# FINECITE: A Novel Approach For Fine-Grained Citation Context Analysis

**Lasse Jantsch[1] Dong-Jae Koh[1] Seonghwan Yoon[1] Jisu Lee[1] Anne Lauscher[2*] Young-Kyoon Suh[1*]**

[1]School of Computer Science and Engineering, Kyungpook National University, South Korea
[2]Data Science Group, University of Hamburg, Germany
{lassejantsch, djkoh, shyoon0214, jisulee74, yksuh}@knu.ac.kr,
anne.lauscher@uni-hamburg.de

## 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 prevalence of vague definitions and restrictive 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 simple context conceptualization might be sufficient to assign intent or function classes, it fails to cover the rich information of scientific discourse. To address this concern, we analyze the context conceptualizations of previous works and, to our knowledge, construct the first comprehensive context definition based on the semantic properties of the citing text. To evaluate this definition, we construct and publish the FINECITE corpus containing 1,056 manually annotated citation contexts. Our experiments on established CCA benchmarks demonstrate the effectiveness of our fine-grained context definition, showing improvements of up to 25% compared to state-of-the-art approaches. We make our code and data publicly available.[1]

## 1 Introduction

Scientific research is inherently collaborative, with each discovery building upon a foundation of prior studies. To acknowledge previous work and provide credit, it is standard practice to include citations that connect past findings to new contributions. By embedding scientific progress and argumentation, citations serve a critical function that has been extensively examined—a research field known as citation context analysis (CCA) (Kunnath et al., 2022; Swales, 1986).

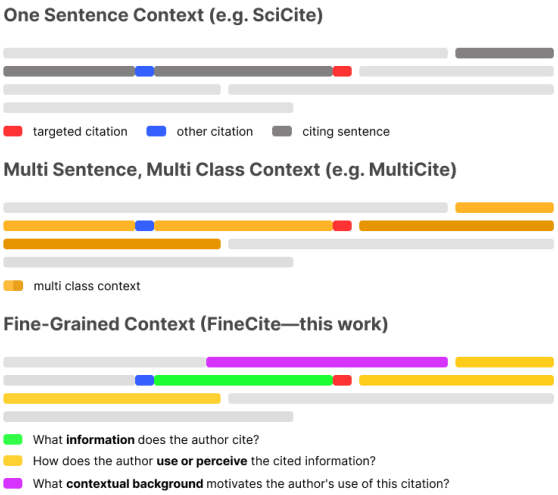



Figure 1: A visual comparison of different citation context conceptualizations in the CCA literature

In computational linguistics, CCA is mainly concerned with the automatic classification of citations along various dimensions, such as citation function (Lauscher et al., 2022; Cohan et al., 2019; Jurgens et al., 2018; Teufel et al., 2006), sentiment (Lauscher et al., 2017; Abu-Jbara et al., 2013; Athar and Teufel, 2012), or influence (Pride and Knoth, 2020; Cohan et al., 2019). Given a passage of text surrounding a citation marker—referred to as the *citation context*—one or more classes from a predefined citation classification scheme are assigned.

Although a considerable amount of research has explored different classification schemes and methods, the citation context has received relatively little attention. This lack of focus has led to an absence of a comprehensive definition and datasets with overly simplistic and coarse-grained citation contexts (Pride and Knoth, 2020; Cohan et al., 2019).

To address this concern, we analyze different context conceptualizations in previous work and provide a new comprehensive definition based on the semantic information of the citing text. A visual comparison is provided in Figure 1.

*Corresponding author
[1]https://github.com/lab-paper-code/FineCite



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 24525–24542
July 27 - August 1, 2025 

With our research and artifacts, we hope to spark new interest in the exploration of citation context information. Given the drastic capabilities of LLMs in zero-shot text understanding (Brown et al., 2020; Lewis et al., 2020; Vaswani et al., 2017), and the emergence of advanced language processing systems (Lewis et al., 2020; Edge et al., 2024), we argue that an improved understanding of contextual citation information is essential for improving interactive exploration of scientific argumentation.

Our contributions are the following:

- We analyze and formalize citation context conceptualizations in previous work.
- We propose, to our knowledge, the first fine-grained citation context definition based on the semantic structure of the citing text.
- We construct and publish the FINECITE corpus comprising 1,056 manually annotated fine-grained citation contexts.
- We evaluate our context definition in two experiments and demonstrate its effectiveness on established benchmarks.

