# Dynamic Context Extraction for Citation Classification

**Suchetha N. Kunnath, David Pride, Petr Knoth**
Knowledge Media Institute (KMi)
The Open University
Milton Keynes
UK
{snk56, david.pride, petr.knoth}@open.ac.uk

## Abstract

We investigate the effect of varying citation context window sizes on model performance in citation intent classification. Prior studies have been limited to the application of fixed-size contiguous citation contexts or the use of manually curated citation contexts. We introduce a new automated unsupervised approach for the selection of a dynamic-size and potentially non-contiguous citation context, which utilises the transformer-based document representations and embedding similarities. Our experiments show that the addition of non-contiguous citing sentences improves performance beyond previous results. Evaluating on the (1) domain-specific (ACL-ARC) and (2) the multi-disciplinary (SDP-ACT) dataset demonstrates that the inclusion of additional context beyond the citing sentence significantly improves the citation classification model's performance, irrespective of the dataset's domain. We release the datasets and the source code used for the experiments at: https://github.com/oacore/dynamic_citation_context

## 1 Introduction

Understanding citation types has served a wide range of applications, including research evaluation (Jurgens et al., 2018), article summary generation (Nanba et al., 2000) and information retrieval (Valenzuela et al., 2015) to name a few. Classifying citation types according to their purpose or intent can make use of a variety of features, the most essential of which is the contextual textual fragment (context window) surrounding the citation marker within the citing article (Abu-Jbara et al., 2013; Jha et al., 2017). This information, also known as citation context, articulates how a cited work is presented in a research paper. Several citation type taxonomies of widely varying granularity have been used for citation type classification in the past (Kunnath et al., 2021). The taxonomy originally introduced by Jurgens et al. has been used across the two largest annotated datasets for citation typing, ACT (Pride et al., 2019) and ACL-ARC (Jurgens et al., 2018) and is shown in Appendix A.

Although evidence indicates that the size of the citation context window matters, there is not yet a consensus about its optimal size. While some researchers argue that multi-sentence context windows only add noise, thus confining their focus to the citing sentence alone (Dong and Schäfer, 2011; Cohan et al., 2019), others emphasise the need to incorporate longer citation context to avoid information loss (Abu-Jbara et al., 2013; Jha et al., 2017; Lauscher et al., 2021).

Most citation intent classification methods rely on a fixed-size contiguous citation context window (most typically one sentence) (Abu-Jbara et al., 2013; Hernandez-Alvarez et al., 2017; Nielsen et al., 2019), or a defined number of characters (Jurgens et al., 2018). Significant variation in contextual lengths however for each citation makes considering fixed context window size less desirable (Kunnath et al., 2021).

The use of a fixed citation context comes also with the risk of either the addition of noise (when the surrounding sentences have one or more citations) or loss of information (when the implicit citation context is beyond the static window size). Additionally, previous research shows that the document structure can influence the citation context window size, where it is more likely that context size is smaller for citations in the introduction section than in other sections, thus questioning the reliability of fixed citation contexts (Bertin et al., 2019b).

The use of adaptive longer than one sentence context methods for determining the optimal context span was also investigated by the earlier works (Rotondi et al., 2018). These methods involving supervised sentence classification require manual annotations for identifying the citation context boundary. Additionally, prior work on citation context

*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
November 20–23, 2022. ©2022 Association for Computational Linguistics

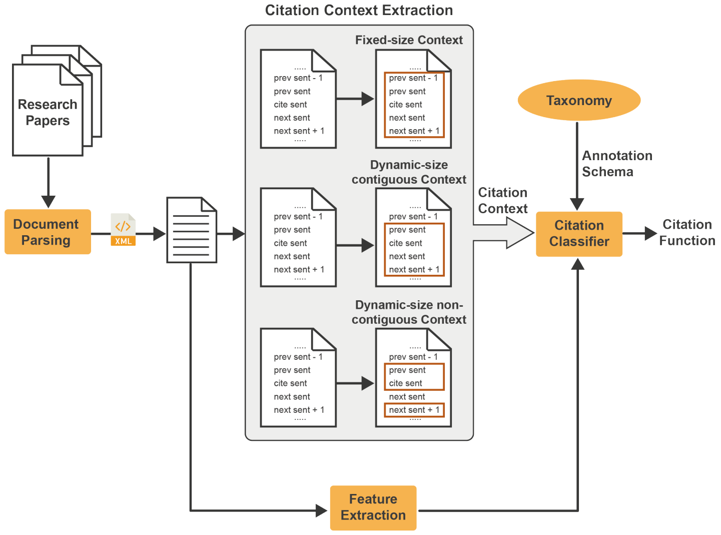


Figure 1: Citation classification pipeline.

extraction is mostly domain-centric, with many previous studies explicitly focusing on articles from computational linguistics. It was shown however in Harwood (2009) that citation behaviour of researchers differs across disciplines.

