Article:
Jiang, X. and Chen, J. (2023) Contextualised segment-wise citation function classification. Scientometrics, 128 (9). pp. 5117-5158. ISSN: 0138-9130

https://doi.org/10.1007/s11192-023-04778-3

Contextualised Segment-Wise Citation Function Classification

Xiaorui Jiang[1], Jingqiang Chen[2]
[1] Coventry University, Coventry, UK xiaorui.jiang@coventry.ac.uk
[2] Nanjing University of Posts and Telecommunications, Nanjing, China cjq@njupt.edu.cn

Abstract

Much effort has been made in the past decades to citation function classification, but noteworthy issues exist. Annotation difficulty resulted in limited data size, especially for minority classes, and inadequate representativeness of the underlying scientific domains. Concerning algorithmic classification, state-of-the-art deep learning-based methods are flawed by generating a feature vector for the whole citation context (or sentence) and failing to exploit the full realm of citation modelling options. Responding to these issues, this paper studied contextualised citation function classification. Specifically, a large new citation context dataset was created by merging and re-annotating six datasets about computational linguistics. A variety of strong SciBERT-based citation function classification models were proposed, and new states of the art were achieved. Through deeper performance analysis, this study focused on answering several research questions about the effective ways of performing citation function classification. More specifically, the study justified the necessity of modelling in-text citations in context and confirmed the superiority of doing citation function classification at citation (segment) level. A particular emphasis was placed on in-depth per-class performance analysis to understand whether citation function classification is robust enough to suit various popular downstream applications and what further efforts are required to meet such analytic needs. Finally, a naïve ensemble classifier was proposed, which greatly improved citation function classification performance.

Keywords

Citation context analysis; citation function classification; deep learning; SciBERT; ensemble

1. Introduction

Citation function[1] is "the author's reason for citing a given paper" (Teufel et al., 2006a), e.g., the cited paper is compared to "as a rival approach", used "as part of the solution", criticised "to justify the current research, or appraised "to motivate the current research" (Teufel et al., 2006b). It is part of the scientific argumentation about the "rhetorical function" of citations in scientific writing (Teufel, 2010). For instance, Ex. 1 in Table 1 illustrates a citation sentence (abbr. citance) with two in-text citations (citation hereafter) of different citation functions. "Prince et al., 1993" is a citation about the weakness of the cited work ("Weak"), while "Kisseberth 1970" is a neutral citation that merely acknowledges an existing study ("Neut").

Citation function classification (CFC), i.e., the recognition of the rhetorical functions of citations, is an important task of scientific text understanding with rich downstream applications (Ding et al., 2014). Significant progress has been made in the past two decades since the early work on automated CFC using rule-based algorithms (Garzone & Mercer, 2000; Nanba et al., 2000) and machine learning (ML) algorithms (Teufel et al., 2006b). The state of the art (SOTA) on the `ACL-ARC` dataset of a 6-class annotation scheme has been improved from around 55% macro F1 by feature engineering approaches (Jurgens et al., 2018) to about 70% by deep learning (DL) without domain-specific language modelling (Cohan et al., 2019) and 71% using the domain-specific large pretrained language model SciBERT (Beltagy et al., 2019).

However, several noteworthy issues exist in current automated CFC research, which motivated the current study. We will discuss them in the following (Sect. 1.1) and then summarise the main contributions we made (Sect. 1.2).

1.1. Issues of Citation Function Classification Research

We discuss the issues from three aspects: dataset annotation, algorithm development and practical application.

Issues with Citation Function Annotations and Datasets. The first issue is about annotation scheme. Many citation function annotation schemes were used, ranging from 12 classes (Teufel et al., 2006a) to only 3 classes (Cohan et al., 2019; Zhang et al., 2022). Sect. 2.1 gives a critical review in detail. Most existing studies worked on a particular annotation scheme, and there was little discussion of mapping between different annotation schemes. The second issue is about dataset size. Due to annotation difficulty, datasets are limited in size. Because dataset distributions are highly skewed, some important functions are very small. For example, the technical modification ("PModi") and ideational basis ("PBas") classes in Teufel et al. (2006a) both only have 60 instances, while in Jurgens et al. (2018), the extension class ("Extends" = "PModi" and "PBas" combined) only has 73 instances. A more extreme case is the criticism/weakness function, e.g., the "hed" class (criticism via hedging) in Hernández-Alvarez et al. (2017) only has 40 instances, and the "Weak" class in Su et al. (2019) only has 30 instances. This makes these datasets less feasible for training large deep learning models. If the annotation schemes of some datasets were cognitively

[1] Generally, we do not see a consensus on the terminology used by the community. Various synonyms of citation function exist, including citation role (Agarwal, et al., 2010), citation purpose (Abu-Jbara et al., 2013; Kunnath et al., 2020), citation intent (Cohan et al., 2019; Ferrod et al., 2021). They have been used interchangeably. Noteworthily, citation motive or citation motivation was not seen as equivalent to citation function. Indeed, Zhang et al. (2013) said citation content analysis "endeavors to describe the citing behavior itself, as well as to interpret and understand the underlying motives for the observed pattern. Namely, it seeks to understand what the purposes, functions, attitudes, dispositions, and sentiments behind the citing behavior are and how these patterns are represented in citations to indicate authors' motivations". Citation motivation was regarded by Zhang et al. as an umbrella of a few operationalised tasks including citation function classification, citation polarity classification, citation sentiment analysis, etc. Citation motivation was mainly used in qualitative studies by social scientists to analyse the subjective motive for authors to cite a paper, while citation function was mainly used in empirical studies by computer scientists with an emphasis, as Teufel (2010) stated, on the rhetorical function of citations in scientific argumentation. However, we also observed cases where citation motivation and citation function were interchangeably used, e.g. in the surveys by Tahamtan and Bornmann (2019; the "Citer motivation surveys or interviews" subsection), and Lyu et al. (2021; the "Scientific motivation" subsection). In addition, we observed that citation motivations and citation functions were often partially mappable. For example, the "Active support", "Active criticism" sub-categories of the motives to cite by Erikson and Erlandson (2014) are equivalent to the "PSup" (supporting) and "Weak" (weakness) functions by Teufel et al. (2006a), while most of their citation motives were not mappable to any meaningful function but could be fit to Teufel et al.'s "Neut" function.

mappable, at least partially, then it would theoretically allow the creation of a larger and more comprehensive citation function dataset by merging and potentially re-annotating part of them.

Issues with Citation Function Classification Algorithms. The first issue about algorithmic classification is that most DL approaches did either citance-level or context-level CFC because they generated one feature vector for either the whole citance or the whole citation context (Munkhdalai et al., 2016; Lauscher e al., 2017; Bakhti et al., 2018; Cohan et al., 2019; Su et al., 2019; Beltagy et al., 2019). This is conceptually flawed. Citance-level CFC will inevitably err on citances with multiple citations of different functions such as Ex. 1-2 in Table 1. Ideally, we do *citation-level CFC*, i.e., model and classify each citation separately. The second issue is that only a limited design space of citation modelling has been explored. When there are multiple citations, it is equally important to understand the meaning of each citation as understanding the meaning of the enclosing citance. Often, citation functions can only be decided in context. For example, the similarity citation ("PSim") in Ex. 2 in Table 1 is decided based on the context sentence S-2, and all "CoCoXY" citations (comparison or contrast between cited studies) in Ex. 3 are decided based on the meta-statement S-1. There are various ways of modelling citance and context representations. However, these modelling options have not been fully explored, nor the effective ways of combining citation, citance and context representations.

Issues with Practical Application of Citation Function Classification. It is plausible that one CFC model can suit and be applied to all downstream applications[2]. However, this may not work in practice. Take Teufel et al.'s difficult "CoCoXY" class for example. "CoCoXY" may be either confused with other comparison classes due to using similar comparative expressions (e.g., compare the italicized expressions in Ex. 3 and Ex. 4), or confused with "Neut" because neither class describes any relationship between the citing and cited papers (e.g., compare Ex. 5 to Ex. 3). Is it better to treat "CoCoXY" as a comparison class or as a neutral class as in Jurgens et al. (2019)? No matter what choice to make, there will be a seesaw effect on the CFC performances, i.e., the CFC performance improvement on one class usually causes performance drop on others. For example, Cohan et al.'s best model got worse performance on the neutral class than their second best model despite of significant improvement on the comparison class (Cohan et al., 2019, Table 5). Seesaw effects occurred on other citation functions too. Indeed, Cohan et al.'s best overall result was obtained at the cost of sacrificing the extension and motivation classes. We believe that a systematic analysis of the impacts of various modelling options on the CFC performances of different citation functions will give us useful insights into the potentials and pitfalls of applying existing CFC models to real-world applications.

1.2. Summary and Main Contributions of This Study

According to the analysis above, this study focused on one central question: What are the most effective citation modelling methods for high-performance citation function classification? Two subquestions arose for us to answer.

- *Q1: Should citation modelling be done in context and at citation level*?
- *Q2: What are the most effective ways of modelling and combining the representations of citation, citance and context?*

To answer these questions, this study comprehensively explored the design space of modelling the representations of citation, citance and context, made in-depth performance analysis of large-scale experiments using the trained CFC models, and built strong ensemble algorithms that achieved new SOTAs of CFC. Meanwhile, we hope that this study provides a good CFC benchmark to facilitate further research in citation context analysis and semantics-driven scientometric and bibliometric analysis based on citation context analysis.

The following summarises the main contributions of this study and accordingly the organisation of the paper. Concerning the issues with dataset annotation, we created a larger citation context dataset by merging and re-annotating six datasets in the

[2] This is what (almost) all existing studies did. Lauscher et al., (2021) can be said the only exception, where multi-label CFC was the focus, which essentially built, or can be seen as building, one classifier per citation function.

computation linguistics domain (Sect. 3). Concerning our central research question about the effective ways of citation modelling, we developed a series of strong DL-based CFC models based on SciBERT (Beltagy et al., 2019) by extensively exploring the options of encoding citance and/or citation context into the citation feature representation (Sect. 4). In addition to achieving new SOTAs of CFC, we made a detailed performance analysis of to derive the answers for the afore-mentioned two questions about the proper level of citation modelling (Q1) and the promising options for citation modelling (Q2) respectively (Sect. 5). Concerning how well the CFC models are applicable to real-world applications, we made an in-depth per-class performance analysis, concluded that there is no single best model that fits all, and discussed the strengths and weaknesses of existing DL-based CFC models in a wide range of scientific applications (Sect. 6). Following the per-class performance analysis, we proposed a naïve ensemble approach and obtained CFC models that significantly improved the overall performance and per-class performances.

Table 1. Citation Context Examples. All Come from Teufel's Annotations or Annotation Guidelines.

Example	**Citation Context**
Ex. 1: Two functions in one citance. From: https://aclanthology.org/W00-1804.	S-1. While Optimality Theory (OT) (Prince et al. 1993) [Weak] has been successful in explaining certain phonological phenomena such as conspiracies (Kisseberth 1970) [Neut], it has been less successful for computation. (...more weaknesses...)
Ex. 2: Context sentence S-2 is needed to infer the "PSim" citation in S-1. From: https://aclanthology.org/J00-1004.	S-1. Formalisms for finite-state and context-free transduction have a long history (e.g., Lewis and Stearns 1968; Aho and Ullman 1972) [PSim], and such formalisms have been applied to the machine translation problem, both in the finite-state case (e.g., Vilar et al. 1996) [Neut] and the context-free case (e.g., Wu 1997) [Neut]. S-2. In this paper we have added to this line of research by providing a method for automatically constructing fully lexicalized statistical dependency transduction models from training examples.
Ex. 3. The meta-statement of comparison or contrast like S-1 indicates the subsequent comparisons between cited studies ("CoCoxY"). Otherwise, all citations would be "Neut". From: https://aclanthology.org/C00-2175.	S-1 *However*, *different* sets of GRs are useful for different purposes. S-2 For example, Ferro et al. (1999) [CoCoXY] is interested in semantic interpretation, and needs to differentiate between time, location and other modifiers. S-3 The SPARKLE project (Carroll et al., 1997) [CoCoXY], *on the other hand*, does not differentiate between these types of modifiers. S-4 As has been mentioned by John Carroll (personal communication) [PSup], this is fine for information retrieval.
Ex.4: Meta-statement of comparison and contrast may apply to all "CoCo" classes (Comparison or Contrast). The meta-statement can appear very far from a particular citation it covers, say "Miller et al" in S-5. The meta-statement about "main parallels" also qualify "CoCo" over "PSim". "CoCoGM" means Comparison or Contrast in Goal and Method. From: Teufel (2010, pp. 434).	S-1 We will outline here the *main parallels* and *differences* between our method and previous work. S-2 In cooccurrence smoothing [Brown et al. 1993] (CoCoGM), as in our method, a baseline model is combined with a similarity-based model that refines some of its probability estimates. S-3 In Brown et al's work, given a baseline probability model P, which is taken to be the MLE, the confusion probability EQN between conditioning words EQN and EQN is defined as EQN and the probability that EQN is followed by the same context words as EQN. S-4 Then the bigram estimate derived by cooccurrence smoothing is given by EQN. S-5 In addition, the cooccurrence smoothing method sums over all words in the lexicon. [Miller et al] (CoCoGM) suggest a similar method... S-6 They do... (and so on …)
Ex.5: "CoCoXY" may be confused with "Neut" if the contrast or difference can be inferred but not explicit enough. From: https://aclanthology.org/A00-2009.	S-1 The line data was recently revisited by both (Towell and Voorhees, 1998) [Neut] and (Leacock et al., 1998) [Neut]. S-2 The former take an ensemble approach where the output from two neural networks is combined; one network is based on a representation of local context while the other represents topical context. S-3 The latter utilize a Naive Bayesian classifier.

2. Related Work

2.1. Citation Function Datasets and Annotation Schemes

The field of citation context analysis has used various names to describe the reasons for authors to cite references in scientific writing, including citation type, citation category, citation purpose, citation role, citation intent and citation function (Hernández-Alvarez & Gómez, 2016; Kunnath et al., 2021; Lyu et al., 2021). From a computational point of view, this study focuses on the annotation schemes and corresponding datasets in empirical studies which were more application oriented, i.e., the studies since the seminal work by Teufel et al. (2006b). Table 2 summarises them into three categories: 1) general-purpose citation function datasets (the majority part), 2) special-purpose datasets for a subset of citation functions or citation functions on specific scientific entities, and 3) a special type of datasets about citation importance (annotated per citing-cited paper pair). The "Fulltext" column says whether all citation contexts and all in-text citations of involved papers were annotated (marked by "•") or not (left blank). For the "Context (size)" column, the notation "[–*l*, +*r*]" specifies that the context consists of *l* and *r* sentences to the *left* and *right* sides of the citance respectively, while a "?" indicates the information is unclear from the paper and "variable" means the context can be of variable length according to user needs (only when full-texts are parsed and annotated). "OA" stands for "open access". The last column "Authoritative" indicates whether the annotations were done by the authors of the involved papers (marked by "•").