The rest of the paper is organized as follows. The subsequent section reviews relevant literature in the field of CCA and provides a formalization of task and context conceptualization. Section 3 introduces our citation context definition. In Section 4, we describe the curation process of the FINECITE corpus and provide core statistics. In Section 5, we assess the effectiveness of our context definition in both context extraction and citation classification. Section 6 summarizes our contributions and outlines directions for future research.

## 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 learning a function $F_C$ representing the relationship of a citation context spanning $s \in S$ to a set of classes $c \in C$. The task can thus be formalized as

$$F_C(s) = \arg\max_{c \in C} P_{F_C}(c \mid s), \qquad (1)$$

where $P_{F_C}$ are the class probabilities emitted by $F_C$. The classes $C$ can represent various citation attributes, such as 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 CCA, refer to (Kunnath et al., 2022).

Despite the continued research in CCA, the introduction of new and larger datasets (Cohan et al., 2019; Jurgens et al., 2018), and updated methodology (Shui et al., 2024; Lauscher et al., 2022; Cohan et al., 2019), the simple modeling paradigm as described in Equation 1 prevailed. The popular SCICITE benchmark (Cohan et al., 2019) even further simplifies the task by reducing the commonly used six-class framework of (Jurgens et al., 2018) to a three-class schema. This simplicity leads to a low task complexity; however, it often fails to adequately represent the rich information present in the scientific texts (Lauscher et al., 2022). To capture a wider range of information, it is necessary to move beyond prevalent context span constraints and conceptualization on mutually exclusive classes. Table 1 compares the relevant research.

**Context Span Constraints.** The optimal context spans $S^*$ can be defined such that

$$S^* = \left\{ \arg\max_{s_i \in S_i} P_F(c_i \mid s_i) \mid i \in I \right\} \qquad (2)$$

where $S_i$ is the set of all possible context spans for one citation instance $i \in I$, and $P_F$ is the probabilities assigned by a function $F$ representing some relationship between $S$ and $C$.

As it is infeasible to solve Equation 2, previous work uses various assumptions to extract an approximate optimal context $\hat{S}^*$. The first common assumption is that $S^*$ can be 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 the whole paragraph (Teufel et al., 2006). Some approaches (Abu-Jbara et al., 2013; Athar and Teufel, 2012) allow for a non-context classification of context-window subsets, introducing a simple form of dynamic context spans. Only recent publications stress the importance of a fully dynamic approximation of $S^*$ (Lauscher et al., 2022; Nambanoor Kunnath et al., 2022) to conform to the situated structure of scientific argumentation.

| AUTHOR (YEAR) | ASPECT | NO. CLS. | EXCL | SEM | DYN | NON-C | SUB-S |
|---|---|---|---|---|---|---|---|
| Lauscher et al. (2022) | function | 7 | ✔ | ✘ | ✔ | ✔ | ✘ |
| Kunnath et al. (2022) | function | 6 | ✘ | ✘ | ✔ | ✔ | ✘ |
| Ferrod et al. (2021) | intent | 5 | ✘ | ✔ | (✔) | (✔) | ✔ |
| Pride and Knoth (2020) | purpose | 6 | ✘ | ✘ | ✘ | ✘ | ✘ |
| Cohan et al. (2019) | intent | 3 | ✘ | ✘ | ✘ | ✘ | ✘ |
| Jurgens et al. (2018) | function | 6 | ✘ | ✘ | ✘ | ✘ | ✘ |
| Abu-Jbara et al. (2013) | purpose | 6 | ✘ | ✘ | (✔) | ✔ | ✘ |
| Athar and Teufel (2012) | sentiment | 3 | ✘ | ✘ | (✔) | ✔ | ✘ |
| Abu-Jbara and Radev (2012) | - | - | - | ✔ | (✔) | (✔) | ✔ |
| Teufel et al. (2006) | function | 11 | ✘ | ✘ | ✘ | ✘ | ✘ |
| **FINECITE (this work)** | - | - | - | ✔ | ✔ | ✔ | ✔ |

Table 1: Structural comparison of previous work in computational linguistics on CCA (NO. CLS. = Number of classes, EXCL = Mutually exclusive labels, SEM = Semantic-based conceptualization, DYN = Dynamic context, NON-C = Non-contiguous context, SUB-S = Sub-sentence context)

The second common assumption is that $S^*$ stretches continuously from the citation marker. Even 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. They directly compared a non-contiguous context window with a smaller contiguous version and found that the former slightly outperforms the latter.