The goal of this study is to answer the following research questions:

RQ1: **To what extent does the performance of citation classification models vary depending on the size of the applied context window?**

Previous studies have not provided a definitive answer to this question. This is largely due to the results from previous studies not being comparable, as they use different datasets, type classifications and methodologies. Our work tests the effect of changing the citation context window size under the same experimental conditions, i.e. using identical state-of-the-art models; across two benchmark datasets, one multidisciplinary and one domain-specific. Accurately measuring this effect then enables us to measure the extent to which the citation intent classification performance varies depending on the context window size. Should we find that such difference is significant, this would motivate us to answer:

RQ2: **How can we create a dynamic-size context extraction model that adaptively identifies sentences in the vicinity of the citation marker that should be semantically part of a given citation context window?**

Such models would constitute a component that dynamically, i.e. adaptively for each citation marker, identifies the boundaries for a semantically coherent and complete citation context. The output of this component could be fed to the input of a citation intent classification model to increase its performance.

## 2 Related Work

Rotondi et al. (2018) categorise citation context determination strategies depending on the size of the context used as follows: (1) Fixed number of characters, (2) Citing sentence, (3) Fixed extended context and (4) Adaptive extended context. For automatic classification of citation functions, Jurgens et al. (2018) utilised fixed context size of 200 characters from either side of the citation, which was extracted using ParsCit (Councill et al., 2008), an open-source scientific document parser. The developers of the SciCite dataset (Cohan et al., 2019) on the other hand, noted that the addition of more context besides citing sentences resulted in the introduction of noise. Using sequence classification approach, Abu-Jbara et al. (2013) experimented with different citation context window sizes for citation purpose and polarity classification. The authors concluded that the best context span constituted the previous, citing and two following sentences.

Sequence classification approaches for context window detection use NLP-based features for identifying dynamic citation contexts. Kaplan et al. (2016) did extensive analysis on citation context

| Teams | Method Used | Context Used | macro f-score |
|---|---|---|---|
| IREL | SciBERT | citing sentence | **0.2670** |
| Duke Data Science | BiLSTM Attention + ELMo | prev sent,citing sent, next sent | 0.2590 |

Table 1: SDP 2021 3C shared task top models and citation contexts used

detection using a set of 35 features. The authors exploited the text coherence property and attained a performance boost by using discourse relation and citation location-based features. Based on the sentence polarity, Athar and Teufel (2012) categorised scientific text to extract implicit context. The primary assumption behind such a multi-class sentence classification system was that the authors are more likely to express their actual sentiment towards a citation, not in the citing sentence but in the sentences following. The findings from AbuRa'ed et al. (2018) shows the importance of features, direct citations and embedding similarity in implicit context detection.

The annotation guidelines of the existing dynamic context datasets require the annotators to choose implicit context from a fixed number of sentences before and after the citing sentence. Jha et al. (2017) introduced a manually annotated dataset, with sentences included using a fixed context window from citing sentences. The annotation guidelines for ACL Anthology Network corpus (AAN) based corpus developed by Xing et al. (2020) mention the need for choosing implicit citation context from three prior to, and three sentences following, the citing sentence. The new multi-intent (citation context annotated with one or more functions) domain-specific MultiCite dataset, developed by Lauscher et al. (2021), used co-reference and scientific entity mentions for manually annotating the dynamic context.

To establish a benchmark for citation classification allowing methods' comparison under the same experimental conditions, Kunnath et al. (2020); N. Kunnath et al. (2021) organised two rounds of the Citation Context Classification (3C) shared task. The shared task used multi-disciplinary author annotated dataset called Academic Citation Typing (ACT) dataset (Pride and Knoth, 2020; Pride et al., 2019). Compared to the first version of the classification task, the 2021 edition [1] saw a significant improvement in results primarily attributed to the application of deep learning-based models and features external to the manuscript in which the citation appears. Table 1 lists the top two systems with the used citation context window sizes and their achieved macro f-score. The winning team used citing sentence alone as input to SciBERT (Maheshwari et al., 2021). However, the runner-up team reported a further post-evaluation macro f-score improvement [2] by using additional fixed-size context beyond the citing sentence demonstrating the importance of the citation context window size for this task (Baig et al., 2021).

## 3 Methodology

Our experiments for RQ1 are designed to systematically test the performance of citation typing classification models on different fixed-size context windows. For this purpose we utilise a state-of-the-art model based on SciBERT (Beltagy et al., 2019), which is the highest performing system from the previous two 3C shared tasks (Kunnath et al., 2020, 2021).

Additionally, to understand the extent to which performance is impacted by the size of the citation context window, we evaluate a non-deterministic oracle approach. This approach assigns the correct label if at least one of the fixed window models make the right prediction. We extract several fixed-size contexts (Table 2), at a sentence level up to the maximum of a paragraph boundary. This boundary is motivated by studies of Kaplan et al. (2016) and Bertin et al. (2019a).

In RQ2, we address the limitations of the existing fixed-size context approach by exploring a new adaptive unsupervised approach for dynamically extracting citation context. As illustrated in Figure 1, there are two types of the dynamic-size context: (1) contiguous and (2) non-contiguous. Our extraction method utilises transformer-based scientific document embedding methods, SPECTER (Cohan et al., 2020) and SciNCL (Ostendorff et al., 2022) and features from the citing and cited article, in addition to the citing sentence. Finally, we evaluate the extracted dynamic context on citation function classification task using a sample of the multi-disciplinary ACT dataset (Pride and Knoth, 2020; Nambanoor Kunnath et al., 2022) and domain-specific ACL-ARC dataset (Jurgens et al., 2018).

