing the best baseline by 0.032. SciBERT$_{Linear}$ and SciBERT$_{CRF}$ perform slightly lower with average $F1_{macro}$ scores of 0.574 and 0.571 respectively.

Noteworthy is that the performance increase on the ACL-ARC and the ACT2 dataset is larger than on the SciCite and MultiCite datasets. The larger increase indicates that a fine-grained context might be especially valuable in circumstances with restricted data availability.

These results demonstrate that the fine-grained context proposed in FINECite captures the citation context better than previous approaches, enhancing citation classification performance.

6 Conclusion and Future work

In this paper, we introduced a foundational approach to defining citation context, aiming to foster new research in citation context analysis. We proposed a conceptual framework that characterizes citation context based on semantic dimensions and structural properties. To the best of our knowledge, we are the first to apply this definition in annotating the FINECite corpus, the first dataset with fine-grained citation context annotations. Our analysis demonstrated that this definition is practically applicable and that incorporating fine-grained context improves performance on established CCA benchmarks compared to state-of-the-art methods.

We will focus on expanding the dataset to cover a wider range of scientific texts and domains, further refining the Back dimension. Additionally, we plan to explore new applications, such as retrieval-augmented generation (RAG) (Lewis et al., 2020; Edge et al., 2024) and question-answering (Q&A) frameworks (Lauscher et al., 2022; Dasigi et al., 2021), to support interactive exploration of scientific argumentation structures.

7 Limitations

This work introduces a very first dataset, though it is limited in size and domain coverage. Consequently, the provided evaluation and analysis should be interpreted within this scope and may not generalize to broader contexts. Our primary focus is establishing a comprehensive definition of citation context rather than addressing the challenge of domain adaptation in CCA. Additionally, the dataset was annotated by individuals without extensive domain expertise.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.

Amjad Abu-Jbara and Dragomir Radev. 2012. Reference scope identification in citing sentences. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 80–90, Montréal, Canada. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Awais Athar and Simone Teufel. 2012. Detection of implicit citations for sentiment detection. In Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, pages 18–26, Jeju Island, Korea. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33:1877–1901.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 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), pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20:37 – 46.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4599–4610, Online. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. ArXiv, abs/2404.16130.

Roger Ferrod, Luigi Di Caro, and Claudio Schifanella. 2021. Structured Semantic Modeling of Scientific Citation Intents. In Extended Semantic Web Conference.

Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation. Science, 178(4060):471–479.

GROBID. 2024. GROBID: A Machine Learning Software for Extracting Information from Scholarly Documents. https://github.com/kermitt2/grobid.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Comput., 9(8):1735–1780.

George Hripcsak and Adam S. Rothschild. 2005. Technical brief: Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association : JAMIA, 12 3:296–8.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics, 6:391–406.

Suchetha N. Kunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. 2022. A meta-analysis of semantic classification of citations. Quantitative Science Studies, 2(4):1170–1215.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In Proceedings of the 5th Workshop on Argument Mining, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2017. Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In Proceedings of the 6th International Workshop on Mining Scientific Publications, WOSP 2017, page 24–28, New York, NY, USA. Association for Computing Machinery.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1875–1889, Seattle, United States. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. ArXiv, abs/2005.11401.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. Preprint, arXiv:1711.05101.

Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. 2021. SciBERT sentence representation for citation context classification. In Proceedings of the Second Workshop on Scholarly Document Processing, pages 130–133, Online. Association for Computational Linguistics.

Suchetha N. Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 SDP 3C citation context classification shared task. In Proceedings of the Second Workshop on Scholarly Document Processing, pages 150–158, Online. Association for Computational Linguistics.

Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2022. Dynamic context extraction for citation classification. In 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), pages 539–549, Online only. Association for Computational Linguistics.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, JCDL '20, page 337–340, New York, NY, USA. Association for Computing Machinery.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network corpus. In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL), pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

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

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

Simone Teufel. 2014. Scientific argumentation detection as limited-domain intention recognition. In Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Neural Information Processing Systems.

David Wadden, Kejian Shi, Jacob Daniel Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hanna Hajishirzi, and Arman Cohan. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. ArXiv, abs/2406.07835.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. Preprint, arXiv:1910.03771.

A Annotation Interface

Figure 4 shows the annotation tool with an annotated example and different features, facilitating an efficient context annotation.