Most datasets provided citation contexts of certain lengths. Teufel et al. and Hernández-Alvarez et al. both annotated all citations in their full contexts (Teufel et al., 2006b; Teufel, 2010; Hernández-Alvarez et al., 2017). Dong and Schäfer (2011) instead annotated all citations only in their citances and replaced each citation string with a pair of empty parentheses as placeholder (Similarly, we used a pseudoword "CITSEG" to replace the consecutive segment of citation strings in our own dataset). Abu-Jbara et al. (2013), also Jha et al. (2016), and Su et al. (2019) used a window of sentences as context, while Jurgens et al. (2018) and Tuarob et al. (2020) used a window of words that were extracted and controlled by ParsCit[3]. Full context of any length is available in our own dataset, but in the experiments, we set $l = 2$ and $r = 3$. We found that the majority of cases could be covered by this setting. Lauscher et al. (2021) made a similar observation. Indeed, during annotation, we encountered a few exceptional cases when the minimal context must go far beyond [–2, +3] to correctly determine citation function. A representative example is Teufel's General Rule 27 about *meta-statements* of the "CoCo" (Comparison or Contrast) class (Teufel, 2010). A meta-statement may appear at the beginning of a paragraph to qualify all subsequent citations as "CoCo", some of which may be very far from the meta-statement, as Ex. 4 in Table 1 shows. For another example, to decide "PMot", Teufel's guidelines require annotators to skim-read the source paper to understand what approach/tool is used/extended to solve what problem, i.e., the *contribution sentences* defined in D'Souza et al. (2021). In this sense, context sentences for "PMot" can appear anywhere, although they will more likely occur in the Title, Abstract, Introduction and Conclusion sections. Therefore, we decided that full text availability was the first prerequisite for creating a citation context dataset.

[3] https://github.com/knmnyn/ParsCit

Table 2. Survey of Existing Citation Function Datasets

Dataset	Fields*	Size	Annotation Scheme	Fulltext	Context	OA	Authoritative
Teufel et al. (2006b, 2010)	CL	4022	Neut, Weak, CoCoXY, CoCoGM, CoCoR0, CoCo-, PSim, PSup, PMot, PUse, PModi, PBas	•	variable	•	
Agarwal et al. (2010)	BM	3491	Background/Perfunctory, Contemporary, Contrast/Conflict, Evaluation, Explanation, Method, Modality, Similarity/Consistency	•	[-1, +1]	•	
Dong and Schäfer (2011)	CL	1728	Level 1: Background, Compare, Fundamental Level 2: Background_GRelated, Background_SRelated, Background_MRelated, Compare, Fundamental_Idea, Technical_Basis	•		•	
Jochim and Schütze (2012)	CL	2008	Aspect 1: conceptual vs operational; Aspect 2: evolutionary vs juxtapositional; Aspect 3: organic vs perfunctory; Aspect 4: confirmative vs negational.	•	variable	•	
Li et al. (2013)	BM	6335	Based_on^{+}, Corroboration^{+}, Discover^{+}, Positive^{+}, Practical^{+}, Significant^{+}, Standard^{+}, Supply^{+}, Contrast$^{=}$, Co-citation$^{=}$, Neutral$^{=}$, Negative^{-} (+/=/-: positive/neutral/negative)	•	[-?, +?]		
Abu-Jbara et al. (2013) also Jha et al. (2016)	CL	2098	Neutral, Criticizing, Comparison, Substantiating, Basis, Use		[-1, +2]	•	
Hernández-Alvarez et al. (2017)	CL	3013	acknowledge, corroborate, weakness, hedge, useful, based	•	variable	•	
Jurgens et al. (2018)	CL	1954	Background, Compare or Contrasts, Motivation, Uses, Continuation (=> Extends), Future	partial**	•***	•	
Cohan et al. (2019)	CS, BM	11020	Background introduction, Method, Result comparison			•	
Su et al. (2019)	CL	1402	Neut, Weak, CoCo, Pos		[-1, +1]	•	
Kunnath, Pride, et al. (2020, 2021)	CS, BM	3000	Background, Compares_Contrasts, Motivation, Uses, Extension, Future			•	•
Pride and Knoth (2020)	various	11233	Background, Compare_Contrast (subclasses: similarities, differences, disagreement), Motivation, Uses, Extension, Future			?	•
Ferrod et al. (2021)	various	1380	Proposes, Analyzes (subclass: critiques), Compares (subclass: contrasts), Uses (subclass: dataset), Extends *Additional aspect: role – subj v.s. obj*			•	
Lauscher et al. (2021) *Multi-label annotation*	CL	12653	Background, Differences, Similarities, Motivation, Uses, Extends, Future Work		variable	•	
Zhang et al. (2021)	CL	9594	Relationship – Motivation, Comparison, Extension, Application; Content – Background, Method, Data, Result; Sentiment – Positive v.s. Negative	•	?	•	
Zhang et al., (2022)	CS, BM, plus CL	9645***	Cohan et al., (2019) enlarged with CL papers: Background introduction, Method, Result comparison			•	
Meyers (2013)	BM	291	Corroborate v.s. Contrast	?	[-?, +?]		
Zhao et al. (2019), Zheng et al. (2021) *Resource citation*	CL, ML, BM	3088	Use, Produce, Introduce, Compare, Extend, Other *Role: Material – Data; Method – Tool, Code, Algorithm; Supplement – Website, Document, Paper, Media, License*		[-2, +2]	•	
Tuarob et al. (2020) *Algorithm citation*	CS	8796	Level 1: UTILIZE v.s. NONUTILIZE Level 2: USE, EXTEND v.s. MENTION, NOTALGO		•****	•	
Jochim and Schütze (2012)	CL	2008	2-grade: organic v.s. perfunctory (citation-level)	•	variable	•	
Wan et al. (2014)	CL	~800	5-grade	N/A	N/A		
Zhu et al. (2015)	various	140+	2-grade: influential v.s. non-influential	N/A	N/A	•	•
Valenzuela et al. (2015)	CL	465	4-grade; 2-grade (important v.s. incidental)	N/A	N/A	•	
Qayyum and Afzal (2019)	CS	488	2-grade	N/A	N/A		•
This study *CITSEG- and citation-level annotation*	CL	4784/ 3854	11-/10-class: Future, Neutral, Weak, CoCoXY, CoCoGM, CoCoRes, Similar, (Support)*****, Motivation, Usage, Basis	•	[-2, +3] or variable	•	

* Field abbreviations: CL – Computational Linguistics; BM – Biomedicine; CS – Computer Science in general; ML – Machine Learning.
** Not all citations and not all citation contexts were annotated.
*** The original size, i.e., number of citation contexts, is 11965. We cleaned it to 9645 non-duplicate contexts. CFC is made on citation contexts rather than citations.
**** A context window of a certain number of characters around the citation were extracted by ParsCit's.context size is not in measured in sentence count.
***** 11- or 10-class depending on whether including a Support class or re-annotating Support into other categories

It is observable that very different aspects were annotated for biomedicine (BM) and computational linguistics (CL) or computer science (CS) domains, which are non-ignorable nuances if we want to map and merge different datasets. For example, BM had an obvious focus on scientific claims in biomedical publications, evidenced by the "confute/contrast" relationships, e.g., "Similarity/Consistency" v.s. "Contrast/Conflict" in Agarwal et al. (2010), and "Corroboration" v.s. "Contrast" in Li et al. (2013) and Meyers (2013). In addition, annotation schemes for BM are less consistent and mappable. Some citation functions are of special interest to biomedical scientists, like "Evaluation", "Explanation" and "Modality" in Agarwal et al. (2010) and "Discover^{+}", "Practical^{+}" and "Standard^{+}" in Li et al. (2013). On the other hand, citation function schemes for engineering science like CS or CL have been more or less "stabilised" to a 6-class scheme since Jurgens et al. (2018). Although Teufel et al.'s 12-class scheme (Teufel et al., 2006a; Teufel, 2010) may be the most cognitively plausible, the 6-class scheme is easier to understand and annotate by scientists who are not specialised in the area of citation context analysis. For the ease of re-annotation, our second prerequisite was that the annotation schemes of source datasets should be at least partially mappable (detailed in Sect. 3.1).

2.2. Citation Function Classification Algorithms

This section presents a critical review of the studies of computational algorithms for citation function classification (CFC) in the past two decades. Some good surveys exist (Hernández-Alvarez & Gómez, 2016; Iqbal et al., 2021; Kunnath et al., 2021). Among them, Kunnath et al. (2021) not only reviewed the preprocessing steps and feature engineering approaches when applying machine learning algorithms for CFC, but also made a meta-analysis of annotation schemes, datasets, benchmarks, and the state of the art of machine learning and deep learning methods on CFC.

The early attempt to building an automated citation function classifier was manually created based on rules called *pragmatic grammar*, which describes a number of handcrafted *cue lists* and certain syntactic constraints and relations around cue words (Garzone & Mercer, 2000). Similarly, by designing a set of 160 cue phrased-based rules, Nanba et al. (2000) developed a citation classifier for three classes, theoretical basis, gap or weakness, and other. In 2006, Teufel et al. provided the first comprehensive and, at the same time, operationalizable 12-class annotation scheme and a dataset suitable for machine learning algorithm (Teufel et al., 2006a, 2006b). The 12-class scheme followed a four-way distinction between citation motivations: Explicit statement of weakness of cited work; contrast or comparison with other work; agreement, usage, or compatibility with other work; and a neutral category (holding all cases that unfit other categories). They designed a large set of features capturing commonly seen cue phrases in expressing scientific ideas as well as the syntactic information around these phrases or the main verbs of the citation sentence, and applied IBk (Instance-Based k-nearest-neighbor classifier) for CFC (Teufel et al., 2006b). Annotation schemes proposed in follow-up studies were greatly simplified, but more or less mappable to Teufel et al.'s scheme (Dong & Schäfer, 2011; Abu-Jbara et al., 2013; Jha et al., 2017; Hernández-Alvarez et al., 2017; Jurgens et al., 2018; Su et al., 2019). Sect. 3.1 details the partial mappability between these annotation schemes and Teufel et al.'s scheme in order to build a larger citation context dataset.

Teufel et al.'s seminal works embarked a lot of research in this line by adapting their 12-class annotation scheme and adding or adjusting the syntactic features and lexical patterns around the manually collated informative cue-phrases for different classes (Agarwal et al., 2010; Dong & Schäfer, 2011; Li et al., 2013; Abu-Jbara et al., 2013; Jha et al., 2017; Hernández-Alvarez et al., 2017; Meng et al., 2017). Among this line, the work by Jochim and Schütze (2012) was special. They followed Moravcsik and Murugesan's quadchotomic approach to dividing citations into four dimensions (Moravcsik & Murugesan, 1975), including conceptual v.s. operational (akin to "Fundamental_Idea" v.s. "Technical_Basis" in Dong and Schäfer (2011)), organic v.s. perfunctory (akin to the meaningful v.s. incidental dichotomy in Valenzuela et al. (2015)[4]), evolutionary v.s. juxtapositional

[4] "The organic vs perfunctory facet – ORG-PERF – distinguishes those citations that form the underpinnings of the citing work from more cursory citations." (Jochim and Schütze, 2012, p. 1345.)

(i.e., “based on” v.s. “alternative to” cited work), and confirmative v.s negational. While this scheme is cognitively plausible, not all aspects appear in every citation. Jochim and Schütze (2012) also concluded on the significance of named entity features in CFC. The problem with previous studies is that they all provided their own schemes, datasets, algorithms but did not evaluate on the same benchmark. The SOTA result of feature engineering approaches was produced by Jurgens et al. (2018) with an easy-to-understand 6-class scheme (see Table 2). New features like topics of citation context, bootstrapped linguistic patterns around the citation, and PageRank rankings were introduced. This scheme was later used in the 3C (Citation Context Classification) shared tasks (Kunnath et al., 2020).

Recently deep learning approaches have been introduced to the CFC task. Earlier works applied Convolutional Neural Network (CNN; Lauscher et al., 2017; Bakhti et al., 2018), Bidirectional Long-Short Term Memory (BiLSTM; Munkhdalai et al., 2016) or CNN stacked over BiLSTM (Yousif et al., 2018) to encode and pool a feature representation from the citation sentence or context, which was then fed into a linear classifier or MLP (Multiple-Layer Perceptron) for classification. Pretrained word embeddings (Cohan et al., 2019; Roman et al., 2021) or contextualised language models (Beltagy et al., 2019; Maheshwari et al., 2021) were used to improve the understanding of citation contexts. Multi-tasking has been a recent trend, which transfers knowledge from semantically related tasks to improve CFC performance through joint learning on both the main task and supplementary tasks. Such tasks included sentiment classification (Yousif et al., 2019), citation worthiness prediction and section type classification (Cohan et al., 2019). A common issue with most existing deep learning solution is that they typically model the whole citation context or citance, instead of modelling individual in-text citations. For example, after encoding using CNN or BiLSTM, typically a max-pooling or self-attention operator was used to pool one single feature vector to summarise the meaning of the whole citation context or citance. While SciBERT achieved significant performance gains over previous DL approaches, Beltagy et al. (2019) only used the encodings of the sequence-level classification symbol “[CLS]” in their CFC experiments. As we mentioned in Sect. 1, this is conceptually flawed. The current study investigated the necessity and explored the promising ways of contextualised citation modelling for CFC, i.e., modelling individual citations in context. Sect. 5 presents more discussions.

A very closely related task, though not being our focus, is citation importance classification, i.e., identifying important, significant or meaningful citations. Wan & Liu (2014), Zhu et al. (2014), and Valenzuela et al. (2015) were the seminal studies embarking on the topic of citation importance classification, after which a lot of studies appeared (Hassan et al., 2017; Pride & Knoth, 2017; Qayyum & Afzal, 2019; Wang et al., 2020; Qayyum et al., 2021; Aljohani et al, 2021b). Citation importance classification can be seen as a special case of CFC with an extremely reduced annotation scheme, because citation importance in essence was defined based on citation function (Lu et al., 2014; Valenzuela et al. 2015). The difference is that CFC was done per each in-text citation, but all studies did citation importance classification on each citing and cited paper pair. Therefore, only paper metadata was used (Wan & Liu, 2014; Valenzuela et al., 2015). Full text features used were also primitive, such as cue phrases and textual similarities (Zhu et al., 2014; Hassan et al., 2018; Qayyum & Afzal, 2019; Ghosh et al., 2022). Deep learning approaches for this task typically suffered the same problem as in CFC that was mentioned in the paragraph above (Yousif et al., 2019; Aljohani et al., 2021b; Maheshwari et al., 2021).

3. Dataset and Annotation

3.1. Dataset Choices

We created a large citation context dataset by merging and re-annotating six available datasets in the computational linguistics (CL) domain. Initially, biomedicine (BM) domain was also considered because most BM papers are freely available in PubMed Central[5], but we found the BM datasets mentioned in Sect. 2 were either not publicly available or inaccessible any more. In addition, BM datasets focused on relationships between scientific claims (Agarwal et al., 2010; Li et al., 2013; Meyers, 2013) and their annotation schemes are less consistent and hard to map. Finally, we were able to obtain six publicly available citation function datasets of CL papers for re-annotation, namely `Teufel2010`[6] (Teufel et al., 2006a; Teufel et al., 2006b; Teufel, 2010), `Dong2011`[7] (Dong & Schäfer, 2011), `Jha2016`[8] (Abu-Jbara et al., 2013; Jha et al., 2016), `Alvarez2017`[9] (Hernández-Alvarez et al., 2017), `Jurgens2018`[10] (Jurgens et al., 2018), and `Su2019`[11] (Su et al., 2019). Aljohani et al. (2021a) made a similar effort, but they only merged `Teufel2010` and `Jurgens2018`, and used Jurgens et al.'s rules to map all instances to the 6-class scheme of the latter. In this sense, they did not do re-annotation. We believed that it would be better to re-annotate all datasets according to a set of consistent annotation guidelines for better data quality. Indeed, many instances of these datasets, even a small number of `Teufel2010`, were re-annotated ("corrected") according to our annotation guidelines that were extended from Teufel (2010).