[1] 22 teams participated in total at the SDP 3C Citation Context Classification shared task - https://www.kaggle.com/c/3c-shared-task-purpose-v2/leaderboard

[2] Team Duke Data Science

| Fixed Context | #Prev sentences | #Next sentences | Description | ABBREVIATION |
|---|---|---|---|---|
| $(sent_{cs})$ | 0 | 0 | citing sentence | FC1 |
| $(sent_{cs-1}, sent_{cs})$ | 1 | 0 | 1 previous sentence + citing sentence | FC2 |
| $(sent_{cs}, sent_{cs+1})$ | 0 | 1 | citing sentence + 1next sentence | FC3 |
| $(sent_{cs-1}, sent_{cs}, sent_{cs+1})$ | 1 | 1 | 1 previous sentence + citing sentence + 1 next sentence | FC4 |
| $(sent_{cs-2}, sent_{cs-1}, sent_{cs})$ | 2 | 0 | 2 previous sentences + citing sentence | FC5 |
| $(sent_{cs}, sent_{cs+1}, sent_{cs+2})$ | 0 | 2 | citing sentence + 2 next sentences | FC6 |
| $(sent_{cs-2}, sent_{cs-1}, sent_{cs}, sent_{cs+1})$ | 2 | 1 | 2 previous sentences + citing sentence + 1 next sentence | FC7 |
| $(sent_{cs-1}, sent_{cs}, sent_{cs+1}, sent_{cs+2})$ | 1 | 2 | 1 previous sentence + citing sentence + 2 next sentences | FC8 |
| $(sent_{cs-3}, sent_{cs-2}, sent_{cs-1}, sent_{cs})$ | 3 | 0 | 3 previous sentence + citing sentence | FC9 |
| $(sent_{cs}, sent_{cs+1}, sent_{cs+2}, sent_{cs+3})$ | 0 | 3 | citing sentence + 3 next sentences | FC10 |
| paragraph | | | Paragraph containing citing sentence | FC11 |

Table 2: Fixed context window sizes used and their descriptions

## 3.1 Datasets

### 3.1.1 ACL-ARC

The ACL-ARC dataset introduced by (Jurgens et al., 2018) uses citation contexts from computational linguistics, annotated for six citation functions. We used the pre-processed version of the ACL-ARC released by Cohan et al. (2019) a split of 85% (1,647 instances) for the training dataset and 15% (284 instances) for the test set. However, due to the significant amount of data leakage[3] and the presence of duplicates, we further cleaned this dataset. We divided the corpus based on the ACL Anthology ID, in such a way that none of the papers used in the training set were utilised by the development and the test sets, as recommended by Jurgens et al. (2018).

### 3.1.2 SDP-ACT

We also utilise the SDP-ACT dataset (N. Kunnath et al., 2021), which was released during the second 3C shared task. This dataset has 4,000 instances (3,000 training and 1,000 test) and is a subset of the largest multi-disciplinary dataset of annotated citations (Pride and Knoth, 2020).

ACT has been sourced from CORE[4] (Knoth and Zdrahal, 2012), a large continuously growing dataset of open access papers. The citation type categories in the dataset are similar to the ACL-ARC dataset(Jurgens et al., 2018), corresponding to the classes depicted in Appendix A. The citation context contains the textual fragment surrounding the citation marker, with the marker masked using the label, #AUTHOR_TAG as shown below:

*"A Decision Tree (DT) algorithm identifies patterns in a dataset as conditions, represented visually as a decision tree (#AUTHOR_TAG, 1986)."*

Note that several previous studies do not mask the citation marker containing the author tag. This subsequently leaks data from the train to the test set, leading to an artificially high model performance caused by over-fitting. The class distributions of the SDP-ACT dataset is in line with the ACL-ARC dataset, with most represented class being BACKGROUND (more than 50%).

## 3.2 Document Parsing

We used GROBID[5] for parsing the PDFs of the citing articles from the ACL-ARC and the SDP-ACT datasets. To ensure the length of the citation context is not more than one sentence, we further cleaned the citation contexts present in both datasets to match the parser's output from sentence segmentation feature. We manually extracted contextual information from papers in the case where citing articles could not be parsed, specifically for the ACL-ARC dataset.

## 3.3 Feature Extraction

Previous methods use discursive properties like text coherence (Kaplan et al., 2016), co-references (Bertin et al., 2019a) and topic mentions (Jebari et al., 2018) as signals for dynamic context extraction. In this work, we utilise semantic context similarity between citing and cited papers as a feature. For extracting citation context dynamically, we utilised the following attributes from citing and cited articles: (1) Cited Title, (2) Cited Abstract, (3) Citing Title and (4) Citation Context. To extract abstracts from the cited papers, we queried CORE[6],

[3] We noted that 49 instances from test set and 42 instances from dev set were already present in the training set.

[4] https://core.ac.uk

[5] https://github.com/kermitt2/grobid

[6] https://core.ac.uk/services/api

| Features Used | |
|---|---|
| **Cited Paper** | **Citing Paper** |
| Cited title | $sent_i{}^+$ |
| Cited title + Cited abstract | $sent_i$ |
| Cited title + Cited abstract | Citing title + $sent_i$ |
| Cited title + Cited abstract | Cited title + $sent_i$ |

$^+$ sentence in citing paragraph

Table 3: Feature vector combinations used for generating cited-citing document embeddings using SPECTER and SciNCL.