B 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 *recall* refers to the proportion of agreement on the annotation of annotator 2.

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 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 Cohen's 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.

C Hyperparamethers for the classification task

We explored the following hyperparameters for both baseline tasks, respectively.

Table 4 shows the hyperparameters (Batch size, Learning rate, Dropout) that resulted in the optimal classification results for the ACL-ARC, ACT2, SciCite, and MultiCite datasets, respectively.

	Batch size	Learning rate	Dropout
ACL-ARC	4	5e-05	0.1
ACT2	16	3e-05	0.1
SciCite	16	3e-05	0.1
MultiCite	8	5e-05	0.1

Table 4: Hyperparameters of each dataset

D Extended Results

The following tables show extended evaluation results. Table 5 shows the extended extraction results on the FINECite dataset. Tables 6, 7, 8, and 9 show the extended classification results for ACL-ARC, ACT2, SciCite, and MultiCite respectively.

E Annotation Guidelines

E.1 Introduction

We want to annotate the citation context of references in scientific literature to build a database for the training of an automatic citation context extraction model.

The scope of the annotation is to mark the context of a citation in a given paragraph. As the citation context, we understand the citation surrounding sentence segments that semantically relate to the target reference.

We use an online platform that supports the annotation process in its structure and functionality. In the following paragraphs, we describe the annotation task and briefly introduce the annotation platform.

E.2 The Task

E.2.1 What does the annotation task look like?

The task is to classify words of several sentences in the same paragraph and determine whether they relate to the citation marked as the target reference. An example annotation task might look like this:

> Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing the modeling of dependencies without regard to their distance in the input or output sequences [GREF]. In all but a few cases [TREF], however, such attention mechanisms are

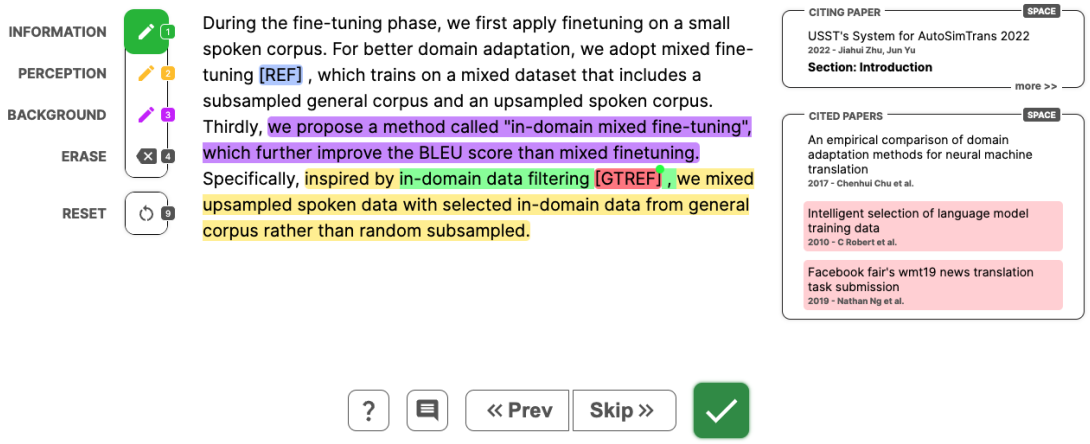


Figure 4: The Annotation Interface: Located on the left is the annotation toolbar, with the color-coded marker for each context scope, an ERASE tool, and the RESET button. The center is the working area where the annotation task is displayed and annotated. On the right side, meta-information regarding 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.

Model	$F1_{macro}$	$F1_{total}$	$F1_{INF}$	$F1_{PERC}$	$F1_{BACK}$
HUMAN (annotation)	0.483	0.758	0.654	0.416	0.338
SCIBERT & Linear	0.557	0.771	0.755	0.495	0.422
SCIBERT & Bilstm	0.56	0.759	0.768	0.496	0.415
SCIBERT & CRF	0.521	0.787	0.738	0.434	0.391

Table 5: Extended extraction results on the FINECITE Dataset.

used in conjunction with a recurrent network.

E.2.2 What is the meaning of the tags?