Re-annotation was possible because the six annotation schemes were conceptually mappable, although they are often only partially mappable (summarised in Table 3). For example, the comparison functions "CoCoGM", "CoCoR0" and "CoCo-" defined in `Teufel2010` are merged into a single function in other datasets. "Technical_Basis" in `Dong2011` subsumes the "PUse" and "PModi" functions in Teufel2010, and "Fundamental_Idea" conceptually subsumes "PBas", "PMot" and "PSim". The "bas" function in `Alvarez2017` is equivalent to "Fundamental_Idea" and "Technical_Basis" combined, while "Pos" in `Su2019` moves instances about similarity to "CoCo". To conclude, we felt it feasible to re-annotate a large portion of each dataset. A notable benefit of re-annotation of six datasets is the wide time span of the merged dataset, ranging from early 1990s to late 2010s, which is believed to better reflect authors' patterns of placing citations and exhibit richer language expressions around citations.

3.2. Dataset Re-annotation

Our dataset, named `Jiang2021`, was created in three steps: dataset preparation, re-annotation and post-processing (Figure 1). To avoid distracting the readers, the details of the whole pipeline were moved to Appendix A. Three postgraduate research students in natural language processing were recruited to re-annotate (parts of) the six datasets with the first author (see Appendix A2 and A4 for more details). The re-annotation was done according to Teufel et al.'s 12-class scheme (Teufel et al., 2006a) plus a "Future" class for future work (Jurgens et al., 2018). See Table 4 for the annotation schemes we used. After re-annotation, we merged consecutive citation strings in each citance into a citation segment, represented by a pseudoword "CITSEG". For example, the citance "SHRDLU (Winogard, 1973) was intended to address this problem." would be tokenized and rewritten to "["SHRDLU", "(", "CITSEG", ")", "was", "intended", "to", "address", "this", "problem", "."]".

---

[5] https://pubmed.ncbi.nlm.nih.gov
[6] https://www.cl.cam.ac.uk/~sht25/CFC.html
[7] https://aclbib.opendfki.de/repos/trunk/citation_classification_dataset
[8] https://github.com/ivder/University-Project/tree/master (The "Citation Sentiment_Purpose Analyser/citation_sentiment_umich/" subfolder)
[9] http://rua.ua.es/dspace/handle/10045/47416

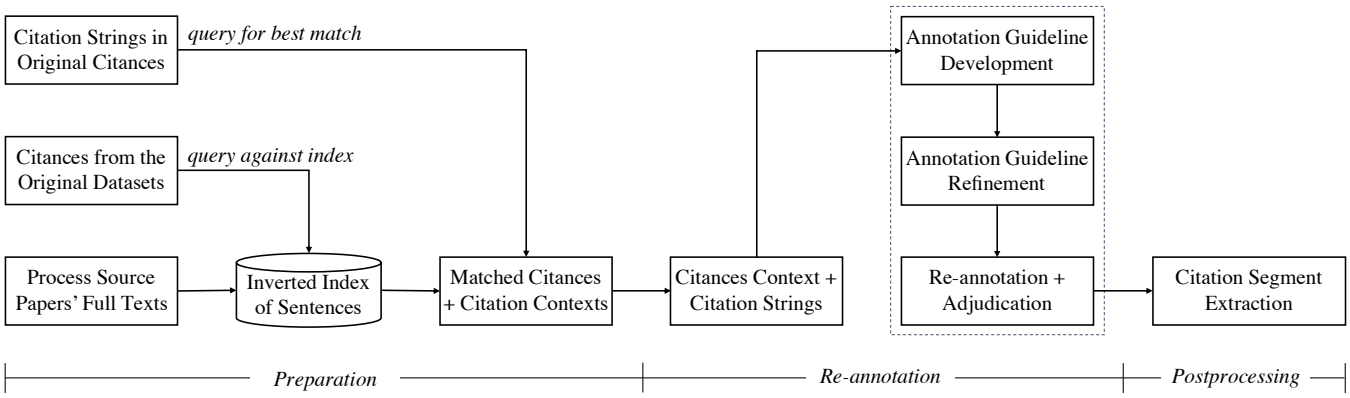


Figure 1. The Dataset Re-annotation Pipeline.

The first author was the main annotator (C_4) who lead the annotation guideline development and refinement, designed the re-annotation protocols, and guided the three annotators (C_1–C_3) in the whole process. In the annotation guideline development stage (Stage 1), we mainly worked on re-annotating Teufel's "PSup", "PSim", "PUse", "PMot", "PModi", "PBas" and "CoCoXY" instances to reach a consensus of understanding her annotation guidelines. Four annotators reached very high inter-annotator agreements (IAA) in term of Fleiss's κ (Fleiss, 1971; Davies & Fleiss, 1982): 0.8252 on the original 13-class scheme and 0.8529 on a condensed 9-class scheme. We must be extremely cautious when interpreting such extremely high agreements. We believe it mainly reflected the high quality of Teufel's annotation guidelines and manual annotations (Teufel et al., 2006a; Teufel, 2010). In the final re-annotation and adjudication stage, the pairwise IAAs between C_4 and C_1–C_3 were 0.6344, 0.6413, 0.5107 respectively on the 13-class scheme and 0.6758, 0.6714, 0.5350 respectively on the 9-class scheme. The overall IAA was about 0.6358 and 0.6614 on the 13-class and 9-class schemes respectively, which were regarded good (Artstein & Poesio, 2008, Table 8). We must remind that these figures only reflected how the four annotators agreed with each other during our special re-annotation process, which should be cognitively less complex than annotating from scratch, so these figures were not comparable with other studies. Appendix A.4 gives more details.

Context matters a lot for annotation. For example, instances about weakness often require reading a few context sentences ahead because a common scientific argumentation pattern is that the citance gives a neutral description while the following sentences point out its weakness. Full text is available in our dataset. A related issue is that there may be several functions appropriate for a citation depending on how we deem the citation in context, e.g., as "Neut" by only looking at the citance or "Weak" by overlooking at context sentences. In such situations, Teufel (2010) chose to let the stronger class overwrite or subsume the weaker, e.g., "Weak" overwrites "Neut". Following Teufel (2010), we defined "overwriting rules" for all citation functions, and always assigned the strongest function to each citation. A good example is the following rule from Teufel's annotation guidelines: "PMot" subsumes "PUse" if the *plausible usage* of something is justified by a *positive statement* (qualifying motivation). Note that, multi-label annotation (Lauscher et al., 2021) is an alternative solution, which was left to future work.

3.3. Statistics of the New Datset

In total, our dataset `Jiang2021` gathered 3356 citation contexts, 4784 in-text citations, and 3854 CITSEGs in total (Table 4). Compared to `Teufel2010`, most small classes like "CoCo-", "CoCoR0" and "PBas" were more than doubled, similar for

[10] https://github.com/davidjurgens/citation-function
[11] https://github.com/WING-NUS/citation_func_n_prov (We combined the "func" and "prov" portions of this dataset)

some other important functions like “PMot” and “PSim”. The most difficult class “PSup” was almost tripled. Because “PModi” and “PBas” were still too small, although much bigger than past datasets, we decided to merge them into “Basis” (equivalent to “Extends”). Then, “CoCo-” was split and re-annotated into either “CoCoGM” or “CoCoR0” due to its small size. These treatments resulted in our own 11-class citation function annotation scheme, which was mapped to 9-class, 7-class and 6-class schemes as Table 4 shows. They would be useful if the scientific analysis task did not require as fine-grained an annotation scheme as the 11-class scheme, e.g., when we did not distinguish between comparisons of methods and results. With different annotations schemes, we were able to run extensive experiments to justify the reasons for good or bad performances, and meanwhile able to find the best CFC models for a particular citation function of interest. Note that, during re-annotation, most of the time we did not find explicit comparison between a weakness citation and the citing paper, which is a requirement by Teufel’s guidelines to qualify a comparison function. Sometimes a weakness citation will be accompanied by a statement or hint that the citing study tries to overcome the weakness, but this is not frequent. Because weakness is an explicit description about the cited work, we decided that it might be more suitable to merge it into “Background” in our 7-class scheme.

Table 4. Citation function scheme mapping and CITSEG-level statistics of the re-annotated dataset

Teufel2010 (12+1 class)				**Jiang2021 (11-class)**			**Jiang2021 (9-class)**			**Jiang2021 (7-class)**			**Jurgens2018 (6-class)**		
label	**size**		**ratio**	**label**	**size**	**ratio**	**label**	**size**	**ratio**	**label**	**size**	**ratio**	**label**	**size**	**ratio**
	citstr	**citseg**	**citseg**												
Future	97	85	2.21%	Future	85	2.21%	Future	85	2.21%	Future	85	2.21%	Future	85	2.21%
CoCoXY	200	152	3.94%	CoCoXY	152	3.94%	Neutral	1615	41.90%	Background	1773	46.00%	Background	1615	41.90%
Neut	1924	1463	37.96%	Neutral	1463	37.96%									
Weak	223	158	4.10%	Weakness	158	4.10%	Weakness	158	4.10%						
CoCoGM	390	299	7.76%	CoCoGM	328	8.51%									
CoCo-	108	80	2.08%				Comparison	479	12.43%	ComOrCon	479	12.43%	ComOrCon	944	24.49%
CoCoR0	107	100	2.59%	CoCoRes	151	3.92%									
PSup	123	100	2.59%	Support	100	2.59%	Support	100	2.59%	Similar*	307	7.97%			
PSim	247	207	5.37%	Similar	207	5.37%	Similar	207	5.37%						
PMot	365	288	7.47%	Motivation	288	7.47%	Motivation	288	7.47%	Motivation	288	7.47%	Motivation	288	7.47%
PUse	794	755	19.59%	Usage	755	19.59%	Usage	755	19.59%	Uses	755	19.59%	Uses	755	19.59%
PModi	72	65	1.69%	Basis	167	4.33%	Basis	167	4.33%	Extends	167	4.33%	Extends	167	4.33%
PBas	134	102	2.65%												
Total	4784	3854			3854			3854			3854			3854	

* Cognitively, it was not plausible to merge “Support” into “Similar” because it has the meaning of “computational plug-in-ability”. Experiments proved this.

Table 3. Annotation Schemes, Statistics, and (Partial) Conceptual Mappings between Six Citation Function Datasets

Teufel2010			**Dong2011**				**Jha2016**				**Alvarez2017**			**Jurgens2018**			**Su2019**		
Type	#	%	Type		#	%	Type		#	%	Type	#	%	Type	#	%	Type	#	%
PSup[2]	46	1.14	**Background**[4] **- GRelated** General	Neu	953	55.15	**Substantiate??** *Unmappable!!*		126	6.01	Background **cor**roborate **ack**nowledge	0	0	**ComOrCon**	-	-	**Neut**	993	70.83
Neut Neutral description, or not fit into other classes	2398	59.6	**- SRelated** Specifics: method, parameter, …				**Neutral**		1283	61.15		982	32.59	**Future**	69	3.53			
CoCoXY Contrast between 2 cited methods	125	3.11	**- MRelated** Methods that may be usable								**deb**ate (0)	0	0	**Background**	999	51.13			
				Pos *PMot?*	149	8.62	**Criticising**	Pos *PMot?*	71	3.38	Use (**use**ful) *Partly PMot?*	857	28.44						
Weak Weakness	127	3.16		Neg	46	2.66		Neg	150	7.15	Critique - **wea**kness	141	4.68	**ComOrCon** (Compare or Contrast)	353	18.07	**Weak**	30	2.14
											- **hed**ge	40	1.33						
CoCo- Unfavourable contrast/comparison (against cited work)	62	1.54	**Compare**		70	4.05	**Comparison**		122	5.82	Contrast (con)	136	4.51				**CoCo**	90	6.42
CoCoGM[1] contrast/comparison in Goals or Methods	187	4.65																	
CoCoR0 comparison in Results	51	1.27																	
PSim similar	133	3.31	**Fundamental - (Fundamental)**		127	7.35					Use (**bas**ed) = PSim + PUse + PModi + PBas + PMot	491	16.30						
PMot positive about approach used or problem studied, as motivation for citing paper	131	3.26	**Idea** *+PSim*				**Basis**		74	3.53				**Motivation**	89	4.55	**Pos** Positive (usage)	289	20.61
PBas starting point	60	1.49												**Extends**	78	3.99			
PModi Adapt or modify tools, algorithms, data etc.	60	1.49	**- (Technical) Basis** *+PSim*		420	24.31													
PUse Use algorithms, tools, data and etc.	642	15.96					**Use**		272	12.96				**Uses**	366	18.73			
Total	4022				1728				2098			3013			1954			1402	

4. Citation Function Classification Algorithms

We designed a series of SciBERT-based DL models for citation function classification. The overall model architecture is shown in Figure 2. To perform segment-wise CFC, the pseudoword "CITSEG" was added to the vocabulary of SciBERT. SciBERT was used to encode the citation context. The CITSEG Encoder used the encodings of CITSEG as the *citation representation* **h**. According to Lauscher et al. (2021), more than 90% citation instances could be annotated based on the citance alone, so we defined the Citance Pooler to generate the *citance representation* **s**. To handle citations requiring multi-sentence contexts, the Context Pooler generated the *context representation* **c**. In this study, we fixed the context window to [-2, +3], i.e., two left and three right sentences. Indeed, Lauscher et al. (2021) showed that only a very tiny portion of citation instances need contexts larger than 6 sentences. The final feature vector **f** was the concatenation of these three parts, i.e., **f** = [**h**; **s**; **c**]. An MLP (Multiple-Layer Perceptron) was used for citation function classification. Citation representation was a mandatory component distinguish different citations in the same citance, but citance and context representations were optional. If only context representation was used, then **f** = [**h**; **c**]. On the contrary, we also tested only using citance representation, i.e., **f** = [**h**; **s**], to prove the indispensability of citation context.

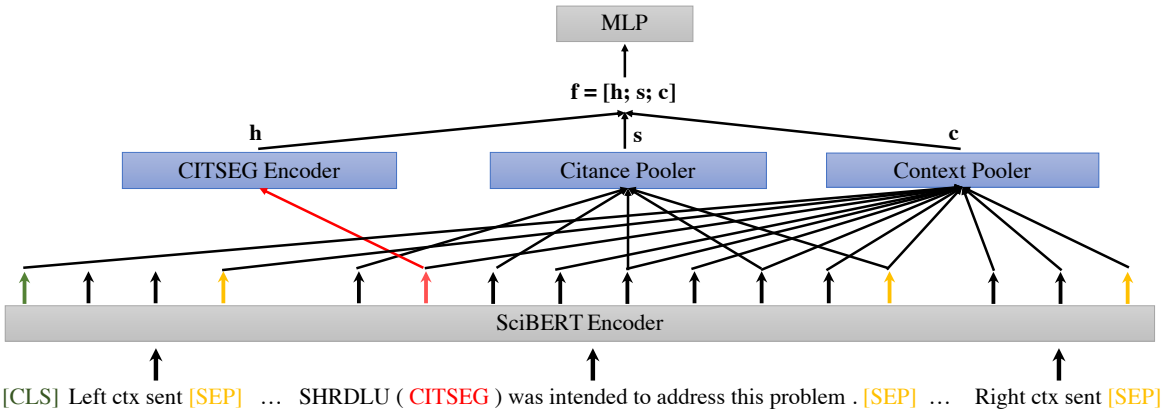


Figure 2. SciBERT-based Citation Semantics Analysis Model (Demonstrated Using a Hierarchical Context).