Semantic Scholar[7] and PubMed Central (PMC)[8] API's using the titles of the cited papers. For the SDP-ACT training and test set, we obtained cited abstracts for $2,697$ and $870$ instances. Similarly, we extracted $1,148$ and $185$ for the ACL-ARC train and test datasets.

### 3.4 Dynamic Context Extraction Method

Let $[.., sent_{cs-2}, sent_{cs-1}, sent_{cs}, sent_{cs+1}, sent_{cs+2}, ..]$ represent a contiguous set of sentences from a citing paper, with $sent_{cs}$ being the citing sentence. The relatedness of each sentence $sent_i$, preceding or following $sent_{cs}$, to the cited article is determined using document embedding similarity. To represent citing and cited articles, we use two transformer-based citation informed scientific document representations – (1) SPECTER (Cohan et al., 2020) and (2) SciNCL (Ostendorff et al., 2022). Both SPECTER and SciNCL build document representations from title and abstract of a paper.

We used several combinations of citing and cited features for generating our embeddings (Table 3), to test their suitability for dynamic context extraction. Our feature selection was motivated by Cohan et al. (2020) and Ostendorff et al. (2022), therefore we chose cited title and cited abstract for representing the cited paper. As our dataset contains several missing values for cited abstracts, we also tested a scenario with cited title alone for document representation.

Initially, the citing sentence alone or in combination with the citing or the cited title is used to represent the citing paper. Similarly, for representing the cited paper, we used one of the four attributes shown in Table 3. The cosine similarity between the two document embeddings determines the threshold for adding other neighbouring sentences. The process of determining the vector representation is repeated for each sentence, $sent_i$, that is preceding or succeeding the citing sentence, followed by the computation of the cosine similarity with the cited embedding. For dynamic non-contiguous citation context, any sentence with a similarity value greater than or equal to the threshold will be included in the dynamic context window. However, in the case of dynamic contiguous citation context, if any of the sentences in the previous or next context does not exceed the embedding similarity threshold, we terminate the search for more context beyond that particular sentence.

For both contiguous and non-contiguous contexts, we extract the preceding context, the following context and the combined context. Similar to the fixed context experiments, if the paragraph starts or ends with the citing sentence, the previous context and the next context will comprise of just the citing sentence.

### 3.5 Experimental Setup

For generating SPECTER and SciNCL document representations for the citing and cited papers, we used the source code from their respective GitHub repositories[9][10]. The missing cited abstracts were treated as empty strings, while presented as inputs for document representation. For all experiments, we chose an embedding sequence length of $512$. To extract abstracts from PuBMed, we used the python package, Biopython (Cock et al., 2009). Since the objective of this research is to analyse the effect of adding citation context dynamically on citation classification results, we chose only the highest performing system from the previous two 3C shared tasks (Kunnath et al., 2020, 2021), which was based on SciBERT (Beltagy et al., 2019). Best results were obtained using the following parameter values: drop out = $0.2$, learning rate = $1e-5$, batch size = $4$ and number of epochs = $5$.

## 4 Results

Tables 4, 5 and 6 show the results we obtained for the domain-specific ACL-ARC and the multidisciplinary SDP-ACT datasets for the fixed-size, dynamic-size contiguous and dynamic-size non-contiguous contexts. It also contains the theoretical performance boundary of the oracle.

From Table 4, we can see that on the single-domain ACL-ARC dataset, performance increases

[7] https://www.semanticscholar.org/product/api

[8] https://www.ncbi.nlm.nih.gov/home/develop/api/

[9] https://github.com/allenai/specter

[10] https://github.com/malteos/scincl

| Model | Fixed Context | ACL-ARC | | SDP-ACT | |
|---|---|---|---|---|---|
| | | Macro F-Score | Micro F-Score | Macro F-Score | Micro F-Score |
| SciBERT | $(sent_{cs})$ | 0.630* | 0.697 | 0.247 | 0.360 |
| | $(sent_{cs-1}, sent_{cs})$ | **0.653** | 0.718 | 0.255 | 0.421 |
| | $(sent_{cs}, sent_{cs+1})$ | 0.600 | 0.697 | 0.275 | **0.448** |
| | $(sent_{cs-1}, sent_{cs}, sent_{cs+1})$ | 0.647 | 0.725 | 0.236 | 0.409 |
| | $(sent_{cs-2}, sent_{cs-1}, sent_{cs})$ | **0.652** | **0.754** | 0.251 | 0.411 |
| | $(sent_{cs}, sent_{cs+1}, sent_{cs+2})$ | 0.627 | 0.718 | **0.284** | **0.447** |
| | $(sent_{cs-2}, sent_{cs-1}, sent_{cs}, sent_{cs+1})$ | 0.613 | 0.700 | 0.258 | 0.441 |
| | $(sent_{cs-1}, sent_{cs}, sent_{cs+1}, sent_{cs+2})$ | 0.590 | 0.693 | 0.260 | 0.444 |
| | $(sent_{cs-3}, sent_{cs-2}, sent_{cs-1}, sent_{cs})$ | 0.561 | 0.704 | 0.281 | 0.433 |
| | $(sent_{cs}, sent_{cs+1}, sent_{cs+2}, sent_{cs+3})$ | 0.576 | 0.679 | **0.287** | 0.445 |
| | paragraph | 0.564 | 0.641 | 0.224 | 0.366 |
| Oracle System | – | **0.831** | **0.894** | **0.560** | **0.743** |

* We noticed a 7.5% drop in score after removing data leakage.