Four different types of tags can occur in the annotation task ([REF],[GREF],[TREF],[GTREF]). The 'REF' part of the tag generally refers to 'Reference,' meaning that each tag is some kind of placeholder for one or multiple references. More particularly, the '[REF]' tag replaces one single reference (e.g. (Goodfellow 2012) → [REF]), and the [GREF] tag replaces a Group of References (e.g. (Cohan et al. 2018, Jha et al. 2016) → [GREF]). Further, there are two different versions of the [REF] and the [GREF] tag, which indicate that they are the Target of the annotation task. The 'T' in the [TREF] and the [GTREF] tag means Target. Each annotation task will have only one target reference, but multiple other single or group references might exist.

E.3 What is the citation context?

The citation context is the text span in the citing document, which describes the information used from the cited document, the way it is used, and how the author of the citing document perceives it. For the annotation process, we distinguish between three scopes:

- Citation information scope: describes the information of the cited document. It answers the question of what is cited. [GREEN]
- Citation intention scope: describes in what way the author perceived, used, or further analyzed the document. It answers the question of how something is cited. [YELLOW]
- Citation background scope: describes additional information required for putting the two previous scopes into the context they are used in. It answers the question of why something is cited. [VIOLET]

E.3.1 General Notes

To make the annotation process possible, we have to assume some facts as given:

1. All reference Markers have been set at the correct position, and none are missing.

Approach	Backgr.			Compare			Extension			Future			Motivation			Use			Macro		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Scaffolds	.682	.764	.720	.551	.311	.393	.285	.138	.177	.095	.160	.180	.147	.240	.180	.615	.745	.673	.396	.393	.377
SciBERT	.754	.849	.798	.613	.368	.460	.755	.807	.780	.475	0.237	.317	.196	.440	.272	.395	.600	.476	.534	.550	.517
GPT 4o	.750	.677	.712	.393	.667	.494	.000	.000	.000	.400	.667	.500	.000	.000	.000	.776	634	.698	.387	.441	.401
SciTulu	.464	.684	.553	.661	.529	.587	.000	.000	.000	.400	.667	.500	.000	.000	.000	.862	.476	.613	.398	.393	.376
SciBERT$_{Linear}$	.775	.804	.789	.727	.489	.582	.415	.213	.265	.566	.760	.633	.190	.440	.263	.714	.852	.775	.565	.593	.551
SciBERT$_{BiLSTM}$	.799	.800	.798	.692	.579	.625	.432	.225	.281	.524	.880	.638	.360	.480	.341	.795	.848	.819	.600	.635	.584
SciBERT$_{CRF}$	.811	.787	.797	.740	.496	.591	.341	.250	.264	.516	.880	.649	.206	.520	.282	.726	.876	.792	.557	.635	.563

Table 6: Extended results of the citation classification task on ACL-ARC.

Approach	Backgr.			Compare			Extension			Future			Motivation			Use			Macro		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Scaffolds	.513	.722	.600	.122	.071	.089	.102	.062	.076	.288	.300	.293	.281	.090	.136	.069	.026	.035	.229	.212	.205
SciBERT	.527	.684	.595	.135	.108	.120	.340	.389	.363	.273	.092	.138	.326	.142	.198	.052	.021	.029	.298	.239	.240
GPT 4o	.773	.511	.615	.017	.020	.018	.000	.000	.000	.000	.000	.000	.038	.308	.068	.000	.000	.000	.138	.139	.117
SciTulu	.753	.507	.605	.068	.053	.060	.000	.000	.000	.000	.000	.000	.000	.000	.000	.026	.014	.018	.141	.096	.114
SciBERT$_{Linear}$	.535	.474	.495	.103	.186	.131	.475	.385	.414	.382	.554	.450	.296	.173	.208	.170	.087	.112	.327	.310	.302
SciBERT$_{BiLSTM}$	.532	.428	.471	.100	.186	.125	.393	.385	.381	.374	.495	.422	.219	.154	.176	.120	.123	.119	.290	.295	.282
SciBERT$_{CRF}$	.512	.320	.387	.087	.139	.104	.355	.354	.342	.324	.589	.417	.299	.250	.265	.113	.164	.128	.282	.303	.274

Table 7: Extended results of the citation classification task on ACT2.