Following the BERT tradition, the token sequence of citation context was prepended with the sequence-level classification symbol "[CLS]" and appended with a sequence separator "[SEP]" to the end. Two types of contexts were tested: *sequential context* without inserting "[SEP]" to separate context sentences and *hierarchical context* with sequence separators inserted after each context sentence. For sequential context, citance representation was pooled from the tokens of the citance by applying a *citance mask* to the context, while context representation was pooled from all context tokens. We opted for two types of citance/context poolers: max-pooling (Eberts & Ulges, 2020) and self-attention (Munkhdalai et al., 2016). For hierarchical context, context representation was pooled from the representations of all enclosed sentences that were generated by a Sentence Pooler. In this case, "[SEP]" was used as the third option for pooling sentence representation.

In summary, the citation function classification model architecture was controlled by several options, as shown in Table 5 and subsequent tables. `Ctx_type` specified whether a sequential context (`ctx_type` = sequential) or hierarchical context (`ctx_type` = hierarchical) was used. `Citance` and `context` defined the citance pooler and context pooler respectively. Valid options included "max_pool", "self_attnd" or "X" (i.e., not used). Context pooler had the last option "[CLS]". With a sequential context, citance and context poolers generated feature representations from the tokens, therefore sentence encoder (the `sentence` option) did not apply ("N/A"). `Sentence` specified the sentence encoder in case of a hierarchical context. Valid options included "max_pool", "self_attnd" and "[SEP]". Finally, `citseg` specified whether CITSEG encoder was used

(i.e., `citseg` = O) or not (i.e., `citseg` = X). The former meant performing segment-wise CFC. The latter was purposed for simulating existing deep learning approaches which performed either citance-level or context-level CFC.

5. Results

5.1. Experimental Implementation

The models were implemented using HuggingFace's Transformers library[12] (version 4.2.2). The pretrained SciBERT model was downloaded from the official website[13] and the special token CITSEG was added to its vocabulary. The word embedding of CITSEG was randomly initialized and learned during the training process. Citation context was built in a "zig-zag" way, i.e., by first concatenating the right context sentence to the citance, then the left, and so on, until the context length reached the 512-token threshold of SciBERT. If citance alone exceeded the threshold (typically due to a failure in sentence segmentation), we centered the context window around the target CITSEG to include as many tokens as possible from both sides. Most SciBERT hyperparameters were unchanged. For self-attention, the attention dimensionality was fixed to 250. The hidden size of the MLP was twice the size of the feature vector (**f**). The AdamW optimizer was used with most parameters set to default. The learning rates for the parameters of SciBERT and MLP were initialised to 5e-5 (lr_pret: learning rate for the pretrained part, i.e., SciBERT) and 5e-4 (lr_cust: learning rate for the customised part, i.e., MLP) respectively. Different initial learning rates, like lr_pret = 5e-5, 1e-4, 2e-4, 5e-4, 1e-3, and 5e-3, were tested for the MLP but no significant difference was seen. The learning rate warmup ratio was fixed to 0.1. The batch size was fixed to 16 for training and validation. The experiments were run on a GeForce RTX 3080 GPU card with CUDA version 11.6. The samples of each citation function were randomly split into training (65%), validation (15%) and test (20%) splits and then merged. Each model was trained for a maximum of 20 epochs with five randomly generated seeds (5171, 13429, 25603, 32491, 47353). The "best" models were picked based on their validation performance. For each model variant, i.e., each combination of the modelling options for citation representation, citance representation and context representation, the best macro F1, average macro F1 and the standard deviation across 5 runs were reported.

5.2. An Additional Dataset and Two Baselines

For a fair comparison (explained below), we also compared with the most recent SciBERT-based contextualised CFC approach (Zhang et al., 2022) and experimented on an additional dataset that Zhang et al. extended from `scicite` (Cohan et al., 2019). The dataset was named `NI-Cite` (Native Information enhanced Citation dataset) by us. The `NI-Cite` dataset made the following extensions to `scicite` by (i) including adding a few thousand more instances from ACL papers to the latter, (ii) complementing each citance with one left context sentence and one right context sentence, (iii) enriching each citation context with a series of metadata, called "native information", such as the functional role of the enclosing section[14], the titles, DOIs and Web URLs of the citing and cited papers. The original dataset has 11195 citation contexts, each labeled with one citation function. During data preprocessing, we found there were a lot of errors and duplicates in the original `NI-Cite` dataset. Thus, we cleaned as many duplicates as detectible and removed as many errors as possible using our in-house scripts followed by manual check. Finally, there were 9645 citation contexts remained in the cleaned `NI-Cite` dataset[15]. Note that this dataset only has three functions: Background, Method and Results. The 3-class annotation scheme is used in Semantic Scholar[16],

[12] https://github.com/huggingface/transformers
[13] https://github.com/allenai/scibert
[14] Such as Abstract, Introduction, Method, Results, Conclusion
[15] Accessible through our GitHub fork of Zhang et al.'s code and data repository: https://github.com/xiaoruijiang/nativeinformation
[16] https://www.semanticscholar.org

technically backed by Cohan et al. (2019) and Beltagy et al. (2019), both coming from the Semantic Scholar group of Allen Institute for Artificial Intelligence. However, we believe this annotation scheme is cognitively incomplete, which limits its use in scientometrics.

The two baselines that were compared to both came from Zhang et al. (2022). We adapted their source codes for training the baselines. We found that the cross-references of a large portions of the original `NI-Cite` dataset were wrong, and so were the titles and DOIs of the cited papers, so we thought it was unreliable to use such information. In addition, the `Jiang2021` dataset only contains section titles but no manually annotated functional roles, so we only compared our methods to Zhang et al.'s two baseline methods which do not use metadata but only use citation context (See the last two rows in Table 5 and the last two columns in Figure 3 in Sect. 5.7). The first baseline "ni-cite w/ context" encoded the citance alone and used "[CLS]" for classification. The second baseline "ni-cite w/o context", considered one left context sentence, the citance, and one right context sentence. Zhang et al. used SciBERT to encode each sentence separately, pooled the citance representation and two context sentence representations using "[CLS]", and concatenated the three sentences' representations for classification. As such, Zhang et al.'s method actually did citance-level CFC. For this group of experiments, most hyperparameters were borrowed from the original implementations reported in their paper. However, we did grid search for learning rate (lr) and loss accumulation steps (acc_step) with the following ranges of values: lr in [1e-4, 5e-5, 1e-5, 5e-6], and acc_step in [1, 10, 20]. Similar to the experiments on `Jiang2021`, 20 epochs were run.

5.3. Citation Function Classification Results: Summary

Table 5 shows the CFC performances on the `Jiang2021` dataset. There are in total 36 model variants (models hereafter when the context is clear): models seq-01 to seq-12 and hie-01 to hie-24. We also ran preliminary experiments on citance-level CFC i.e., using citance alone. They are models cita-01 to cita-03, where cita-01 simulates previous studies based on SciBERT which used the sequence classification symbol "[CLS]" for classification (Beltagy et al., 2019; Varanasi et al., 2021)[17]. In addition, we also ran five more models which did contextualised encoding but used either citance or context representation alone for CFC, i.e., models seq-x07 to seq-x11. They are the CITSEG-agnostic counterparts (i.e., `citseg` = X) of models seq-07 to seq-11 respectively. The difference between seq-x07/x08/x09 with cita-01/02/03 is that the former encoded the whole context while the latter only encoded the citance. Finally, the difference between seq-x10 (resp. seq-x11) and cita-02 (resp. cita-03) is that the former encoded the citance in its context while the latter encoded the citance alone. To prove that citations should be encoded in context, we also tested CITSEG-only variants (i.e., model seq-12 and hie-24). The top-3 models (from seq-01 to seq-12 and hie-01 to hie-24) in term of best macro F1 were in **<u>bold underlined</u>**, **bold**, and <u>underlined</u> fonts respectively. If a top-3 model falls in seq-x07 to seq-x11, then it is highlighted in ***bold italic***.

On the 11-class annotation scheme, the best F1 was as high as 66.16% and the average F1 could reach more than 63.5%. Considering the cognitive complexity of the 11-class scheme, the performance figures were considered very promising. From the results on all four annotation schemes in Table 5, we observed a consistent trend that the more concise the annotation scheme, the better the overall classification performance. The best F1 was improved by 1.62% to 67.78% on the 9-class scheme, by about 6.65% to 72.81% on the 7-class scheme, and further to 74.03% on the 6-class scheme, a 7.87% absolute improvement from the 11-class scheme. Correspondingly, in term of average F1, the best performance was improved from around 63.5% on the 11-class scheme to about 71.2% on the 6-class scheme, an approximately 7.7% absolute improvement. These performance results could be deemed rather strong compared to two recent SOTAs: 67.9% by Cohan et al. (2019) and 70.98% by Beltagy et al., (2019). Concerning the latter SciBERT baseline, we will present more experimental results in Sect. 5.7. Note that the

[17] Models cita-01 to cita-03 are CITSEG-agnostic. This is however what most SciBERT-based SOTAs did. In fact, we also tested the CITSEG-aware versions: Encode the citance alone and set the feature vector as **f** = **h**, or **f** = [**h**; **s**] (citance = CLS, max_pool or self_attend). The performances were not good. The highest F1 was only around 59%, demonstrating the necessity of modelling citation context for CFC.

results were only indicative and not directly comparable because (1) they were obtained from our own dataset `Jiang2021`, and (2) we did CFC for each in-text citation segment, i.e., segment-wise CFC, but both SOTAs did citance-level or context-level CFC (and used different randomly generated seeds). Our dataset absorbed `Teufel2010` and `Alvarez2017`, which contain all citations in all sentences. On the contrary, `Jurgens2018` did not annotate all citations in a paper, even not all citations in the same citance. This might give citance-level CFC a small unfair advantage. Citance-level CFC is likely to stumble when seeing more citances with multiple citations of different functions. This claim was partially supported by the fact that the CITSEG-agnostic models seq-x07/x08/x09 performed consistently worse than all CITSEG-aware models on all annotation schemes, and models seq-x10/x11 were most of the time worse than their CITSEG-aware counterparts. Note that model seq-x07 simulated Beltagy et al.'s approach, so we are **confident to conclude the superiority of contextualised citation-level CFC**. More discussions are in Sect. 5.4-5.5.

5.4. Effectiveness of Citation-Level Classification

In this section, we try to answer part of research question *Q1: Should citation modelling be done at citation level?* In Table 4, models cita-01/02/03 reported the CFC performances by only encoding citance and using citance representation alone, i.e., no context sentences were considered. This is how Beltagy et al. (2019) reported their results on the `scicite` dataset. On the contrary, models seq-x07/x08/x09 encoded the surrounding context and used context representation alone for classification. This is how Beltagy et al. reported their results on the `Jurgens2018` dataset. A consistent phenomenon was observed across all four annotation schemes that the performances of the context-level CFC models, i.e., the CITSEG-agnostic models seq-x07/x08/x09, were worse than their context-level counterparts, i.e., the CITSEG-aware models seq-07/08/09 which used the pooled context representation to enhance citation representation. The poor performances of CITSEG-agnostic models justified our statement that it is conceptually flawed to use the summarised context representation for CFC. The citance-level models cita-01/02/03 got even worse performance than all other models we tested. From the above observations, we can partially conclude that **citation function classification should be done per citation rather than per citation sentence or context**.

However, when only citance representation was used to enhance citation representation this trend seemed to disappear. By comparing models seq-09/10 with seq-x09/x10, it appeared that the CITSEG-agnostic counterparts performed on par with the CITSEG-aware models. both won in some scenarios and the performances could be said close. It seemed that in certain cases, citance alone could provide strong enough signals for CFC. Indeed, Beltagy et al. (2019) reported that their first SciBERT baseline performed very well on `scicite`, which contains only a single citance for each sample and thus only allows CFC at citance level. From the above, it seems hard to draw a convincing conclusion. But, we can still observe the fact that all (top-2) best-performing models on all annotation schemes came from the family where citation was properly encoded in its context, e.g., models seq-08 and seq-06 on the 11-class scheme, models seq-12 and hie-08 on the 9-class scheme, models hie-14 and hie-19 on the 7-class scheme, and models seq-01 and seq-12 on the 6-class scheme (model seq-x10 is an exception; the fact that its avg F1 is not competitive implies that it might not be a very stable model). Therefore, it stills seems valid to conclude that **citation function classification should be done per citation rather than per citation sentence**.

Table 5. Citation Function Classification Performances on Different Annotation Schemes