Table 4: Results using different fixed citation context windows and their comparison with oracle system

| Dataset | Model | Features Used | Context used | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Macro | | | Micro | | |
| | | | prev | next | prev+ next | prev | next | prev+ next |
| ACL-ARC | SciBERT+ SPECTER | (cited_title) + $(sent_i)$ | 0.682 | 0.593 | 0.574 | 0.742 | 0.665 | 0.644 |
| | | (cited_title, cited_abstract) + $(sent_i)$ | **0.708** | 0.630 | 0.651 | **0.778** | 0.704 | 0.750 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.639 | 0.689 | 0.653 | 0.679 | 0.735 | 0.739 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.682 | 0.620 | 0.550 | 0.750 | 0.654 | 0.634 |
| | SciBERT + SciNCL | (cited_title) + $(sent_i)$ | **0.673** | 0.636 | 0.580 | **0.750** | 0.701 | 0.644 |
| | | (cited_title, cited_abstract) + $(sent_i)$ | 0.627 | 0.584 | 0.666 | 0.686 | 0.644 | 0.725 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.669 | 0.623 | 0.665 | 0.739 | 0.679 | 0.746 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.588 | 0.566 | 0.588 | 0.665 | 0.676 | 0.676 |
| SDP-ACT | SciBERT+ SPECTER | (cited_title) + $(sent_i)$ | 0.247 | 0.275 | 0.238 | 0.402 | 0.410 | 0.417 |
| | | (cited_title, cited_abstract) + $(sent_i)$ | 0.207 | 0.264 | 0.245 | 0.330 | **0.458** | 0.417 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.249 | 0.266 | 0.246 | 0.411 | 0.433 | 0.396 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.247 | **0.277** | 0.266 | 0.456 | 0.438 | 0.449 |
| | SciBERT+ SciNCL | (cited_title) + $(sent_i)$ | 0.267 | **0.285** | 0.267 | 0.446 | 0.445 | 0.406 |
| | | (cited_title, cited_abstract) + $(sent_i)$ | 0.259 | 0.274 | 0.252 | 0.421 | 0.441 | 0.402 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.282 | 0.246 | 0.263 | **0.471** | 0.435 | 0.430 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.266 | 0.271 | 0.253 | 0.466 | 0.439 | 0.453 |

Table 5: Dynamic contiguous citation context results on citation function classification

by adding the previous sentence to the citing sentence. However, on the multi-disciplinary SDP-ACT dataset, models perform well when using the immediate sentences following the citing sentence. In both cases, we can see that the theoretical performance boundary, represented by the Oracle approach, performs substantially better. This empirically shows high dependence of classification performance on the context window size, indicating a strong potential for improvement with the dynamic-size context approaches.

The results for the three context window approaches are as follows:

**Fixed-size context** – The highest macro and micro f-score for the ACL-ARC dataset is obtained by adding up to one or two previous sentences from the citing sentence. However, surprisingly, the performance drops when the subsequent sentences from the citing sentence are added to the citation context. This contrasts with the findings of Abu-Jbara et al. (2013) who previously reported that "...the related context almost always falls within a window of four sentences. The window includes the citing sentence, one sentence before the citing sentence, and two sentences after the citing sentence.." (Abu-Jbara et al., 2013, p. 599), where the authors performed experiments using papers from computational linguistics, similar to the ACL-ARC dataset. In the case of multi-disciplinary SDP-ACT corpus, the sentences from the next context proved to be more valuable for citation classification. The highest performance was reported when up to three

| Dataset | Model | Features Used | Context used | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Macro | | | Micro | | |
| | | | prev | next | prev+ next | prev | next | prev+ next |
| ACL-ARC | SciBERT+ SPECTER | (cited_title) + ($sent_i$) | 0.637 | 0.623 | 0.625 | 0.725 | 0.676 | 0.711 |
| | | (cited_title, cited_abstract) + ($sent_i$) | **0.684** | 0.613 | 0.614 | **0.764** | 0.683 | 0.683 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.626 | 0.568 | 0.683 | 0.679 | 0.616 | 0.750 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.660 | 0.594 | 0.576 | 0.725 | 0.661 | 0.647 |
| | SciBERT + SciNCL | (cited_title) + ($sent_i$) | **0.672** | 0.654 | 0.513 | **0.739** | 0.679 | 0.595 |
| | | (cited_title, cited_abstract) + ($sent_i$) | 0.646 | 0.603 | 0.505 | 0.704 | 0.658 | 0.602 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.609 | 0.555 | 0.586 | 0.679 | 0.641 | 0.704 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.622 | 0.641 | 0.516 | 0.655 | 0.718 | 0.669 |
| SDP-ACT | SciBERT+ SPECTER | (cited_title) + ($sent_i$) | 0.241 | 0.267 | 0.245 | 0.395 | **0.472** | 0.435 |
| | | (cited_title, cited_abstract) + ($sent_i$) | 0.243 | 0.273 | 0.239 | 0.392 | 0.448 | 0.404 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.249 | **0.284** | 0.258 | 0.435 | 0.459 | 0.433 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.263 | 0.259 | 0.236 | 0.424 | 0.465 | 0.414 |
| | SciBERT+ SciNCL | (cited_title) + ($sent_i$) | 0.280 | 0.263 | 0.262 | 0.505 | 0.452 | 0.456 |
| | | (cited_title, cited_abstract) + ($sent_i$) | 0.255 | **0.291** | 0.259 | 0.440 | **0.500** | 0.411 |
| | | (cited_title, cited_abstract) + (citing_title, $sent_i$) | 0.263 | **0.292** | 0.262 | 0.441 | 0.444 | 0.427 |
| | | (cited_title, cited_abstract) + (cited_title, $sent_i$) | 0.235 | 0.281 | 0.235 | 0.463 | 0.465 | 0.422 |