2. Group references have the same (or at least sufficiently similar) information.

3. All the information mentioned in connection with the reference is from the cited document.

E.4 General Rules

1. Articles (*a*, *this*, and *the*) must be included in the scope of the following noun.

✘ **The** architecture of the system is very similar to **a** large system built for the NIST Arabic/English task [TREF]

✔ **The** architecture of the system is very similar to **a** large system built for the NIST Arabic/English task [TREF]

2. The reference marker ([REF], [TREF], etc.) must be marked as well (adjacent scope).

✘ BERT is a large language model (LLM) **[TREF]**

✔ BERT is a large language model (LLM) [**TREF**]

✘ Following [**TREF**], the loss is a sum of binary cross-entropy losses over all entity types T over all training examples D.

✔ Following [**TREF**], the loss is a sum of binary cross-entropy losses over all entity types T over all training examples D.

3. Only marks what is relevant to the targeted reference marker in case one reference is mentioned multiple times.

4. If the text is ambiguous, it should be marked in the following hierarchy: Information scope, Perception scope, and Background scope.

5. In cases where it is unclear whether the information is a contribution of the cited paper or the author, it should be marked as the author's contribution.

6. Conjunctions like "however," "in fact," "furthermore," "hence," "therefore," "in that," "on the other hand," etc., should not be included.

✘ **However**, BERT is a large language model (LLM) [TREF]

✔ **However**, BERT is a large language model (LLM) [TREF]

E.5 What is the citation information scope?

The citation Information scope of the target citation is the part of the paragraph that describes objective facts directly from the cited paper. This information is objectively true and does not involve any judgment from the author. They can be attributed as a finding of the cited paper or describe a process or judgment in the cited paper.

Approach	Backgr.			Method			Result			Macro		
	P	R	F	P	R	F	P	R	F	P	R	F
Scaffolds	.863	.873	.868	.792	.792	.792	.827	.784	.804	.827	.816	.821
SciBERT	.894	.862	.805	.805	.834	.819	.805	.855	.829	.835	.850	.842
GPT 4o	.860	.810	.834	.725	.821	.770	.671	.719	.694	.785	.784	.766
SciTulu	.803	.857	.829	.832	.726	.775	.720	.768	.743	.782	.784	.782
SciBERT$_{Linear}$	.886	.867	.876	.819	.812	.815	.796	.870	.831	.834	.850	.841
SciBERT$_{BiLSTM}$	.898	.862	.880	.823	.836	.829	.782	.875	.826	.834	.858	.845
SciBERT$_{CRF}$	.890	.863	.876	.827	.820	.822	.780	.874	.823	.832	.852	.841

Table 8: Extended results of the citation classification task on SciCite.

Approach	Backgr.			Motivation			Uses			Extends			Similarity			Differen.			Futur			Macro		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F			
Scaffolds	.732	.762	.746	.140	.093	.106	.598	.670	.631	.303	.279	.285	.415	.356	.382	.573	.485	.523	.186	.200	.189	.421	.406	.409
SciBERT	.821	.799	.810	.241	.334	.280	.740	.758	.749	.482	.552	.515	.607	.529	.565	.695	.644	.669	.437	.564	.492	.584	.597	.584
GPT 4o	.514	.715	.598	.053	.227	.086	.702	.554	.619	.436	.473	.454	.195	.556	.289	.667	.574	.617	.273	.600	.375	.406	.528	.434
SciTulu	.489	.712	.580	.011	.100	.019	.728	.557	.632	.257	.743	.382	.054	.440	.096	.699	.438	.539	.182	.286	.222	.346	.468	.353
SciBERT$_{Linear}$	.840	.788	.812	.404	.294	.338	.789	.706	.744	.536	.507	.518	.652	.464	.539	.737	.576	.643	.652	.582	.641	.659	.560	.602
SciBERT$_{BiLSTM}$	.830	.777	.802	.428	.323	.366	.753	.727	.739	.524	.525	.522	.622	.460	.526	.720	.601	.655	.600	.545	.571	.640	.564	.597
SciBERT$_{CRF}$	.827	.776	.799	.388	.415	.395	.784	.685	.729	.545	.515	.529	.647	.443	.526	.722	.598	.654	.690	.545	.606	.658	.568	.606

Table 9: Extended results of the citation classification task on MultiCite.

E.5.1 INCLUDE

Information about the contribution of the cited paper:

CONTRIBUTION

This can also be seen in BERT [TREF].