Model	Model options					Macro F1 (%)											
	citseg	ctx_type	Encoding methods			11-class			9-class			7-class			6-class		
			citance	context	sentence	best	avg	std	best	avg	std	best	avg	std	best	avg	std
seq-01	O	sequential	max_pool	CLS	N/A	63.93	62.72	1.11	66.53	63.89	1.94	70.70	69.03	1.45	**74.03**	70.88	1.87
seq-02	O	sequential	max_pool	max_pool	N/A	63.21	62.61	0.45	64.84	63.60	1.08	71.39	68.13	1.89	70.23	68.25	1.60
seq-03	O	sequential	max_pool	self_attnd	N/A	64.26	62.82	1.04	65.61	63.66	1.91	70.19	69.24	0.64	70.99	68.86	1.71
seq-04	O	sequential	self_attnd	CLS	N/A	63.12	62.07	1.00	65.16	63.86	1.03	68.56	67.54	1.46	69.96	68.22	1.58
seq-05	O	sequential	self_attnd	max_pool	N/A	64.12	62.82	1.20	64.69	64.19	0.47	68.86	66.80	1.62	71.56	69.05	1.85
seq-06	O	sequential	self_attnd	self_attnd	N/A	**65.12**	63.05	1.60	64.84	62.52	1.48	70.63	69.16	1.43	72.19	69.81	1.37
seq-07	O	sequential	X	CLS	N/A	64.65	61.01	2.21	65.38	62.20	1.78	70.35	68.28	1.33	71.48	69.75	1.07
seq-08	O	sequential	X	max_pool	N/A	**66.16**	63.53	1.55	66.03	62.98	2.05	69.89	67.98	1.90	70.98	69.90	1.21
seq-09	O	sequential	X	self_attnd	N/A	63.92	62.80	0.89	65.41	64.18	0.75	70.80	69.78	0.85	71.91	69.66	1.47
seq-10	O	sequential	max_pool	X	N/A	63.93	62.72	1.11	66.19	63.72	2.74	69.16	67.87	1.85	71.89	70.18	1.77
seq-11	O	sequential	self_attnd	X	N/A	64.42	63.01	0.89	66.92	64.58	1.45	68.83	67.22	1.75	71.32	69.69	1.01
seq-12♣	O	sequential	X	X	N/A	64.93	63.50	1.04	**67.78**	64.74	1.88	70.65	69.28	1.30	**73.56**	70.22	2.44
seq-x07*	X	sequential	X	CLS	N/A	60.20	58.93	1.06	60.28	59.34	0.87	62.74	61.68	0.94	68.07	66.20	1.73
seq-x08	X	sequential	X	max_pool	N/A	59.54	57.89	1.40	61.36	59.34	1.68	63.97	62.81	1.18	65.56	64.43	1.15
seq-x09*	X	sequential	X	self_attnd	N/A	60.55	58.72	1.22	59.96	59.02	0.92	65.10	63.95	0.99	68.31	65.90	2.48
seq-x10	X	sequential	max_pool	X	N/A	64.09	62.23	1.70	65.04	63.62	1.6	68.68	67.85	0.62	***73.52***	69.31	3.12
seq-x11	X	sequential	self_attnd	X	N/A	64.38	62.46	1.13	***67.08***	64.21	2.38	69.34	67.31	1.90	69.48	68.85	0.59
cita-x01	X	citance	CLS	N/A	N/A	58.16	56.20	1.64	60.30	58.75	1.38	60.30	58.75	1.38	63.58	62.39	1.16
cita-x02	X	citance	max_pool	N/A	N/A	57.47	55.77	1.36	59.07	58.00	1.06	59.07	58.00	1.06	63.88	61.81	1.58
cita-x03	X	citance	self_attnd	N/A	N/A	59.49	58.13	1.11	56.99	56.01	1.17	56.99	56.01	1.17	62.54	61.51	0.95
hie-01	O	hierarchical	SEP	max_pool	SEP	62.78	61.76	0.89	65.39	63.24	1.40	69.18	67.35	1.50	69.39	68.42	1.25
hie-02	O	hierarchical	SEP	self_attnd	SEP	61.42	61.42	0.96	63.12	61.95	1.60	70.00	67.76	1.73	71.08	69.87	1.51
hie-03	O	hierarchical	max_pool	max_pool	SEP	63.30	63.30	1.12	65.39	63.24	1.40	69.18	67.35	1.50	71.71	69.60	1.36
hie-04	O	hierarchical	max_pool	self_attnd	SEP	63.79	63.79	1.71	63.12	61.95	1.60	70.00	67.76	1.73	72.10	70.25	1.69
hie-05	O	hierarchical	self_attnd	max_pool	SEP	63.69	63.69	2.21	64.96	62.95	1.50	67.77	66.39	0.84	70.09	67.83	1.74
hie-06	O	hierarchical	self_attnd	self_attnd	SEP	63.79	63.79	1.71	63.12	61.95	1.60	70.00	67.76	1.73	72.10	70.25	1.69
hie-07	O	hierarchical	max_pool	max_pool	max_pool	62.63	62.16	0.51	62.37	61.25	1.00	70.76	68.71	1.60	70.22	67.94	1.38
hie-08	O	hierarchical	max_pool	self_attnd	max_pool	65.02	62.10	2.24	**67.49**	64.51	1.97	69.53	67.47	1.73	69.77	68.24	1.33
hie-09	O	hierarchical	max_pool	max_pool	self_attnd	63.38	62.45	0.59	64.69	62.92	1.16	69.38	67.66	1.49	72.11	70.07	1.8
hie-10	O	hierarchical	max_pool	self_attnd	self_attnd	63.31	62.44	0.89	65.45	63.11	2.21	69.45	68.75	0.41	71.40	70.02	1.03
hie-11	O	hierarchical	self_attnd	max_pool	max_pool	64.46	62.17	1.99	64.00	62.80	1.62	68.76	67.09	1.50	72.38	69.33	3.07
hie-12	O	hierarchical	self_attnd	self_attnd	max_pool	63.43	62.26	0.83	65.36	64.28	0.97	69.66	68.27	1.60	70.78	69.56	1.57
hie-13	O	hierarchical	self_attnd	max_pool	self_attnd	64.99	63.56	1.15	64.97	63.97	0.80	70.10	67.99	1.88	71.49	69.52	1.66
hie-14	O	hierarchical	self_attnd	self_attnd	self_attnd	63.16	62.09	1.02	65.69	64.44	1.29	**72.81**	69.47	2.64	71.32	68.35	2.22
hie-15	O	hierarchical	X	max_pool	SEP	61.17	59.98	1.14	65.53	63.07	1.66	68.71	67.12	1.45	73.24	70.19	2.41
hie-16**	O	hierarchical	X	self_attnd	SEP	63.22	62.25	0.89	65.24	63.79	1.09	69.57	67.97	1.90	71.56	70.40	1.18
hie-17**	O	hierarchical	X	max_pool	max_pool	64.56	64.16	0.39	65.96	62.81	2.29	69.35	67.96	1.31	70.90	70.04	0.94
hie-18↑	O	hierarchical	X	self_attnd	max_pool	64.95	62.82	1.64	66.07	63.76	1.56	70.05	68.87	0.97	72.09	69.35	2.11
hie-19	O	hierarchical	X	max_pool	self_attnd	62.62	61.61	1.18	65.35	64.16	1.08	**72.39**	68.40	2.47	71.89	70.48	1.04
hie-20	O	hierarchical	X	self_attnd	self_attnd	63.15	62.39	0.60	66.25	63.79	1.97	70.88	69.54	1.10	70.72	69.75	1.1
hie-21	O	hierarchical	SEP	X	N/A	63.48	61.27	1.39	65.36	64.11	0.96	69.81	68.19	1.04	72.81	70.96	1.32
hie-22	O	hierarchical	max_pool	X	N/A	63.48	61.27	1.39	66.88	63.38	2.06	69.47	67.89	1.93	72.81	70.96	1.32
hie-23	O	hierarchical	self_attnd	X	N/A	62.55	61.09	1.05	64.60	61.89	1.56	68.82	66.55	2.23	70.38	69.28	1.19
hie-24	O	hierarchical	X	X	N/A	64.37	62.80	1.51	65.68	64.97	0.73	70.44	69.01	1.29	72.07	71.21	0.70
ni-cite w/o context			(Zheng et al., 2022)			51.94	(lr = 1e-5)		52.44	(lr = 1e-5)		58.47	(lr = 5e-5)		59.83	(lr = 5e-5)	
ni-cite w/ context						52.55	(lr = 5e-5)		51.99	(lr = 5e-5)		55.92	(lr = 5e-5)		60.27	(lr = 5e-5)	

* Models seq-x07 (note: no intermediate "[SEP]") and seq-x09 simulate the CFC approach in Beltagy et al. (2019) and Cohan et al. (2019) respectively.

** Difference between models hie-16/18 and seq-10/11: The former takes into consideration intermediate "[SEP]" symbols.

5.5. Effectiveness of Contextualised Encoding

Now we try to answer the second part of research question Q1: *Should citation modelling be done in context?* To answer this question, we should look at several aspects. Firstly, consider the citance-only CITSEG-agnostic models, i.e., cita-x01/x02/x03. The highest F1 was only around 59%. These performances were much worse than the models which encoded citance in context, i.e., seq-x07/x08/x09. This demonstrated the necessity of contextualised CFC, i.e., modelling citation context for CFC. Recall that the best feature engineering approach by Jurgens et al. (2018) got a 54.6% F1 while BiLSTM reported a 54.3% F1 on the `Jurgens2018` dataset (Cohan et al., 2019). Despite being not directly comparable, the results still proved the power of domain-specific contextualised word embeddings like SciBERT. Note that, we also ran the CITSEG-aware counterparts of models cita-x01/x02/x03 (not shown in Table 5), which encoded the citance and used citance representation to enhance citation representation (**f** = [**h**; **s**]). Their performances were very close to cita-x01/x02/x03, thus much worse than their contextualised counterparts seq-07/08/09. The results are not shown in the table as they are not our focus. Anyway, these results partially support our conjecture that **citations should be encoded in context**.

Secondly, there was a confirmative fact that, for all annotation schemes, the best models all encoded the whole context and used context representation (**c**) and/or citance representation (s) to enhance citation representation, i.e., the encodings of "CITSEG" (h). For example, the best model on the 11-class scheme seq-08 used the max-pooled context representation to enhance citation representation. The best model on the 6-class scheme seq-01 used "[CLS]" as the pooled context representation to enhance citation representation. For the 7-class scheme, hie-14 was the best model, which used both citance representation and context representation, for which the citation representation was pooled from the encodings of all citance words while the context representation was pooled from the representations of all context sentences. The only "exception" was the 9-class scheme, where the best model seq-12 only used citation representation. However, "CITSEG" was still encoded in its full context. In summary, we could argue that **citations should better be encoded in context**. This conclusion can be further supported by the fact that the contextualised citance-level models seq-x10/x11 reported unexpected strong performances but the context-agnostic counterparts cita-x02/x03 performed very poor.

5.6. Effectiveness of Citation Modelling Options

It is more complex for research question Q2: *What are the most effective ways of modelling and combing the representations of citation, citance and context?* It becomes very hard for us to draw meaningful conclusions. For different annotation schemes, the best encoding combination (of citance pooler, sentence encoder and context pooler) has to be determined case by case. **Using sequential context, "self_attnd" was most of the time a stronger context pooler than "max_pool" when context representation was used to enhance citance representation**, e.g., by comparing models seq-03/06 against models seq-02/05 respectively. However, we see that model seq-01, i.e., `ctx_type` = "sequential", `citance` = "self_attnd" and `context` = "[CLS]", was very strong across all four annotation schemes. This corroborates with the results of experiments on scientific named entity recognition, where this combination also produced highly competitive results (Eberts & Ulges, 2020; Jiang, 2021). It would be interesting to investigate more NLP tasks and more datasets to see whether this phenomenon was a coincidence or is a common law. In general, it is unfortunately unable for us to say whether "self_attnd" or "max_pool" is a better context pooler; more mixed behaviours happened to models seq-08/09 and seq-10/11, including models seq-x08/x09 and seq-x10/x11.

The same also applies to hierarchical context. No conclusions could be made to which is the better context pooler, "self_attnd" or "max_pool"; more mixed behaviours happened. The only regularity we find is that "self_attnd" worked better than "max_pool" as context pooler in the following setting: `citance` = "X" and `sentence` = "max_pool" (comparing models hie-18 against hie-17). Further, we are unable to conclude from Table 5 whether "self_attnd" or "max_pool" is a better sentence encoder; mixed behaviours happened across all annotation schemes. To see this clearly, we need to do a bit of re-arrangement of Table 5. See Table B1 in Appendix B, where we used upward arrow (↑) or downward arrow (↓) to indicate

performance gain or loss when changing one option while fixing the others, a pair of upward and downward arrows (↑↓) to indicate mixed behaviours in terms of best F1 and avg F1 or equal performances, and a yellow trèfle (♣) to indicate the cases where "[SEP]" performed the best as sentence encoder. For the latter case, it is also difficult to conclude whether "[SEP]" is a good sentence encoder. However, what we can confirm is that "[SEP]", as sentence encoder, sometimes brought competitive performances, such as models hie-15 on the 6-class scheme, and hie-04/06 on both the 7-class and 6-class schemes (See the yellow trèfles in Table B1). It would be interesting to investigate if we could further pre-train "[SEP]" to make it a better sentence encoder, for example, by following the pre-training paradigm used for long document extractive summarisation (Xu et al., 2020).

Concerning citance pooler, still we are unable to answer which is better, "self_attnd" or "max_pool". To see this better, we need to re-arrange the rows about models seq-04/05/06 against seq-01/02/03 respectively. See Table B2 in Appendix B. The only regularity we find is that **"max_pool" worked better than "self_attnd" as citance pooler together with "[CLS]" as context pooler**. This strengthens our conjecture on the compatibility between the settings `citance` = max_pool" and `context` = "[CLS]". Similarly, there were mixed behaviours with hierarchical context. The only regularity occurred when `context` = "X", where "max_pool" outperformed "self_attnd" as citance pooler. Overall, "max_pool" outperformed "self_attnd" in more cases as citance pooler (Table B2). Finally, we may be able to conclude that, across both sequential and hierarchical contexts, **it is NOT effective to integrate citance representation alone with citation representation**, when the latter is properly encoded in its context (comparing model seq-12 against seq-10/11, model hie-24 against hie-21/22/23). Often, the best (top-2) models were either the models using context representation to enhance citation representation or the models integrating both context and citance representations, with only two exceptions from model seq-12 on the 9-class and 6-class schemes. Anyway, even in these two exceptional cases, the citance tokens were contextually encoded too. This **re-iterates the importance and usefulness of encoding citation in its context**.

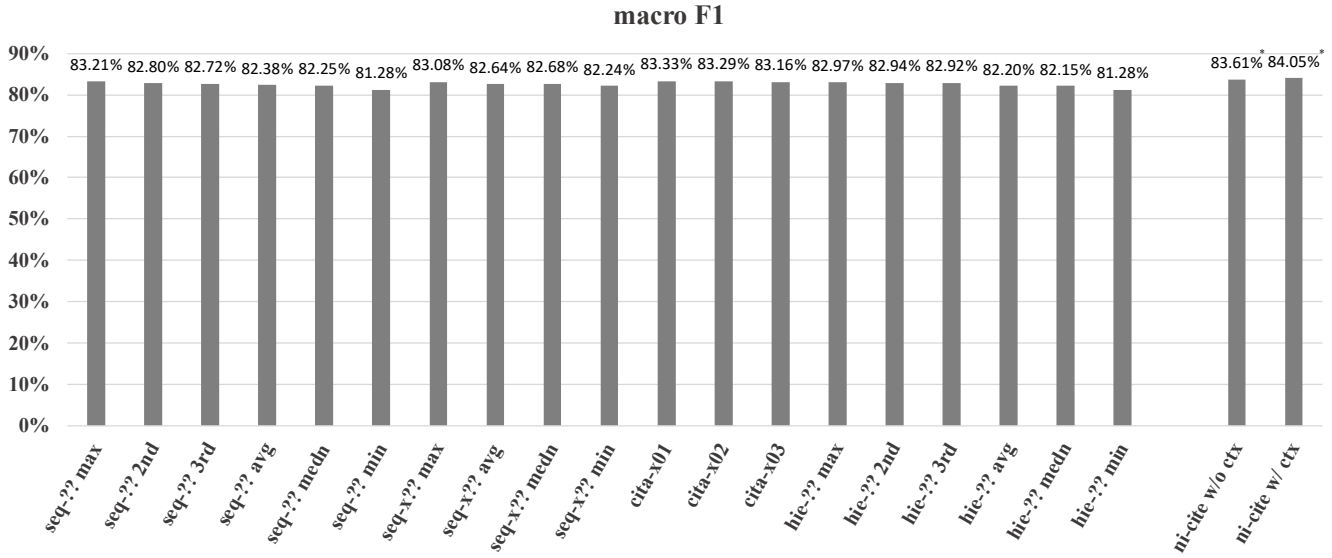
5.7. Additional Experiments on NI-Cite

To further demonstrate the necessity of modelling citations in their contexts, this section presents our additional experimental results on the `NI-Cite` dataset (Figure 3). We reported the top-3 best performances (with suffices "max", "2nd" and "3rd"), the mean (with suffix "avg") and median (with suffix "medn") of all CITSEG-aware models with a sequential context (with prefix "seq-??"), and the mean and median of all CITSEG-aware models with a hierarchical context (with prefix "hie-??"). We also reported the max, mean and median performances of all five CITSEG-agnostic models with a sequential context (with prefix "seq-x??") and the performances of the three CITSEG-agnostic citance-level models (cita-01/02/03). For comparison purpose, the rightmost two columns reported the performances of the two baseline models of Zhang et al. (2022) (with prefix "ni-cite"): One ignores citation context (with suffix "w/o ctx"), and the other encodes context sentences (with suffix "w/ ctx").