Table 6: Dynamic non-contiguous citation context results on citation function classification

sentences following the citing sentence were added to the fixed citation context. The experimental results across both datasets (Table 4) reveal that citation classification models benefit from additional context beyond the citing sentence, suggesting that the sentences surrounding the citing sentence frequently contain relevant information.[11]

**Dynamic-size contiguous context** – The similarity of embeddings from SPECTER, between the cited article title + abstract and the sentences from the paragraph produced the highest macro f-scores for both datasets. In the case of ACL-ARC dataset, the increase in macro f-score using the above system was nearly 8.5% in comparison with the highest fixed-size citation context. Contiguous context for SDP-ACT also obtained comparable scores. However, the highest micro f-score resulted from the previous context. In the majority of the cases, using bidirectional contexts is associated with lower model performance. This might be due to these contexts being too long, introducing unnecessary noise to the model.

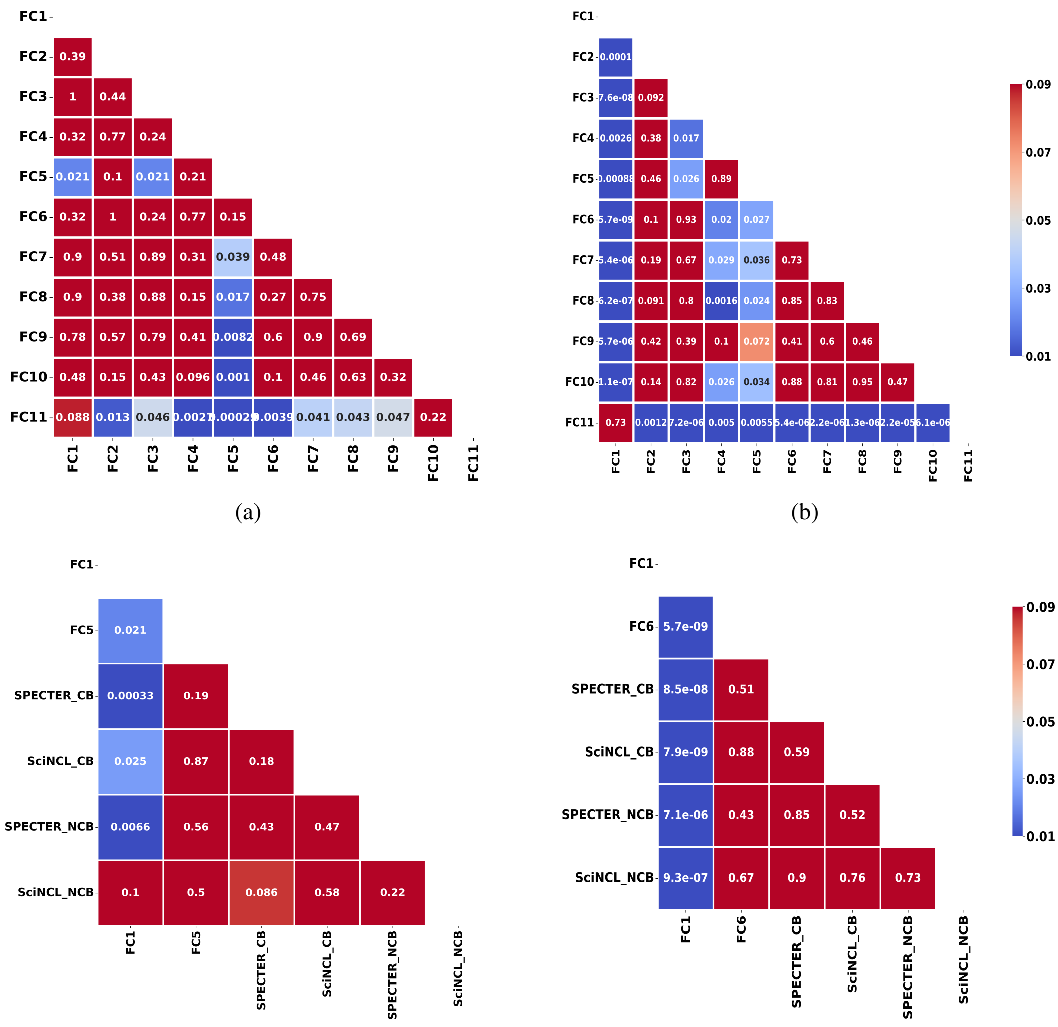
**Dynamic non-contiguous context** – The performance of the non-contiguous context on the ACL-ARC citation classifier falls by 3.4% when compared to its contiguous counterpart (Table 6). However, our non-contiguous approach outperforms the contiguous one on the SDP-ACT data, when used in conjunction with the SciNCL embeddings and the features - cited title, cited abstract and with or without citing title, with a 6% improvement in micro f-score. This validates our assumption that dynamic-size citation context approach has the potential to improve citation classification performance over fixed-size contexts and that there might be potential for further gains with the non-contiguous approach.