CONTRIBUTION + FACT

BERT is a large language model (LLM) [TREF].

CONTRIBUTION + PURPOSE

The architecture of the system is very similar to a large system built for the NIST Arabic/English task [TREF].

CONTRIBUTION + OUTCOME

[TREF] trains a new model called BERT, and they can show it outperforms the current state-of-the-art model.

NOTE If slightly judgmental verbs (emphasizes, stresses-out, underlines) are in an otherwise non-judgemental sentence, they should be marked as information scope.
Keywords that are referenced by they, this, etc., and belong to the information scope.

SLIGHT JUDGEMENT

[TREF] does not discuss LSP costs for internal MT development. He emphasizes on margin shrinking, which is directly linked to investment gain.

REFERENCED KEYWORDS

Recently, many reports have described studies using deep learning for dialogue systems that have achieved good performance. They can generate fluent sentences based on a user's utterances [GTREF].

E.5.2 INCLUDE

Information about used processes in the cited paper:

PROCESS

[TREF] trains their proposed mode.

PROCESS + FACT

[TREF] trains their proposed model on a classification task.

PROCESS + PURPOSE/REASON

[TREF] trains their proposed model to achieve superior performance.

E.5.3 INCLUDE

Information about outcomes or judgments in the cited paper: It should only be marked as information scope when it is clear that the judgment is from the cited paper and not from the author.

JUDGEMENT

[TREF] shows their model works well.

JUDGMENT + COMPARISON

They show their model works better than the BERT model [TREF].

JUDGMENT + FACT

[TREF] have shown how parallel suffix arrays can be used to significantly reduce the large memory footprints that phrased-based SMT systems suffer from when attempting to use longer phrases.

E.5.4 INCLUDE

Information about when, where, and by whom the paper was published: All information that gives clues about temporal, locational, or personal facts about the paper but does not judge the content in any way.

PERSONNEL

The same research team developed BERT [TREF].

TEMPORAL

Recently, BERT was introduced [TREF].

LOCATIONAL

In a paper from the ACL Conference BERT is introduced [TREF].

E.5.5 EXCLUDE

Further Information:

on SIBLING SOURCES

On a larger scale, event extraction has extended to many languages beyond English, including French [REF], Spanish [REF], Italian [TREF] and very recently, Hindi [REF].

E.5.6 EXCLUDE

Non-attributable facts: Information that can not be clearly attributed to the cited paper.

RESULTS/FINDING

Furthermore, the word embedding techniques used by [REF] or [TREF] have been shown to work well. (The position of the judgment after the ref marker makes it unsure).

E.6 What is the citation perception scope?

The citation perception scope relates to the author's subjective perception and use of the information in the cited document or a concept, the cited document is provided as an example.

E.6.1 INCLUDE

Use of the referenced information:

PROCESS

We use a BERT model pre-trained on classification [TREF].

PROCESS + FACT

We analyze a BERT model pre-trained on classification [TREF] on our dataset.

PROCESS + PURPOSE

We use a BERT model pre-trained on classification [TREF] for classifying our dataset.

PROCESS + REASON (for/against)

To increase model performance, we use the text segmentation approach suggested by [TREF].

E.6.2 INCLUDE

Judgment of the referenced information

PERFORMANCE JUDGMENT

[TREF] develop a promising classification method.

The proposed BERT model [TREF] is not reliable.

RELATIONAL JUDGEMENT

Recently Neural Networks are getting more attention. An example of this trend is BERT [TREF].

SCOPING JUDGEMENT

On a larger scale, ...; In particular...; Other common methods ..; Most of...

NOT-MENTIONED JUDGMENT

[TREF] does not discuss LSP costs for internal MT development.

JUDGMENT + COMPARISON

[TREF] shows that BERT is a reliable model. Compared to RoBERTa [REF], which employs other metrics, it is less reliable.

E.6.3 INCLUDE

A concept the citation is an example of that is strongly judged (reason for a decision): These rules only apply when the concept is subjectively judged by the authors. Only if there is a strong connection between the concept and the example strong connection words: such as, like, etc.

CONCEPT + USE

We analyze automated metrics such as BLEU [TREF].

CONCEPT + JUDGEMENT

We consider actual human judgments to be preferable to automated metrics such as BLEU [TREF].