Thirdly, $S^*$ is often conceptualized with the sentence assumed to be the atomic unit of information (Cohan et al., 2019; Nambanoor Kunnath et al., 2022; Lauscher et al., 2022). In certain cases, however, this is not necessarily the case. Abu-Jbara and Radev (2012), for instance, shows evidently that sub-sentence segmentation is necessary to approximate $S^*$ for sentences with multiple citations. While their focus lies on the multi-citation setting, we also observe sub-sentence context granularity in other settings.

**Conceptual Restraints.** Next to the restrictive assumptions imposed on the context span, there are conceptual restraints limiting the expressiveness of citation contexts. In nearly all previous work, the context is conceptualized as

$$\hat{S}_C^* \approx \left\{ \arg\max_{s_i \in S_i'} P_{F_C}(c_i \mid s_i) \mid i \in I \right\}, \quad (3)$$

where

$$S_i' = \{ s_i \in S_i \mid \exists c \in C : F_C(s_i) = c \}. \quad (4)$$

This formulation captures that the context approximation $\hat{S}_C^*$ only contains spans $S'$ that have a clear association with a class in $C$. In other words, the citation context is conceptualized based on the classification schema represented through $F_C$ and not based on the semantic information of the text.

Most previous works additionally restrain their conceptualization by defining the relationship between $S$ and $C$ as mutually exclusive (Pride and Knoth, 2020; Cohan et al., 2019; Jurgens et al., 2018). This restricts the citation context further, as scientific discourse is faceted and can have multiple explanations (Lauscher et al., 2022). Lauscher et al. addressed this by creating the MULTICITE dataset, designed for multi-sentence, multi-function classification. They find that nearly one in five citations have at least two classes, with some reaching up to four. While this represents a step forward, it does not resolve the underlying limitation of defining citation context solely through the lens of the classification schema.

The only previous publication that defines a context based on semantic information from the vicinity of the citation marker is from Ferrod et al. (2021). They distinguish between the *citation object* and the *context*, where the former is the cited concept and the latter background information, or constraints on the *citation object*. While this goes in a similar direction to this work, their definition lags in completeness and only works on a subset of instances. To our knowledge, we are the first to propose a comprehensive citation context definition that is disjoint from the classification task.

## 3 Fine-Grained Citation Context

In this section, we propose and formalize our fine-grained context definition.

**Semantic Dimensions.** We base our context definition on previous research on argumentative structures in scientific texts. Teufel (2014) categorizes scientific argumentation along four principal dimensions: (i) statements about the author's own work (citing paper), (ii) properties of existing solutions (cited papers), (iii) the relationships between existing solutions and the author's contribution, and (iv) general properties of the research space. We apply this framework to the field of CCA and define the following three context dimensions.

The first dimension of the citation context is the information the citing author references from the cited paper. In the example

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

the phrase, "*the citation labeling scheme of <CITATION>*," describes here *what* information from the cited paper the author is referring to. This dimension highly correlates with (ii)—the properties of existing solutions, and is somewhat related to the *citation object* of Ferrod et al. (2021). In the following, we refer to this dimension as the *Information Dimension* (INF).

The second dimension describes the relationship between the citing and the cited work and corresponds to (iii) in Teufel (2014). 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 substantial fraction of occurring relations, this dimension also includes other forms of perception, such as comparison, evaluation, and judgment. In the following, 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 of *why* the author chose to include a citation.

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

Here, the reason the author included this citation is to emphasize a novel property of the citing paper's contribution. In Teufel's (2014) framework, this falls under the semantic class (i)—properties of the citing work—and is neither considered in the INF nor the PERC dimension. In other instances, such a motivating factor could be related to a property of the research space (iv) or other direct citations (ii, iii). We categorize these passages, which explain *why* a citation was included, as the *Background Dimension* (BACK) of a citation.

**Formal Definition.** To formalize our fine-grained citation context, we expand upon Equation 2 by removing the task dependency and incorporating semantic dimensions outlined above. Specifically, we define the task-independent, approximately optimal citation context $\hat{S}^*$ as:

$$\hat{S}^* := \{s_i^* \mid i \in I\}, \tag{5}$$

where

$$s_i^* = \{s_i \in S_i \mid \exists d \in D : F_D(s_i) = d\}, \tag{6}$$

$$D = \{\text{INF}, \text{PERC}, \text{BACK}\} \tag{7}$$

is the set of semantic dimensions defined in this Section, and $F_D$ represents the semantic relationship of the surrounding text to the citation.