The best F1 score we achieved was 83.21%. It was almost on par with the best performances reported by running the method of Zhang et al. (2022) using their best seed, which was 83.61% without considering citation context (the "ni-cite w/o context" column in Figure 3). We also see that the three citance-level models, cita-x01 to cita-x03, were among the top models, surpassing their contextualised counterparts, i.e., seq-x07 to seq-x11. These results implies that the `NI-Cite` dataset (Zhang et al., 2022), similarly the `scicite` dataset (Cohan et al., 2019) it extends on, is "problematic" in the sense that the citations could be recognised using the citance alone. Indeed, the `scicite` dataset, as we discussed in Sect. 3, does not contain citation context information. We guess that the dataset might be annotated using citance alone. Another nuance is that `scicite` and its extended version `NI-Cite` both assign one label to each citance. Therefore, it did not help much by encoding citation context, although the best performance of Zhang et al. (2022) was improved a bit to 84.06% by searching the best learning rate. This may also explain why our citance-level models slightly outperformed citation-level models on `NI-Cite`, although the latter proved to be much stronger on `Jiang2021` (described in the next paragraph). Note that, our contextualised CFC models

were not hyperparameter-tuned; due to high computational overloads, we used the same learning parameters that were used in the experiments on `Jiang2021`, i.e., lr_pret = 5e-5 and lr_cust = 5e-4. Excluding the impact of random seed, we also conjecture that the slight performance disadvantage might also be caused by the fact that, for the `NI-Cite` dataset with the easiest 3-class annotation scheme, the 3-layer MLP used in our models might be harder to learn than the linear classifier used by Zhang et al. Anyway, our contextualised CFC models were competitive.

On the contrary, the picture on `Jiang2021` was totally different (see the last two rows in Table 5). We found that acc_step = 1 always produced the best performances on `Jiang2021`. We can see that Zhang et al.'s methods reported extremely poor performances, even after extensive hyperparameter tuning. This not only demonstrated that citation context provides indispensable information for CFC, but also proved that an appropriate citation context for CFC can often be quite large. Again, we can claim the **importance that citations should be encoded in context** and **that citation function classification should be done at citation level rather than per citance** (answers for Q1). In addition, citation function annotation should also be done at context level rather than relying on citance alone. Recall that Zhang et al.'s method dealt with only one context sentence at each side of a citance, while our methods dealt with the 2 left context sentences and 3 right context sentences. Indeed, it is a difficult problem to decide the proper context size. The ideal is to determine a "dynamic" context, i.e., the minimal context around the citance which provides enough information for determining the citation function (Abu-Jbara et al., 2012; Aggarwal et al., 2016). We leave this line of research to future work.



* Note: The last two columns "ni-cite w/o ctx" and "ni-cite w/ ctx" are the baseline method by Zhang et al. (2022) which do not consider citation context and encode context sentences respectively. There were finetuned by different learning parameters and reported using the best parameter lr = 5e-6 for both baselines.

Figure 3. CFC Performances on the `NI-Cite` Dataset.

6. Analysis

In this section, we take a deeper dive into the performances of the CFC models to see what implications we can derive for different scientific analysis tasks and the potential directions of improving CFC to suit the needs of these tasks.

6.1. Per-Class Performance Analysis

Table 6-9 present the per-class performances of a few selected models that performed well with at least one annotation scheme. Note that the first three models in each table are the top-3 models on the corresponding annotation scheme. Citation functions that are large or cognitively less complex were easier to recognise, such as neutral citations ("Neutral"/"Background") and usage citations ("Usage"/"Uses"). Teufel et al. (2006b) and Cohan et al. (2019) made similar observations. For example, the best models for neutral citations achieved the highest per-class performance 78.85% F1 with the 11-class scheme (model hie-08 in Table 6), 79.66% F1 with the 9-class scheme (model hie-14 in Table 7), 82.64% F1 with the 7-class scheme (model seq-06 in Table 8), which was close to the highest per-class performance 83.45%, and 79.88% F1 with the 6-class scheme. "Future" was a small and easy class. Indeed, we found the linguistic features were more obvious than most other categories for re-annotation. The highest F1 could reach 90.91% on the 11-class scheme (model seq-06) and the 9-class scheme (model hie-18), and 90.32% with 100% precision on the 7-class scheme (model seq-06, seed = 25603). This allows accurate bibliometric analysis of the **impact and role of future work** in scientific development (Teufel, 2017; Hao et al., 2020).

"Basis" was a difficult citation function. Language patterns like "following <cited>" and "based on <cited>" might cause confusion between "Basis" and "Usage". Human annotators often disagreed on this class too. Indeed, Teufel (2010) reported a low inter-annotator agreement on "PBas" ($\alpha = 0.41$) and "PModi" ($\alpha = 0.55$)[18]. This low inter-annotator agreement was partially why we decided to also check all `Teufel2010` instances that were merged into our new dataset. Concerning algorithmic classification, good models could achieve 66.67% F1 on the 11-class scheme (model seq-08 in Table 6). Note that the best performance 69.70% F1 was obtained by a citation-agnostic model seq-x10. This was probably because usually one approach, i.e., one CITSEG, in a citation sentence was "based on" the citing paper. On the 7-class scheme, we were excited to see an obvious performance improvement to 70% F1 (model seq-01 in Table 8, "Extends" = "Basis"). For "Motivation" we could get higher than 71.70% F1. This is promising. The best performance obtained for the motivation class was 73.21% F1 on the 6-class scheme (model hie-10, the "best of all models" column in Table 9). The overall best performing models, in term of macro F1, could not produce the best performances for "Motivation". In summary, these results are promising because they make it possible to screen out perfunctory citations using a good "Neutral"/"Background" model or to keep organic citations (Jochim & Schütze, 2012) using good "Usage"/"Uses" and/or "Basis"/"Extends" models as well as "Motivation" models. This is an important first step for **semantic analysis of scientific knowledge flows** (Ghosal et al., 2022; Jiang et al., 2022).

Comparison or contrast functions often recorded good performances. With the 11-class scheme (Table 6), we got high F1 scores for "CoCoRes" (78.12%, by model seq-12) and "CoCoGM" (71.83%, by model seq-08). With the 9-class and 7-class schemes, the F1 scores were able to reach 77.23% (model seq-12 in Table 7) and 77.84% (model hie-21, the "best of all models" column in Table 8) respectively. We see that the "Background" performance on the 7-class scheme was greatly improved to 83.45%. This seemed to empirically support the reason for merging "Weakness" into "Background" that we stated in Sect. 3, where we also explained the cognitive plausibility of doing so. On the contrary, the 6-class scheme merged both "Weakness" and "Similar" (including "Support") into the comparison classes (Jurgens et al., 2018). The performance on "Background" was about 3.5% worse compared to the 7-class scheme. We believe that this again gave partial empirical support to the decision of absorbing "Weakness" into "Background". However, it did not worsen the performance on "ComOrCon" by merging both "Weakness" and "Similarity" into it. This result is promising to applications that do not distinguish similarity from contrast

[18] Krippendorff's alpha (α): https://en.wikipedia.org/wiki/Krippendorff%27s_alpha.

and difference between two studies. Cognitively, both similarity and comparison functions (also including the supporting class "Support") imply high topical or technical relatedness between studies, so these results are very useful for building **academic recommendation systems** to identify related studies for many down-stream applications, such as assisting peer review and performing systematic review.

A related class was "Support" (equiv. Teufel et al.'s "PSup" class), which was also the most difficult class. The best F1 values on "Support" were only 50% and 51.16% on the 11-class and 9-class schemes respectively. Even such low performances were rarely seen. Similarly, Teufel (2010) reported a 0.47 F1 by her machine learning algorithm on "PSup" and the lowest inter-annotator agreement 0.27 among all her classes. Since "Support" caused an extremely low recognition rate, it was acceptable to merge it into other classes, but Table 8 shows that simply absorbing "Support" into "Similar" made it even more difficult to correctly recognise the similarity class. According to Teufel's annotation guidelines, there are two distinct meanings of "PSup"/"Support", i.e., compatibility between scientific knowledge claims or computational plug-in-ability between approaches (viewed as technical compatibility), which have quite different language uses. Therefore, we tried to re-annotate all "Support" instances into other categories and re-ran all experiments. Supplementary Sect. C presents these additional results.

"CoCoXY" was another confusing category. One possible reason is that, for certain cases, we need a *meta-statement* about comparison in a long context, which however often falls out of our context window. Ex. 3 in Table 1 illustrates this case, where the first sentence is the meta-statement. Without seeing such a meta-statement, the "CoCoXY" instance could be mis-recognised as either "CoCoGM" or "Neutral". This class often confuses with Teufel's Rule 40 about "List of Approaches", which says "Neut(ral)" should be applied if no meta-statement of comparison exists. Although in certain cases comparison can be inferred from juxtaposed citations, Teufel's annotation guideline says that only explicit comparison expressions qualify "CoCoXY". Ex. 5 in Table 1 illustrates the latter case. We followed Teufel's principle in our own annotation guidelines.

6.2. No Single Model Fits All

An important observation is that no single model variant and no single trained model (with a specific seed) could beat others on all citation functions or on all annotation schemes. This phenomenon is especially obvious on the 6-class scheme (Table 9), where the model with the best overall performance (74.03% F1) was not the best model in term of per-class performance for any citation function. On the contrary, the second best model (73.56% F1) won on the "Future" class, while the third best model (73.22% F1) won on "Background", "ComOrCon" and "Uses" classes by large margins. By now, we can conclude that **no single CFC model is robust enough for performing all types of scientific analysis tasks based on citation context analysis**. Our opinion is that, for each different task, it is better to choose or develop a bespoke CFC model tailored to that task.

For example, the best CFC model for "Future" should be chosen to analyse the **scientometric value of the future work sections** of a paper (Teufel, 2017; Hao et al., 2020). To bibliometrically analyse **the usage of scientific entities**, such as algorithm usage (Wang & Zhang, 2020), method usage (Wang et al., 2022), software usage (Li et al., 2019), or dataset usage (Fan et al., 2022), we will need to turn to the best "Usage" model. However, the annotation schemes adopted in this study are rooted in Teufel et al.'s 12-class scheme, and does not annotate the type of cited entity, such as algorithm, method, software, and dataset etc. To facilitate fine-grained scientometric analysis, it would be better to employ a two-level annotation scheme (Lu et al., 2014; Zhang et al., 2021) or multi-level annotation scheme (Budi & Yaniasih, 2022), which consider not only why something is cited but also what specific scientific entity is cited. To analyse **scientific research lineage** (Ghosal et al., 2022) or extract **technology dependency roadmap** (Zha et al., 2019; Yin et al., 2021) we will need a strong CFC model for "Basis"/"Extends" citations. The best model across all four annotation schemes reported a 70.00% F1. While the overall performance can be said good, there is still a problem of trade-off between precision and recall. As this is the most important class (Lu et al., 2014; Valenzuella et al., 2015), there is large room of improvement and demand of further research. In Sect. 7.1, we will see how a simple ensemble method could improve the performance to around 75% F1.

For the purpose of scientific ranking, we may wish to either suppress incidental citations (Valenzuella et al., 2015) or remove such perfunctory citations (Jochims & Schütze, 2012). Luckily, "Future" and "Neutral"/"Background" both reported good performances. Recall that the best "Background" model absorbed "Weakness" and reported an 83.45% F1. We believe that **it is valid to rely on citation function classification to screen out incidental/perfunctory citations, or weight citations based on citation function for scientific ranking**. In addition, it would be interesting to do **main path network analysis** (Jiang et al., 2020) in a citation semantics-aware way. To do this, we can choose to keep only *organic* citations by use of strong "Usage" and "Basis"/"Extends" modes. Optionally, we can also only rely on good "Basis"/"Extends" models if we emphasise on the "*evolutionary* v.s. *juxtapositional*" aspect of citations (Jochims & Schütze, 2012), which characterizes whether a citing study "builds on the cited work" or "presents an alternative to the cited. Jiang and Liu (2022) presented an initial attempt of this idea.

Table 6. Per-Class Performances of Selected Models with the 11-Class Scheme

ID (seed)	seq-08 (5171)			seq-06 (47353)			hie-18 (13249)			hie-08 (32491)			seq-12 (5171)			seq-11 (47353)			seq-01(13249)			best of all models		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Macro Avg	68.50	65.17	66.16	67.74	64.05	65.12	67.10	64.17	64.95	65.18	65.31	65.02	65.59	65.11	64.93	65.28	64.00	64.42	65.91	63.28	63.93	--	--	--
Future	88.24	88.24	88.24	93.75	88.24	**90.91**	93.33	82.35	87.50	66.67	82.35	73.68	70.00	82.35	75.68	87.50	82.35	84.85	81.25	76.47	78.79	93.75	88.24	**90.91**
Neutral	72.24	78.16	75.08	74.76	79.86	77.23	75.60	75.09	75.34	78.19	79.52	**78.85**	77.48	79.86	78.66	75.33	78.16	76.72	77.70	78.50	78.10	78.19	79.52	**78.85**
Weakness	65.52	59.38	62.30	70.37	59.38	**64.41**	73.91	53.12	61.82	65.52	59.38	62.30	71.43	46.88	56.60	66.67	56.25	61.02	75.00	46.88	57.69	76.92	62.50	**68.97**
CoCoXY	69.23	58.06	63.16	60.00	48.39	53.57	59.46	70.97	**64.71**	57.58	61.29	59.38	69.23	58.06	63.16	51.61	516.1	51.61	58.82	64.52	61.54	78.57	70.97	**74.58***
CoCoGM	67.11	77.27	**71.83**	70.97	66.67	68.75	72.41	63.64	67.74	63.38	68.18	65.69	62.16	69.70	65.71	67.65	69.70	68.66	52.53	78.79	63.03	67.11	77.27	**71.83**
CoCoRes	65.62	67.74	**66.67**	62.50	80.65	70.42	81.82	58.06	67.92	63.64	67.74	65.52	75.76	80.65	**78.12**	62.50	80.65	70.42	68.75	70.97	69.84	80.00	77.42	**78.69**
Similar	67.74	50.00	57.53	60.87	66.67	63.64	60.00	57.14	58.54	68.42	61.90	**65.00**	59.18	69.05	63.74	67.57	59.52	63.29	60.98	59.52	60.24	71.05	64.29	**67.50**
Support	46.15	30.00	36.36	46.15	30.00	36.36	34.62	45.00	39.13	36.84	35.00	35.90	38.10	40.00	39.02	42.11	40.00	**41.03**	29.41	25.00	27.03	47.53	55.00	**51.16**
Motivation	65.00	67.24	66.10	59.42	70.69	64.57	51.85	72.41	60.43	65.08	70.69	67.77	62.90	67.24	65.00	61.29	65.52	63.33	74.07	68.97	**71.43**	65.71	79.31	**71.88**
Usage	77.94	70.20	73.87	76.39	72.85	74.58	71.71	72.19	71.95	77.62	73.51	**75.51**	77.21	69.54	73.17	75.54	69.54	72.41	79.86	73.51	76.55	79.17	75.50	**77.29**
Basis	63.16	70.59	**66.67**	70.00	41.18	51.85	63.33	55.88	59.38	74.07	58.82	65.57	58.06	52.94	55.38	56.25	52.94	54.55	66.67	52.94	59.02	71.88	67.65	**69.70***

* This result comes from model hie-13.
** This result comes from model seq-x10.