### 4.1 Ablation Study

We study the significance of different citation context windows using statistical McNemar's test ($p \leqslant 0.05$). Figure 2 represents the statistical significance scores for the different fixed-size as well as the best performing dynamic-size citation context spans on both datasets. For ACL-ARC, adding two previous sentences significantly improves classification scores in comparison to seven different context window sizes including the single citing sentence. Most of the fixed citation contexts, except ($sent_{cs}$) (FC1) and ($sent_{cs}, sent_{cs+1}, sent_{cs+2}, sent_{cs+3}$) (FC10) are significant when compared to the entire paragraph as context. For the SDP-ACT dataset, all citation contexts except the paragraph are significant with respect to citing sentence. This validates the need for contexts beyond the citing sentence, yet of a lower granularity than an entire paragraph.

Investigating dynamic-size context extraction, except the best non-contiguous citation context ex-

[11] For the SDP-ACT, we also extracted fixed number of words (10, 50, 100) from both sides of #AUTHOR_TAG. The results obtained for these citation contexts window sizes were in consistent with what we obtained for various fixed sentence windows. The highest score was obtained for 50 words (marco f-score: 0.28, micro f-score: 0.46).



(a)

(b)

(c)

(d)

Figure 2: Statistical significance on (1) ACL-ARC fixed contexts, (2) SDP-ACT fixed contexts, (3) ACL-ARC fixed and dynamic best contexts and (4) SDP-ACT fixed and dynamic best contexts. FC represents Fixed Context as shown in Table 2; CB and NCB are the Contiguous Best and Non-Contiguous Best

tracted using SciNCL (for ACL-ARC), all the highest scoring citation contexts from fixed-size and dynamic-size contexts are statistically significant when compared to the citing sentence. Despite the improvement in evaluation scores with respect to the best fixed-size citation context, the p-value indicates that the dynamic-size contiguous and non-contiguous models are not statistically significant. However, as one doesn't typically know what the best context size for a given dataset is, our unsupervised dynamic-size approaches remain valuable as they provide a statistically significant improvement over the typical scenario of relying on the citing sentence and do not require manual annotation of the citation context boundary.

## 5 Discussion

Citation type classification based on purpose reflects the author's citing intention and is therefore important for a wide range of applications, including research evaluation and scholarly document retrieval. Prior citation classification research has primarily been restricted to specific domains, notably computer science, computational linguistics and bio-medicine. This has severe drawbacks as methods developed for a singular discipline cannot capture the varying differences in citation practices across disciplines. This is why we conducted all our experiments on a domain-specific as well as on a multi-disciplinary corpora.

The outcome that adding further contexts beyond one sentence significantly improve results is important for further practice. As the optimal size of the citation context window for a given dataset is not known in advance, as can be seen from our experiments on the SDP-ACT and ACL-ARC dataset, there are two options: 1) to manually annotate the citation boundaries (which may be tedious) or 2) to apply a dynamic-size context extraction approach prior to feeding data into the citation type classifier. We argue that option 2 is well suited in situations where manual annotation of the boundaries is not available, which is the case on all current citation type datasets, except MultiCite (Lauscher et al., 2021), and whenever one needs to apply the model in practice across large volumes of citations.

One potential limitation of this work is the usage of a restricted set of contextual features for dynamic boundary detection. As a direction for future work, we would be interested in applying additional scientific features (both contextual and non-contextual) to further improve the dynamic non-contiguous method and verify the performance against the existing manually annotated MultiCite corpus (Lauscher et al., 2021). Also, the challenges involved in extracting features resulted in a considerable number of missing values for the cited abstract, which is another limitation of this paper. We believe employing additional sources for meta-data extraction might reduce the missing feature values in the future.

The ACL-ARC and SDP-ACT datasets used in these experiments were chosen for comparison due to their similarities, notably the usage of the six-way classification system. The most significant difference however is the range of domains from which the citation contexts are drawn. The ACL-ARC dataset uses data from just one domain, computational linguistics, whereas the SDP-ACT dataset is compiled from citations across 36 domains. The significant differences in the evaluation scores for the ACL-ARC and SDP-ACT datasets suggest that citation classification models trained on a specific domains are less effective when used to classify a multi-disciplinary dataset. This is an important direction for future work.