CONCEPT + REASON

Because we care about the adequacy of post-edited translations, we consider actual human judgments to be preferable to automated metrics such as BLEU [TREF].

E.7 What is the citation background scope?

The citation background scope includes information about neither the contribution of the cited document nor how it is perceived or used but is essential for understanding its use.

E.7.1 INCLUDE

Background Information

SCOPING BACKGROUND

Text segmentation has been getting more attention recently. For example, [TREF] uses BERT to do text segmentation.

PROCESS BACKGROUND

We adopt the Lexical Conceptual Structure (LCS) of Dorr's work and use a parameter-setting approach to account for the divergences. [TREF] describes a parametric approach.

THIRD PARTY PROCESS/FACTS

Following the SAMT approach, CCG-augmented HPB SMT [REF] uses CCG [TREF] to label non-terminals.

BACKGROUND + JUDGEMENT

In fact, several GANs have recently been proposed for text generation [GREF] and have achieved encouraging results in particular, RelGAN [TREF] has outperformed state-of-the-art (SOTA) results.

BACKGROUND + COMPARISON

In fact, several GANs have recently been proposed for text generation [GREF] and have achieved encouraging results in comparison to comparable maximum likelihood approaches, in particular, RelGAN [TREF] has outperformed state-of-the-art (SOTA) results.

BACKGROUND + REASON

For comparison with the most dominant coreference dataset, OntoNotes [REF], we also measure the MUC score on our dataset. The MUC score on our dataset is 83.6, compared to 78.4-89.4 in OntoNotes, depending on the domain [TREF].

E.7.2 INCLUDE

Further information

as EXAMPLE of CONCEPT

Text segmentation [TREF] describes the process of segmenting text. An example of this would be to segment a sentence into two parts.

on COMPARISON

[TREF] shows that BERT is a reliable model. Compared to RoBERTa [REF], which employs other learning metrics, it is less reliable.

on JUDGMENT + FACT

We train another model on 80,000 Amazon kitchen reviews [TREF], and apply it on the kitchen review dev set and the Amazon electronics dev set, both having 10, 000 reviews.

as SIBLING

The use of BERT has been shown to be reliable [REF] and effective [TREF].

on PROCESS + FACT

For comparison with the most dominant coreference dataset, OntoNotes [REF], which only reported the MUC agreement score [TREF].

on LOCATION IN PAPER

Table 1 displays the result of our BERT Model. We use BLUE for evaluation. BLUE [TREF] is a metric to evaluate.... The use of BLUE is described in the following section.

on USE of JUDGMENT

..service has over 50 million users [TREF]. As native speakers of English, both authors judged the documentation to be of reasonable quality and well-formed. These initial assumptions would be tested in the project.

on USE/JUDGEMENT in THIRD PAPER

[TREF] released XY. This method was later expanded by [REF], who did xx.

E.7.3 EXCLUDE

Background of Background

BG + FACT (further information on the background)

For comparison with the most dominant coreference dataset, OntoNotes [REF], which only reported the MUC agreement score [REF], we also measure the MUC score on our dataset. The MUC score on our dataset is 83.6, compared to 78.4-89.4 in OntoNotes, depending on the domain [TREF].

EXAMPLES of BACKGROUND

Automatic extraction of events has gained sizable attention in subfields of NLP and information retrieval such as automatic summarization, question answering, and knowledge graph embeddings [GREF], as events are a representation of temporal information and sequences in text. [TREF] applies BERT for event extraction.

SIBLINGS of BACKGROUND

We adopt the Lexical Conceptual Structure (LCS) of Dorr's work and use a parameter-setting approach to account for the divergences. [TREF] describes a parametric approach.

LOCATION, PERSONA, TIME of BACKGROUND

In 2016, [REF] published Roberta based on BERT [TREF].

on LOCATION of non-attributed facts IN PAPER (it is not sure whether the part is from the paper)

Following the SAMT approach, CCG-augmented HPB SMT [REF] uses CCG [TREF] to label non-terminals. This section gives a brief introduction to CCG followed by a description of the approach of extracting non-terminal labels using the same.

E.7.4 EXCLUDE

Further information

on Siblings

They [TREF] and JBNU-CCLab (Lee and Na, 2022) achieved much higher performances thanks to SciBERT tokenizer because it is trained on scientific literature.