Table 7. Per-Class Performances of Selected Models with the 9-Class Scheme

ID (seed)	seq-12 (47353)			hie-08 (47353)			seq-11 (47353)			hie-18 (13491)			seq-08 (32491)			hie-14 (5171)			hie-14 (25603)			best of all models		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Macro Avg	69.25	67.13	67.78	67.51	67.89	67.49	68.24	66.31	66.92	72.21	62.51	66.07	70.71	63.02	66.03	67.90	64.62	65.69	69.37	63.58	65.67	--	--	--
Future	93.33	82.35	87.50	86.67	76.47	81.25	86.67	76.47	81.25	93.75	88.24	**90.91**	92.31	70.59	80.00	83.33	88.24	85.71	76.19	94.12	84.21	93.75	88.24	**90.91**
Neutral	77.31	79.94	78.60	76.53	73.46	74.96	46.88	75.93	76.40	73.94	85.80	79.43	74.86	83.64	79.01	76.88	75.93	76.40	73.88	86.42	**79.66**	73.88	86.42	**79.66**
Weakness	65.38	53.12	58.62	60.61	62.50	61.54	66.67	56.25	61.02	94.44	53.12	**68.00**	70.00	43.75	53.85	78.26	56.25	65.45	62.50	46.88	53.57	80.00	62.50	**70.18**
Comparison	68.81	80.41	**77.23**	63.25	76.29	69.16	64.55	73.20	68.60	60.00	71.13	65.09	67.31	72.16	69.65	67.77	84.54	75.23	66.98	73.20	69.95	68.81	80.00	**77.23**
Similar	62.79	64.29	63.53	61.36	64.29	62.79	60.87	66.67	**63.64**	64.86	57.14	60.76	58.54	57.14	57.83	57.78	61.90	59.77	68.57	57.14	62.34	76.47	61.90	**68.42**
Support	45.45	50.00	47.62	50.00	55.00	52.38	52.94	45.00	**48.65**	30.00	30.00	30.00	61.54	40.00	48.48	46.67	35.00	40.00	50.00	35.00	41.18	66.67	40.00	**50.00**
Motivation	59.46	75.86	66.67	62.69	72.41	67.20	59.46	75.86	66.67	79.17	65.52	**71.70**	68.42	67.24	67.83	69.23	62.07	65.45	79.55	60.34	68.63	79.17	65.52	**71.70**
Usage	79.84	68.21	73.57	76.47	68.87	72.47	78.26	71.52	74.74	82.26	67.55	74.18	81.68	70.86	**75.89**	71.15	73.51	72.31	80.00	66.23	72.46	76.13	78.05	**77.12**
Basis	65.38	50.00	56.67	70.00	61.76	**65.62**	67.86	55.88	61.29	71.43	44.12	54.55	61.76	61.76	61.76	60.00	44.12	50.85	66.67	52.94	59.02	82.61	55.88	**66.77**

Table 8. Per-Class Performances of Selected Models with the 7-Class Scheme

ID (seed)	hie-14 (13249)			hie-19 (13249)			seq-02 (32491)			seq-01 (47353)			seq-12 (32491)			seq-06 (25603)			seq-06 (13249)			best of all models		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Macro Avg	73.04	73.04	72.81	73.72	71.26	72.39	71.18	71.80	71.39	76.77	66.83	70.77	70.00	71.57	70.65	74.60	68.23	70.63	70.08	70.66	70.35	--	--	--
Future	82.35	82.35	82.35	86.67	76.47	81.25	82.35	82.35	82.35	100.0	58.82	74.07	72.22	76.47	74.29	100.00	82.35	**90.32**	76.47	76.47	76.47	93.33	82.35	**87.50**
Background	80.28	81.18	80.73	80.91	83.99	82.37	81.34	82.02	81.68	77.25	86.80	81.75	83.67	80.62	82.12	81.08	84.27	**82.64**	82.15	81.46	81.81	80.10	87.08	**83.45**
ComOrCon	83.12	65.98	**73.56**	73.96	73.20	**73.58**	70.71	72.16	71.43	71.29	74.23	72.73	76.40	70.10	73.12	59.35	75.26	66.36	73.12	70.10	71.58	81.82	74.23	**77.84**
Similar	63.08	66.13	**64.57**	64.41	61.29	62.81	63.16	58.06	60.50	58.62	54.84	56.67	55.56	64.52	59.70	58.93	53.23	55.93	54.10	53.23	53.66	63.08	66.13	**64.57**
Motivation	61.43	74.14	67.19	62.90	67.24	65.00	66.67	65.52	66.09	68.52	63.79	66.07	61.19	70.69	65.60	71.70	65.52	68.47	68.25	74.14	**71.07**	80.85	65.52	**72.38**
Uses	76.32	76.32	76.32	79.58	74.83	**77.13**	77.93	74.83	76.35	80.95	67.55	73.65	75.32	76.82	76.07	79.71	72.85	76.12	76.47	77.48	76.97	83.85	72.19	**77.58**
Extends	64.71	64.71	64.71	67.14	61.76	64.62	56.10	67.65	61.33	80.77	61.76	**70.00**	65.62	61.76	63.64	71.43	44.12	54.55	60.00	61.76	60.87	80.77	61.76	**70.00**

Table 9. Per-Class Performances of Selected Models with the 6-Class (`Jurgens2018`) Scheme

ID (seed)	seq-01 (47353)			seq-12 (5171)			hie-15 (13249)			hie-15 (5171)			seq-06 (5171)			hie-18 (47353)			hie-19 (25603))			best of all models		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Macro Avg	77.27	71.53	74.03	75.86	71.82	73.56	72.80	73.81	73.24	72.80	72.17	72.17	75.27	69.69	72.19	75.91	70.06	72.09	72.34	71.55	71.89	--	--	--
Future	92.86	76.47	83.87	93.33	82.35	**87.50**	83.33	88.24	85.71	81.25	76.47	78.79	8667	76.47	81.25	82.35	82.35	82.35	76.47	76.47	76.47	88.24	88.24	**88.24**
Background	75.00	78.70	76.81	75.29	79.01	77.11	78.92	80.86	**79.88**	75.83	77.47	76.64	74.71	80.25	77.38	76.21	73.15	74.65	77.13	78.09	77.61	78.92	80.86	**79.88**
ComOrCon	73.37	76.44	74.87	73.80	72.25	73.02	77.22	72.77	**74.93**	72.19	70.68	71.43	71.07	73.30	72.16	62.55	82.20	71.04	73.23	75.92	74.55	76.68	77.49	**77.08**
Motivation	67.24	67.24	67.24	61.90	67.24	64.46	61.54	68.97	65.04	55.84	74.14	63.70	72.00	62.07	66.67	72.55	63.79	**67.89**	62.30	65.52	63.87	75.93	70.69	**73.21**
Uses	78.26	71.52	74.74	77.78	74.17	75.93	78.23	76.16	**77.18**	78.36	69.54	73.68	76.81	70.20	73.36	84.55	68.87	75.91	77.14	71.52	74.23	76.97	77.48	**77.23**
Extends	76.92	58.82	66.67	73.08	55.88	63.33	57.58	55.88	56.72	73.33	64.71	**68.75**	70.37	55.88	62.30	77.27	50.00	60.71	67.74	61.76	64.62	77.78	61.76	**68.85**

7. Ensembling

In Sect. 6.2, we discussed that there was no single best model that worked the best on all citation function annotation schemes for all citation functions. We saw a seesaw phenomenon that, typically, when a model worked well on some functions it became less effective on the remaining. From Sect. 6.1 sometimes the best model in term of overall performance could not produce the best performance for any citation function. The best performances for different functions could only be obtained by different models. In addition, we also saw drastic differences in the behaviours of different models, i.e., the prediction results of different models bore high degree of diversity. These observations are all the basis of utilising multiple trained models to build an ensemble classifier to achieve better CFC performance. This section presents our preliminary results in this direction. Figure 4 illustrates the idea of the *naïve ensemble* classifier. Refer to Zhou (2014) for more details of ensemble learning.

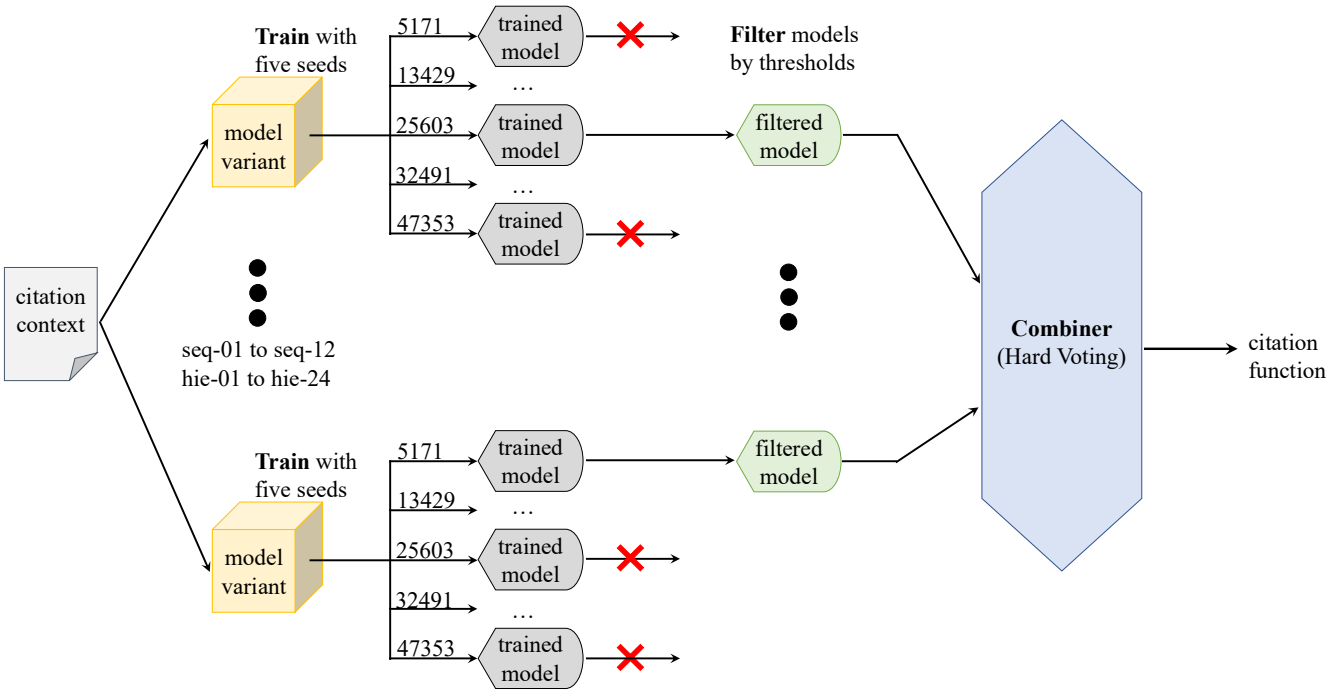


Figure 4. Naïve Ensemble Model for Citation Function Classification

7.1. Building Ensembles

The first step was base classifier selection. Recall that, we proposed in total 36 model variants, i.e., seq-01 to seq-12, and hie-01 to hie-24. For each model variant, we trained five models with five seeds, each trained with 20 epochs. We obtained one best trained model for each seed and each model variant according to validation performance. There were in total $5 \times 36 = 180$ trained models as base classifiers. To build the ensemble classifier, we first filtered the base classifiers according to (either overall or per-class) test F1 score by adjusting performance threshold such that no less than 20 models were kept as candidate base classifiers. Then we sorted the base classifiers in descending order of their performances (i.e., test F1 score) and chose the top T models as the base classifiers. Note that, this is why we call our approach *naïve ensemble* because typically ensembling choices need be made based on analysing classifier diversity. However, we skipped this step but simply chose the top T as base

classifiers. This simplification, or *naïve* treatment, might cause some problems when base classifiers have similar behaviours. However, we will see that this naïve ensemble method worked pretty well most of the time.

The second part of the ensemble approach was the combiner. We left more in-depth study of ensembling to future work but focused on the simplest approach, *hard majority voting*. This is the second reason we call our approach *naïve ensemble*. Please refer to Zhou (2014; Ch. 4) for details about different combination methods. Here, three types of information should be considered to derive ensemble decisions: (1) *predictions* of each base classifier, i.e., the predicted class labels, (2) base classifier's *confidence*, i.e., the posterior probabilities of each base classifier for each instance, and (3) base classifier's *reliability*, i.e., the overall performances of each base classifier. The last two types of inputs were used to break ties when two or more classes got the same number of votes. In case of ties, the accumulated confidence for each base classifier was used first, and if in rare cases ties still happened, then base classifier reliability was used to break ties. From the top-T base classifiers, we selected K (K = 2, 3, …, T) from them in descending order of their performances to build ensembles and found the best K which produced the best ensemble performance. We were aware that we did by no means exhaust all possibilities of choosing K base classifiers for the purpose of building ensemble; there should be C_T^K possible ways to combine K base classifiers out of a pool of T models. Note that, exhaustive searching is usually avoided by anlaysing *classifier diversity*. Please refer to Zhou (2014; Ch. 5) for details about classifier diversity. An in-depth analysis of the various available ensembling choices deserves a separate paper, so we leave this line of research to future work.

7.2. Ensemble Results

Experiments were done on each annotation scheme targeting at improving either the overall CFC performance or the CFC performance of difficult functions, e.g., "Basis", "Motivation", "Similar", "Support", "Weakness". In addition, we also tested "Future" as this class also has a special bibliometric analysis purpose. To improve the overall performance, base classifiers were selected, filtered, and sorted according to macro F1. When targeting at improving the recognition rate of a specific function, base classifiers were selected, filtered, and sorted according to per-class F1 of the function. T was set to 20 in all experiments. Table 10 summarises the results of each ensemble classifier together with the corresponding best K. We see that our naïve ensemble method brought in non-trivial performance boosts to almost all cases, except on the "Future" class with the 9-class and 8-class scheme. We reported a small performance drop in the former case and recorded an ensemble performance on par with the best base classifier in the latter case. We conjecture that this might be caused by not performing classifier diversity analysis. "Future" was the class gaining the highest recognition rate. High recognition rate means relatively low classifier diversity, which in turn may bring adverse impact rather than positive impact on ensemble performance (Sesmero et al., 2021).

Huge improvement happened to the difficult classes, e.g., raising the performance of "Weakness" to 75.88%, "Similar" to 72.94%, and the important "Motivation" class and "Basis" class to 77.31% and 75.41% respectively. On all these four classes, the performance improvements were very significant. Although the biggest improvement happened to "Support", which recorded a 9.3% absolute improvement (a 18.15% relative improvement) to 59.46% F1, the performance was still too low, which re-iterates the importance of treating the annotation and recognition of "Support" (in the sense of "mutual compatibility") relationships as a specific machine learning task, as in the recent work by Nicholson et al. (2021). Obviously, we could further merge the models trained on various annotation schemes, if the annotation schemes share the same class. We shall leave further analysis to future work. Note that, all our base classifiers were multi-class CFC models, which were trained in a multi-class way but were used as binary classifiers. If binary CFC models were developed, we should be able to anticipate even better performances. In addition, the multi-class models developed in the current study should be good starting points for training the binary CFC models bespoke to specific citation functions. Overall, these results are very promising as they prove that decent recognition performances are achievable by contextualised citation modelling based on cutting edge deep learning and machine learning techniques. The ensemble models with decent performances for "Basis", "Usage", "Motivation", and "Similar" classes

allow us to perform various types of scientific analysis tasks that were discussed in Sect. 6.2. This would be one very important and interesting future direction.