## 6 Conclusion

This work provides the first comprehensive study of the effect of different citation context window sizes on citation type classification performance. Our results on fixed-size contexts conclusively shows that using only the citing sentence, as it is common in previous work (Cohan et al., 2019), leads to lower performance than what can be achieved with longer citation contexts. Furthermore, our analysis of fixed-size context reveals that the optimal citation context size is domain-dependent. This emphasises the need for determining context dynamically. We therefore present the first unsupervised adaptive dynamic-size context extraction method for contiguous and non-contiguous context extraction. This method significantly improves performance of citation classification models compared to using the citing sentence only. The results from our performance boundary test using the oracle system suggest a large scope for further improvement which can be achieved in the future with the use of dynamic-size context extraction methods.

## Ethical Considerations

The datasets used for this research work do not contain sensitive information and we foresee no further ethical concerns with the work.

## Acknowledgements

This research received funding from Jisc under Grant Reference: 4133, OU Scientometrics PhD Studentship, covering the contributions of Suchetha N. Kunnath and Petr Knoth. Additional funding that contributed to the creation of the manuscript, covering the contribution of David Pride, was received from NRC, Project ID: 309594, the AI Chemist under the cooperation of IRIS.ai with The Open University, UK.

## 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.

Ahmed AbuRa'ed, Luis Chiruzzo, and Horacio Saggion. 2018. Experiments in detection of implicit citations. In *WOSP 2018. 7th International Workshop on Mining Scientific Publications; 2018 May 7; Miyazaki, Japan.[Paris (Francce)]: European Language Resources Association; 2018. 7 p.* ELRA (European Language Resources Association).

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American*

*Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601, Montréal, Canada. Association for Computational Linguistics.

Yasa M. Baig, Alex X. Oesterling, Rui Xin, Haoyang Yu, Angikar Ghosal, Lesia Semenova, and Cynthia Rudin. 2021. Multitask learning for citation purpose classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 134–139, Online. 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.

Marc Bertin, Pierre Jonin, Frédéric Armetta, and Iana Atanassova. 2019a. Determining citation blocks using end-to-end neural coreference resolution model for citation context analysis. In *17th International Conference on Scientometrics & Informetrics*, volume 2, page 2720.

Marc Bertin, Pierre Jonin, Frédéric Armetta, and Iana Atanassova. 2019b. Identifying the conceptual space of citation contexts using coreferences. In *4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) at the 42ndInternational ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 2414, pages 138–144. CEUR-WS. org.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

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.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Isaac Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: an open-source CRF reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 623–631, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Nigel Harwood. 2009. An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, 41(3):497–518.

Myriam Hernandez-Alvarez, José M Gomez Soriano, and Patricio Martínez-Barco. 2017. Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4):561–588.

Chaker Jebari, Manuel Jesús Cobo, and Enrique Herrera-Viedma. 2018. A new approach for implicit citation extraction. In *International conference on intelligent data engineering and automated learning*, pages 121–129. Springer.

Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R Radev. 2017. Nlp-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1):93–130.

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.

Dain Kaplan, Takenobu Tokunaga, and Simone Teufel. 2016. Citation block determination using textual coherence. *Journal of Information Processing*, 24(3):540–553.

Petr Knoth and Zdenek Zdrahal. 2012. Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12):1–13.

Suchetha N. Kunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. 2021. A meta-analysis of semantic classification of citations. *Quantitative Science Studies*, pages 1–46.

Suchetha Nambanoor Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 75–83, Wuhan, China. Association for Computational Linguistics.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *ArXiv*, abs/2107.00414.

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, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev, and Petr Knoth. 2022. Act2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3398–3406, Marseille, France. European Language Resources Association.

Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. 2000. Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1):117–134.

Boris Lykke Nielsen, Stefan Lavlund Skau, Florian Meier, and Birger Larsen. 2019. Optimal citation context window sizes for biomedical retrieval. In *CEUR Workshop Proceedings*, volume 2345, pages 51–63. CEUR Workshop Proceedings.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. *arXiv preprint arXiv:2202.06671*.

David Pride, Jozef Harag, and Petr Knoth. 2019. Act: An annotation platform for citation typing at scale. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL '19, page 329–330. IEEE Press.

David Pride and Petr Knoth. 2020. *An authoritative approach to citation classification*, page 337–340. Association for Computing Machinery, New York, NY, USA.

Agata Rotondi, Angelo Di Iorio, and Freddy Limpens. 2018. Identifying citation contexts: a review of strategies and goals. In *CLiC-it*.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190, Online. Association for Computational Linguistics.

# A Appendix

The following describes the classification schema first suggested by (Jurgens et al., 2018). The more fine-grained labels for the COMPARE_CONTRAST classification were first introduced by (Pride and Knoth, 2020)

| **Class Label** | **Description** |
|---|---|
| BACKGROUND | The cited paper provides relevant background information or is part of the body of literature. |
| USES | The citing paper uses the methodology or tools created by the cited paper. |
| COMPARE_CONTRAST<br>- similarities<br>- differences<br>- disagreement | The citing paper expresses similarities to or or differences from, or disagrees with, the cited paper. |
| MOTIVATION | The citing paper is directly motivated by the cited paper. |
| EXTENSION | The citing paper extends the methods, tools, or data of the cited paper. |
| FUTURE | The cited paper is a potential avenue for future work. |