We further formalize three structural properties of citation context spans $\hat{s^*} \in \hat{S}^*$:

- **Dynamicity:** The length $|\hat{s^*}|$ is dynamic and adapts to the situated structure of the enclosing argumentation.
- **Non-Contiguity:** $\hat{s^*}$ may consist of multiple disjoint spans allowing for skip-structured selection of relevant information.
- **Sub-Sentence Granularity:** $\hat{s^*}$ is constructed on sub-sentence granularity, enabling a fine-grained representation of the argumentative structure.

These properties collectively define a flexible and semantically motivated citation context that diverges from the constrained approximations of previous works. We provide a detailed empirical discussion of their relevance in Section 4.2 and 5.

## 4 FINECITE Corpus

Using the definition in Section 3, we create the FINECITE corpus. 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.** The FINECITE dataset was built from a subset of ACL Anthology Network Corpus (Radev et al., 2009). The ACL Anthology Network contains over 80K papers from several ACL conferences and other venues in computational linguistics. We extracted the full paper text, including citations, using GROBID (GROBID, 2024). Documents containing faulty meta-information, languages other than English, and miscellaneous documents with $<$3 sections and $<$5 references were excluded. From the remaining documents, we sampled 1,056 paragraphs, each containing one citation marker highlighted for annotation.

**Step 2: Guideline creation.** The annotation guidelines comprise best practices and rules for the context annotation. The instructions were created based on the definition presented in Section 3 and further iteratively refined to better handle ambiguous cases. For each iteration, several annotators completed five to ten tasks separately and subsequently discussed differences. Afterwards, the guidelines were updated to reduce ambiguity for the next iteration. In total, five iterations were performed. The complete Annotation Guidelines can be found in Appendix E.

**Step 3: Annotation.** The annotation was performed for each paragraph separately. The annotator was asked to read the paragraph carefully and annotate the context of the highlighted citation based on the guidelines. To provide further information in case of ambiguity, additional information, like the surrounding paragraphs and metadata about citing and cited papers, was provided in the annotation tool. A detailed description of the annotation platform is provided in Appendix A.

All annotators had previous experience with scientific literature and were carefully trained on the Annotation Guidelines. The compensation followed locally typical rates for research assistants.

**Step 4: Validation.** To ensure the annotation quality, we monitored several inter-annotator agreement (IAA) metrics on 10% of the annotations. We measured both F-measure commonly used for span annotations with open bounds (Hripcsak and Rothschild, 2005), and Cohens $\kappa$ (Cohen, 1960) for the agreement on label assignment above that expected by chance. To capture different aspects of the annotation process separately, we provide IAA for the whole context ($F1_{total}$), and for each scope separately ($F1_{inf}$, $F1_{perc}$, $F1_{back}$). The $F1_{macro}$ is the mean over $F1_{inf}$, $F1_{perc}$, and $F1_{back}$. While the metrics follow the standard definition, we provide a formal definition in Appendix B.

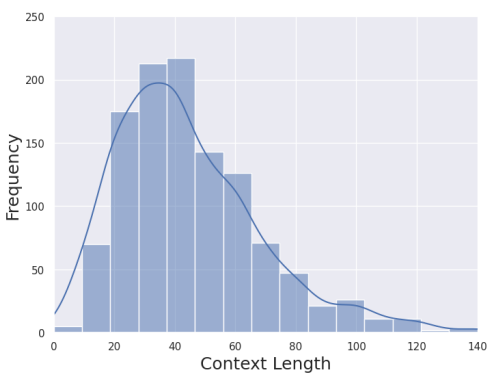
The $F1_{total}$ after the annotation process was 0.75, indicating an overall good agreement. The separate IAA on the context dimensions, however, is considerably lower. While the $F1_{inf}$ is with a score of 0.65 the highest, the $F1_{perc}$ is at 0.42 and the $F1_{back}$ at 0.34. The $F1_{macro}$ lies at 0.48 and the $\kappa$ on the validation samples was 0.55.

While these values are in the typical range for annotation of scientific literature (Lauscher et al., 2022; Ferrod et al., 2021; Lauscher et al., 2018), they highlight the task complexity. The moderate $F1_{macro}$, despite a rather high $F1_{total}$, indicates that while annotators often struggle to clearly distinguish between the dimensions, they have a good sense of what constitutes relevant information. PERC and BACK seem especially ambiguous.

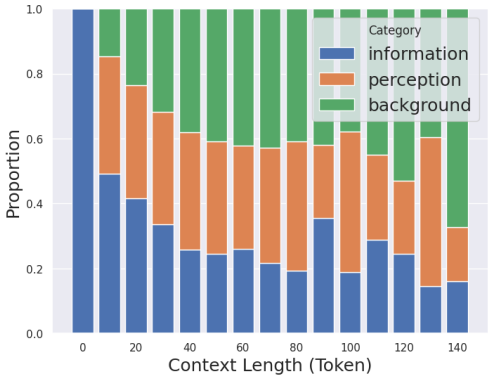
### 4.2 Corpus Statistics

The FINECITE corpus contains 1,056 fine-grained citation contexts for paragraphs from 72 scientific papers. Overall, INF accounts for 27%, PERC for 35%, and BACK for 38% of the annotated words. The average context length is $\sim$45 words and is approximately normally distributed with a long tail toward the upper end. The main contribution to the longer contexts is the BACK dimension. While BACK comprises about 8 words (30%) in contexts shorter than 40 words, it expands to an average of 54 words (43%) in contexts exceeding 100 words. Combined with the low agreement score on BACK, this might indicate that a clearer delimitation of the dimension might be helpful. Figure 2 provides an expanded visualization of the corpus statistics.

To evaluate the context span properties presented in Section 3, we apply context restrictions commonly used in prior work to the contexts in FINECITE and compare them to the fine-grained gold labels. We report F1-score and %-Match met-




(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.

| RESTRICTION | F1-Score | %-Match |
|---|---|---|
| One Sentence | 0.679 | 30.4 |
| Two Sentences | 0.716 | 23.6 |
| Four Sentences | 0.704 | 18.6 |
| Contiguous | 0.863 | 64.6 |
| Sentence Segments | 0.951 | 70.7 |

Table 2: Experiments on context restrictions compared to the gold context from FINECITE

rics for fixed-size windows of one, two, and four sentences, as well as for the longest contiguous sub-context and contexts extended to the next sentence boundary. We ignore dimension classes to highlight the structural properties and allow a two-word tolerance in the %-Match metric. Results are shown in Table 2.

Restricting the context to a fixed number of sentences results in a considerable error in both the F1-score and %-Match. The %-Match scores suggest that single-sentence contexts offer the best performance among fixed-size context windows; however, they fall short of capturing a majority of instances. 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. Surprisingly, the total F1-score error induced through sentence segmentation is relatively small. For the assignment of fine-grained dimension labels, sub-sentence segmentation is, however, a necessary 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 Experiments

In this section, we evaluate the FINECITE dataset on two tasks: (i) extraction of fine-grained citation contexts, and (ii) citation classification on standard CCA benchmarks using fine-grained context information extracted in (i).

### 5.1 Citation Context Extraction

Ensuring that common extraction models can reliably learn to identify citation contexts is crucial for the effective application of the presented fine-grained context definition.

**Data preparation.** We utilized the same samples used in the evaluation of the annotation process as the test set, with the remaining samples reserved for training and validation. We evaluate extraction on (i) uniform token labels and (ii) commonly used IOB (Inside–Outside–Beginning) labels.

**Extraction model.** For all extraction approaches, we use a SCIBERT (Beltagy et al., 2019) encoder model. SCIBERT is a BERT-like encoder-only transformer, pre-trained on scientific literature. To cover several common sequence extraction approaches, we evaluate three different classification heads: a linear, a Bi-LSTM (Hochreiter and Schmidhuber, 1997), and a conditional random field (CRF) (Lafferty et al., 2001) classifier.

**Experiment setup.** We used the pre-trained weights of SCIBERT from huggingface transformers (Wolf et al., 2020) and finetuned the whole model (encoder + cls-head) using AdamW (Loshchilov and Hutter, 2019) with a

| APPROACH | $F1_{total}$ | $F1_{macro}$ |
|---|---|---|
| *Inter Annotator Agreement* | | |
| Human (Annotation) | 0.75 | 0.48 |
| *Extraction Task* | | |
| SCIBERT w. Linear | 0.77 | 0.557 |
| SCIBERT w. BiLSTM | 0.759 | 0.56 |
| SCIBERT w. CRF | 0.787 | 0.521 |

Table 3: Extraction results on the FINECITE dataset

linear warm-up ratio of 5% and a peak learning rate of 5e-5. All models were fine-tuned using early stopping with a patience of three epochs, a batch size of 4, and a dropout rate of 0.1. To address class imbalance, we additionally applied weighted cross-entropy loss. The training was conducted on an NVIDIA A100 GPU. We evaluated the F1 scores described in Section 4.1.

**Result.** Table 3 shows the results of $F1_{total}$ and $F1_{macro}$. See Appendix D for extended results. We observe that all three extraction approaches reach higher F1 scores than those measured during the annotation process. The variance between the different classifiers is rather small. The CRF classifier exhibits the highest $F1_{total}$ score of 0.787, while the Bi-LSTM classifier dominates the $F1_{macro}$ metric with 0.56. The linear classifier achieves an $F1_{macro}$ of 0.557 and an $F1_{total}$ of 0.77, only slightly lower than the other approaches. The best results were achieved using IOB labels for linear and Bi-LSTM classifiers, whereas the CRF classifier worked better with uniform labels.

### 5.2 Citation Context Classification

To showcase the benefits of fine-grained contexts in a competitive setting, we provide a broad comparison with previous work using the citation classification task.

**Data.** We evaluate fine-grained contexts on four commonly used CCA benchmarks.

- **ACL-ARC** (Jurgens et al., 2018) comprises 1,933 labeled citations following a six-label classification schema introduced in the paper. All samples originate exclusively from the computational linguistics domain.
- **ACT2** (N. Kunnath et al., 2021) is a larger, mixed-domain collection with 4,000 annotated citations labeled with the schema used for the ACL-ARC dataset.
- **SCICITE** (Cohan et al., 2019), also multi-domains, contains 11,000 samples, annotated with a simple three-class annotation 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 citing sentence alone, we perform extraction on an extended window containing multiple sentences before and after the citation. SCICITE does not provide text exceeding the citing sentence, which drastically restricts the extraction of our fine-grained context.

To reduce the model's tendency to memorize author names, we conceal the targeted and other citations behind `<TARGET_CITATION/>` and `<CITATION/>` tags, respectively. Each dataset is divided into approximately 85% training and 15% testing. For the FINECITE approaches, we extract the fine-grained context using the extraction approach 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 instruction-tuned LLM for scientific literature. (i) and (ii) were fine-tuned on the training split, and (iii) and (iv) were evaluated in a zero-shot setting. The FINECITE approaches use SCIBERT(Beltagy et al., 2019) embeddings and a linear classification head similar to (ii). Instead of using CLS pooling, we use mean pooling over tokens belonging to the same dimension. The resulting dimension embeddings were concatenated and passed to the 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

| Approach | ACL-ARC | | ACT2 | | SciCite | | MultiCite | | Mean |
|---|---|---|---|---|---|---|---|---|---|
| | macro | st. dev. | macro | st. dev. | macro | st. dev. | macro | st. dev. | |
| *Baseline Approaches* | | | | | | | | | |
| 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 *Approaches* | | | | | | | | | |
| SciBERT (Linear) | 0.572 | 0.018 | 0.302 | 0.02 | 0.84 | 0.002 | 0.603 | 0.021 | 0.579 |
| 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 4: 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.

conducted on an NVIDIA A100 GPU. The optimal learning rate, batch size, and dropout for each dataset are provided in Appendix C. For all fine-tuned models, the performance was evaluated over five consecutive seeds.

**Result.** Table 4 exhibits the macro-F1 and standard deviation for each dataset. Detailed results including class scores are shown in Appendix D.

Among the baselines, SciBERT achieves the highest average macro-F1 (0.546), followed by the Scaffolds approach (0.453). Both GPT-4o (0.43) and SciTulu 70B (0.405) perform lower. These results show that finetuned encoder models have a considerably better conceptualization of the citation task than LLMs in a zero-shot setting. We further observe that the Scaffolds approach exhibits a high standard deviation on the ACL-ARC tasks, as it struggles to predict minority labels correctly on the smaller dataset.

The FineCite models introduced in this work outperform the baselines across all datasets. Among them, the context extracted with the Linear classification head achieves the best overall performance, with an average macro-F1 of 0.579. The context from the BiLST and CRF classifier only perform slightly lower with an average macro-F1 of 0.574 and 0.571, respectively. Comparing the performance on a per-dataset basis reveals a more nuanced pattern. The largest increase can be observed on the ACT2 dataset with a 25% increase over the strongest baseline, followed by a 13% increase on the ACL-ARC dataset. We explain the relatively low performance increases on MultiCite by considering that the dataset already provides a dynamic context, leaving limited advantage for fine-grained contexts. The performance on the SciCite benchmark further stresses that for the extraction of comprehensive fine-grained context, the citing sentence is not sufficient.

| Approach | ACL-ARC | | ACT2 | |
|---|---|---|---|---|
| | macro | st. dev. | macro | st. dev. |
| *Context Dimensions* | | | | |
| w/o Inf | 0.556 | 0.017 | 0.277 | 0.013 |
| w/o Perc | 0.563 | 0.019 | 0.259 | 0.036 |
| w/o Back | 0.56 | 0.019 | 0.253 | 0.024 |
| *Mean Pooling* | | | | |
| Dimensions | 0.584 | 0.014 | 0.302 | 0.02 |
| Weighted$_{tok}$ | 0.542 | 0.013 | 0.281 | 0.019 |
| Weighted$_{dim}$ | 0.573 | 0.015 | 0.28 | 0.015 |

Table 5: Ablation on context dimensions and pooling

Overall, the results demonstrate that the fine-grained citation context proposed in this work captures a more comprehensive citation representation than other conceptualizations in previous work.

**Ablation.** We provide ablation on the context dimensions, pooling method, and domain shift for a further analysis of the proposed fine-grained citation contexts. The dimension and pooling ablation were done on the ACL-ARC and ACT2 datasets. We create two new datasets ($ACT2$, $ACT2'_D$) for the evaluation on domain shift.

With the ablation on the citation dimensions (Table 5) we investigate the significance of the Inf, Perc, and Back dimensions for classification performance. Our analysis shows that removing any of the three citation dimensions leads to a performance drop for both datasets. While the decrease in performance on the ACL dataset is similar for all three dimensions, for the ACT2 benchmark Perc

| Approach | ACT2′ | ACT2′$_D$ | Δ | % |
|---|---|---|---|---|
| SciBERT | 0.345 | 0.228 | -0.117 | -33.9% |
| FineCite | 0.404 | 0.263 | -0.141 | -34.9% |
| Difference in Difference | | | -0.024 | -1.0% [2] |

Table 6: Ablation on Domain shift.

and Back exert greater influence. This highlights that despite the low extraction performance, Perc and Back contain essential information for the citation classification task.

The ablation on pooling strategies (Table 5) evaluates whether pooling citation dimensions separately improves performance over simpler alternatives. We compare this approach to token-weighted pooling, which ignores citation dimensions, and a dimension-weighted method. On both datasets, separate dimension pooling yields better results. Although the performance gap is modest, it indicates that modeling citation dimensions individually enhances representation quality, reinforcing the value of our context definition.

As the FineCite dataset only consists of samples from the computational linguistics domain, there might be a domain bias in the context extraction. To evaluate whether this compromises domain adaptation performance on the classification task, we provide an ablation on two new datasets (ACT2′, ACT2′$_D$) constructed from the multi-domain ACT2 benchmark (Table 6). The ACT2′$_D$ contains samples from computational linguistics and STEM domains in its training split, and samples from medicine and social sciences in its test split, thus evaluating domain adaptation. The ACT2′, on the other hand, contains samples from all domains in both splits while following the same split sizes. We provide the macro-F1 results on the test set for the strongest baseline and our approach, and analyze the difference-in-difference estimator between the two approaches.

For both models, we observe a substantial drop in performance when evaluated out-of-domain. Our approach retains a slightly larger margin, leading to a negative difference-in-difference estimate of -0.024. Despite this indicating that our model approach performs slightly worse on domain adaptation, the performance gains of using fine-grained contexts outweigh this drawback in overall effectiveness.

[2]Percentage Points

## 6 Conclusion

In this paper, we introduced a novel approach to defining citation contexts, 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. Subsequently, we described the curation of the FineCite corpus—a first resource for fine-grained citation contexts—and analyzed core statistics. Our experiments demonstrated that our context definition is practically applicable and leads to improved performance on established CCA benchmarks compared to state-of-the-art methods.

In future work, we will focus on expanding the dataset to a wider range of scientific texts and domains and further refining our context definition. Additionally, we plan to explore 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.

## Limitations

This work presents the first dataset of its kind, albeit with limitations in both size and domain coverage. The accompanying evaluation and analysis should be understood within this restricted scope and may not generalize to broader contexts. The objective is to establish a comprehensive definition of citation contexts and provide a resource and baseline for further analysis. Additionally, although our context definition is intended to be task-independent, our evaluation is limited to a subset of tasks due to constraints in space and resources.

## Acknowledgments

We thank Prof. Seonghyeon Lee for his insightful comments, and Hwanseong Joo and Jooyoung Yoon for their valuable feedback. We also acknowledge Joel Thomas, Minjoon Jin, and Jialun Zheng for their initial contributions. This study was supported by the NRF grant (RS-2018-NR031059) and the BK21 FOUR program (41202420214871), both funded by the Ministry of Education of Korea. Access to the A100 GPU used in the experiments was provided by the Department of Data Convergence Computing at Kyungpook National University, where Young-Kyoon Suh serves as an adjunct professor.

## 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.

Zeren Shui, Petros Karypis, Daniel S. Karls, Mingjian Wen, Saurav Manchanda, Ellad B. Tadmor, and George Karypis. 2024. Fine-tuning language models on multiple datasets for citation intention classification. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16718–16732, Miami, Florida, USA. Association for Computational Linguistics.

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 3 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 ($\kappa$), 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 7 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.

| Dataset | 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 7: Hyperparameters of each dataset

## D Extended Results

The following tables show extended evaluation results. Table 8 shows the extended extraction results on the FineCite dataset. Tables 9, 10, 11, and 12 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 of the 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*

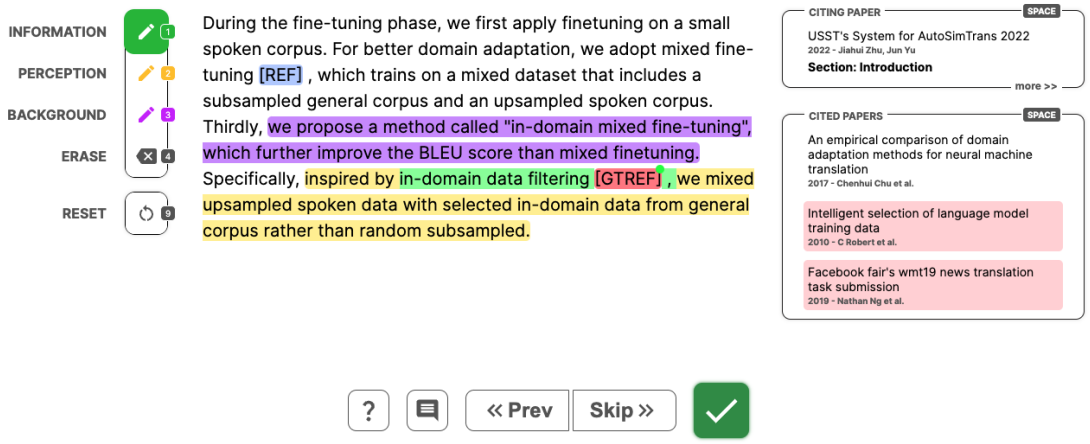



Figure 3: The Annotation Interface: Located on the left is the annotation toolbar, with the color-coded marker for each context scope, an ERASE tool, and the RESET button. The center is the working area where the annotation task is displayed and annotated. On the right side, meta-information regarding the citing and cited papers 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.

| APPROACH | $F1_{macro}$ | $F1_{total}$ | $F1_{inf}$ | $F1_{perc}$ | $F1_{back}$ |
|---|---|---|---|---|---|
| *Inter Annotator Agreement* | | | | | |
| HUMAN (annotation) | 0.483 | 0.758 | 0.654 | 0.416 | 0.338 |
| *Extraction Task* | | | | | |
| SCIBERT w. Linear | 0.557 | 0.771 | 0.755 | 0.495 | 0.422 |
| SCIBERT w. BiLSTM | 0.56 | 0.759 | 0.768 | 0.496 | 0.415 |
| SCIBERT w. CRF | 0.521 | 0.787 | 0.738 | 0.434 | 0.391 |

Table 8: 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 that 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 perception 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.

| 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 |
| *Baseline Approaches* | | | | | | | | | | | | | | | | | | | | | |
| 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 |
| FineCite *Approaches* | | | | | | | | | | | | | | | | | | | | | |
| 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 9: 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 |
| *Baseline Approaches* | | | | | | | | | | | | | | | | | | | | | |
| 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 |
| FineCite *Approaches* | | | | | | | | | | | | | | | | | | | | | |
| 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 10: Extended results of the citation classification task on ACT2.

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.
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.

| Approach | Backgr. | | | Method | | | Result | | | Macro | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F | P | R | F |
| *Baseline Approaches* | | | | | | | | | | | | |
| 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 |
| FineCite *Approaches* | | | | | | | | | | | | |
| 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 11: 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 | | | |
| *Baseline Approaches* | | | | | | | | | | | | | | | | | | | | | | | | |
| 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 |
| FineCite *Approaches* | | | | | | | | | | | | | | | | | | | | | | | | |
| 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 12: Extended results of the citation classification task on MultiCite.

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.

#### 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-judgmental 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 phrase-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 unclear).

### 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 have been 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.