Table 10. Performances of Naïve Ensembles of Citation Function Classification Models

	11-class							**9-class**							**7-class**							**6-class**						
	best model			**best ensemble**				**best model**			**best ensemble**				**best model**			**best ensemble**				**best model**			**best ensemble**			
	P	R	F1	P	R	F1	K	P	R	F1	P	R	F1	*K*	P	R	F1	P	R	F1	*K*	P	R	F1	P	R	F1	*K*
Macro Avg	68.50	65.17	66.16	72.76	68.30	**69.98**	19	69.25	67.13	67.78	74.17	67.77	**70.40**	5	73.04	73.04	72.81	76.71	74.80	**75.66**	5	77.27	71.53	74.03	78.63	74.61	**76.47**	6
Future	93.75	88.24	90.91	100.0	88.24	**93.75**	9	93.75	88.24	90.91	100.0	82.35	90.32	5[*]	93.33	82.35	87.50	93.33	82.35	87.50	4	88.24	88.24	88.24	100.0	88.24	**93.75**	15
Neutral/Background	78.19	79.52	78.85	78.06	84.98	**81.37**	16	73.88	86.42	79.66	75.80	87.96	**81.43**	19	80.10	87.08	83.45	81.94	87.92	**84.82**	7	78.92	80.86	79.88	78.20	83.02	**80.54**	3
Weakness	76.92	62.50	68.97	78.57	68.75	**73.33**	3	80.00	62.50	70.18	84.62	68.75	**75.86**	9	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Similar	71.05	64.29	67.50	72.09	73.81	**72.94**	3[**]	76.47	61.90	68.42	72.50	69.05	**70.73**	7	63.08	66.13	64.57	73.68	67.74	**70.59**	2[***]	--	--	--	--	--	--	--
Support	47.53	55.00	51.16	64.71	55.00	**59.46**	4	66.67	40.00	50.00	64.71	55.00	**59.46**	14	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Motivation	65.71	79.31	71.88	75.41	79.31	**77.31**	9	79.17	65.52	71.70	75.86	75.86	**75.86**	5	80.85	65.52	72.38	75.86	75.86	**75.86**	10	75.93	70.69	73.21	78.18	74.14	**76.11**	7
Usage/Uses	79.17	75.50	77.29	82.52	78.15	**80.27**	12	76.13	78.05	77.12	84.29	78.15	**81.10**	11	83.85	72.19	77.58	82.64	78.81	**80.68**	15	76.97	77.48	77.23	82.61	75.50	**78.89**	17
Basis/Extends	71.88	67.65	69.70	80.00	70.59	**75.00**	5	82.61	55.88	66.77	81.48	64.71	**72.13**	5	80.77	61.76	70.00	88.00	64.71	**74.58**	10	77.78	61.76	68.85	85.19	67.65	**75.41**	5

[*] Here we reported the only performance drop from the experiments on "Future" on the 9-class scheme, and no performance improvement on the 7-class scheme.
[**] This result was reported after removing a duplicate model (hie-18), i.e., a model which makes 100% the same predictions as another model. Without removing it, the performance degraded.
[***] Hard voting worked with even two base classifiers because we also considered base classifiers' confidence and reliability to break ties.

8. Concluding Remarks

This paper studied contextualised segment-wise citation function classification and analysed the implications of the results for scientific analysis applications. Several contributions were made. The first contribution was a larger citation context dataset that was created by merging and re-annotating six datasets in the computational linguistics domain whose annotation schemes were identified partially mappable by a critical review. The new dataset contains 3356 citation contexts, 4784 in-text citations and 3854 citation segments (consecutive block of in-text citation strings). The new dataset has three-fold prominent benefits: (i) The important minority classes are at least doubled, making it better suit deep learning, (ii) the computational linguistics domain is better represented by covering a wide time range from early 1990s to late 2010s, and (iii) the availability of full citation contexts allows for the development of strong citation function classification algorithms based on better citation modelling. Secondly, we performed extensive experiments to study the effective ways of citation modelling – the main research question to answer in this study. To do this, we proposed a series of contextualised segment-wise citation function classification models based on SciBERT by exploring a comprehensive set of options for modelling and combining the feature representations of in-text citation (segment), citation sentence and citation context. New states of the art were achieved. The best model reported a 66.16% macro F1 on the 11-class scheme and improved it to 74.03% on the 6-class scheme.

Concerning the main research question, we were able to empirically conclude that citation function classification should be done at citation level, i.e., by modelling individual citation (segments), rather than per citation sentence or context. Notably, ignoring citation representation led to very poor performance most of the time, and using citation representation alone produced surprisingly competitive results in many cases. Experiments also justified the importance of contextualised citation modelling, i.e., citations should better be encoded in context. This also means that often the best performances were only achievable by appropriately modelling citation sentence and/or context into citation representation. However, it was hard to conclude what are the consistently effective ways of modelling and combining the representations of in-text citations, citation sentences and citation contexts. They had to be decided case by case for different annotation schemes. Even though, we were able to observe some minor patterns. For example, “self_attnd” was most of the time a stronger context pooler than “max_pool” to enrich citation representation in case of sequential context. On the contrary, “max_pool” was a better citance pooler than “self_attnd” when “[CLS]” was used as context pooler.

An in-depth per-class performance analysis revealed that large functions like the neutral and usage citations got the best and most stable results. The weakness class was more difficult because the weak point of something is often pointed out after a neutral description in the citation context, which re-iterates the necessity of contextualised citation function classification. Two citation functions were difficult to recognise: (i) functions whose language expressions overlap other categories like similarity v.s. usage citations, and (ii) functions whose definitions embrace two or more distinct meanings such as the citation relation about mutual support. The per-class performance analysis also pointed out an important reality of citation function classification: NO single best general-purpose citation function classifier exists for all citation functions and all scientific analysis tasks. For each task, a bespoke model tailored to a specific function is needed. Good news was that, thanks to the versatility of well-performing models, we were able to build a naïve ensemble model for citation function classification, which significantly improved not only the overall performance but also the performances of all difficult classes. We believe that there is much room to explore in this direction.

In summary, we were able to conclude that, although not perfect for all citation functions, existing citation function classification models allow for a wide range of scientific applications. For example, the best models were extremely strong in screening out incidental or perfunctory citations, such as citations about neutral description, background introduction and future work etc. This would allow us to perform scientific knowledge flow analysis and academic ranking in a semantics-rich way based on citation context analysis. Citations about comparison or contrast were able to be recognised rather correctly, so would

be promising to be applied to recommending related studies to facilitate many useful applications such as peer review and systematic review etc. Our citation function classification models also obtained decent performances on citations about technical extension or ideational basis or citations about inspiration and motivation. Especially our naïve ensemble models significantly improved performances for these two difficult classes. They greatly facilitate analysing scientific research lineage. The third, but not the last, interesting application would be analysing the pattern of scientific entity usage, including dataset, software, algorithm, method and so on. Our models were strong enough on recognising usage citations for such applications, but the citation context dataset must be extended by citation object.

9. FUNDING AND CONFLICTS OF INTERESTS

The first author Xiaorui Jiang is partially supported by National Planning Office for Philosophy and Social Sciences of China (18ZDA238). Both authors have no competing interests to declare that are relevant to the content of this article.

10. ACKNOWLEDGMENT

The authors deliver their most sincere gratitude to Prof. Simone Teufel for kindly sharing her annotation guidelines and the valuable discussions about citation function annotation.

11. REFERENCES

Abu-Jbara, A., Erza, J., & Radev, D. (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 (NAACL-HLT'13), 596–606. https://aclanthology.org/N13-1067

Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically Classifying the Role of Citations in Biomedical Articles. In *Proceedings of the 2010 Annual Symposium of the American Medical Informatics Association (AMIA'10)*, 11–15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041379

Aljohani, N.R., Fayoumi, A., & Hassan, S.-U., (2021a). A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations. *Journal of Information Science* (Mar, 2021), 1–14. https://doi.org/10.1177/0165551521991022

Aljohani, N.R., Fayoumi, A., Hassan, S.-U., (2021b). An in-text citation classification predictive model for a scholarly search system. *Scientometrics*, *126*, 5509–5529. https://doi.org/10.1007/s11192-021-03986-z

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596. https://aclanthology.org/J08-4004/

Bakhti, K., Niu, Z., Yousif, A., & Nyamawe, A.S. (2018). Citation Function Classification Based on Ontologies and Convolutional Neural Networks. In: L. Uden, D. Liberona, J. Ristvej (Eds.) *Communications in Computer and Information Science*: *Vol 870. Learning Technology for Education Challenges. LTEC 2018* (pp. 105–115). Springer, Cham. https://doi.org/10.1007/978-3-319-95522-3_10

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP'19)*, 3615–3620. https://aclanthology.org/D19-1371

Budi, I., & Yaniasih, Y. (2022). Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*. Published Online on 14 November 2022. https://doi.org/10.1007/s11192-022-04567-4

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 3856–3896. https://aclanthology.org/N19-1361

Davies,M., & Fleiss, J. L. (1982). Measuring Agreement for Multinomial Data. *Biometrics*, *38*(4), 1047–1051. https://doi.org/10.2307/2529886

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833. https://doi.org/10.1002/asi.23256

Dong, C., & Schäfer, U. (2011). Ensemble-style Self-training on Citation Classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, 623–631. https://aclanthology.org/I11-1070

Eberts M., & Adrian Ulges, A. (2020). Span-based Joint Entity and Relation Extraction with Transformer Pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI'20)*. https://ecai2020.eu/papers/1283_paper.pdf

Erikson, M. G., & Erlandson, P. (2014). A taxonomy of motives to cite. *Social Studies of Science*, *44*(4), 625–637. https://doi.org/10.1177/0306312714522871

Fan, W.-M., Jeng, W., & Tang, M.-C. (2022). Using data citation to define a knowledge domain: A case study of the Add-Health dataset. *Journal of the Association for Information Science and Technology*, Online Publishing. https://doi.org/10.1002/asi.24688

Ferrod, R., Di Caro, L., & Schifanella, C. (2021). Structured Semantic Modeling of Scientific Citation Intents. In *Proceedings of the 2021 Extended Semantic Web Conference (ESWC'21)*, 461–476. https://doi.org/10.1007/978-3-030-77385-4_27

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://psycnet.apa.org/doi/10.1037/h0031619

Ghosal, T., Tiwary, P., Patton, R., & Stahl, C. (2022). Towards establishing a research lineage via identification of significant citations. *Quantitative Science Studies* (Advance publication). https://doi.org/10.1162/qss_a_00170

Garzone, M., & Mercer, R.E. (2000). Towards an Automated Citation Classifier. In *Proceedings of the 2000 Conference of the Canadian Society for Computational Studies of Intelligence (Canadian AI'20)*, 337-346. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45486-1_28

Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying Important Citations Using Contextual Information from Full Text. In *Proceedings of the 2017 IEEE/ACM Joint Conference on Digital Libraries (JCDL'17)*, 41–48. https://doi.org/10.1109/JCDL.2017.7991558

Hao, W., Li, Z., Qian, Y., Wang, Y., & Zhang, C. (2020). The ACL FWS-RC: A Dataset for Recognition and Classification of Sentence about Future Works. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL'20)*, 261–269. https://doi.org/10.1145/3383583.3398526.

Hernández-Alvarez, M., & Gómez, J.M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, *22*(3), 327–349. https://doi.org/10.1017/S1351324915000388

Hernández-Alvarez, M., Gómez, J.M., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*, *23*(4), 561–588. https://doi.org/10.1017/S1351324916000346

Iorio, A.D., Nuzzolese, A.G., & Peroni, S. (2013). Towards the automatic identification of the nature of citations. In *Proceedings of the 3rd Workshop on Semantic Publishing (SePublica'13) at the 10th Extended Semantic Web Conference (ESWC'13)*, 63–74. http://ceur-ws.org/Vol-994/paper-06.pdf

Iqbal, S., Hassan, S.-U., Aljohani, N.R., Alelyani, S., Nawaz, R., & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics*, *126*, 6551–6599. https://doi.org/10.1007/s11192-021-04055-1

Jha, R., Abu-Jbara, A., Qazvinian, V., & Radev, D.R., (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130. https://doi.org/10.1017/S1351324915000443

Jiang, X., Zhu, X., & Chen, J. (2020). Main path analysis on cyclic citation networks. *Journal of the Association for Information Science and Technology, 71*(5), 578–595. https://doi.org/10.1002/asi.24258

Jiang, X. (2021). An Empirical Study of Span Modeling in Science NER. In *Proceedings of the 2021 International Conference on Theory and Practice of Digital Libraries (TPDL'21)*, 41–48. https://doi.org/10.1007/978-3-030-86324-1_4

Jiang, X., & Liu, J. (2022). Extracting the Evolutionary Backbone of Scientific Domains: The Semantic Main Path Network Approach Based on Citation Context Analysis. Preprint. https://pureportal.coventry.ac.uk/en/publications/extracting-the-evolutionary-backbone-of-scientific-domains-the-se

Jochim, C., & Schütze, H. (2012). Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, 1343–1358. https://aclanthology.org/C12-1082

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistic*, *6*, 391–406. https://doi.org/10.1162/tacl_a_00028

Kilicoglu, H., Peng, Z., Tafreshi, S., Tran, T., Rosemblat, G., & Schneider, J. (2019). Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics*, *91*, 103123. https://doi.org/10.1016/j.jbi.2019.103123

Kunnath, S.N., Pride, D., Gyawali, B., & Knoth, P. (2020). Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications (WOSP'2020)*, 75–83. https://aclanthology.org/2020.wosp-1.12

Kunnath, S.N., Herrmannova, D., Pride, D., & Knoth, P. (2021). A meta-analysis of semantic classification of citations. *Quantitative Science Studies* (Advance publication). https://doi.org/10.1162/qss_a_00159

Lauscher, A., Glavaš, G., Ponzetto, S.P., & Eckert, K. (2017). Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. *Proceedings of the 6th International Workshop on Mining Scientific Publications (WOSP'17)*, 24–28. https://doi.org/10.1145/3127526.3127531

Lauscher, A., Brandon, K., Kuehl, B., Johnson, S., Jurgens, D., Cohan, A., & Lo, K. (2021). MULTICITE: Modelling realistic citations requires moving beyond the single-sentence single-label setting. *Preprint.* https://arxiv.org/abs/2107.00414

Li, X., He, Y., Meyers, A., & Grishman, R. (2013). Towards Fine-grained Citation Function Classification. In *Proceedings of the 2013 Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'13)*, 402–407. https://aclanthology.org/R13-1052

Li, K., Chen, P.-Y., & Yan, E. (2019). Challenges of measuring software impact through citations: An examination of the lme4 R package. *Journal of Informetrics*, *13*(1), 449–461. https://doi.org/10.1016/j.joi.2019.02.007

Lu, W., Meng, R., & Liu, X. (2014). A Deep Scientific Literature Mining-Oriented Framework for Citation Content Annotation. *Journal of Library Science in China*, *40*(214), 93–104. (in Chinese) https://doi.org/10.13530/j.cnki.jlis.140029

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: a meta-synthesis. *Scientometrics*, *126*, 3243–3264. https://doi.org/10.1007/s11192-021-03908-z

Maheshwari, H., Singh, B., & Varma, V. (2021). SciBERT Sentence Representation for Citation Context Classification. In *Proceedings of the Second Workshop on Scholarly Document Processing (SDP'21)*, 130–133. https://aclanthology.org/2021.sdp-1.17

Meng, R., Lu, W., Chi, Y., & Han, S. (2017). Automatic Classification of Citation Function by New Linguistic Features. In *Proceedings of iConference 2017*, 826–830. https://doi.org/10.9776/17349

Meyers, A. 2013. Contrasting and corroborating citations in journal articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'13)*, 460–466. https://aclanthology.org/R13-1060

Moravcsik, M. J. and Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*, 86–92. https://doi.org/10.1177/030631277500500106

Munkhdalai, T., Lalor, J., & Yu, H. (2016). Citation Analysis with Neural Attention Models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI'16)*, 69–77. https://aclanthology.org/W16-6109

Nanba, H., Kando, N., & Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proceedings of the 11th ASIS SIG/CR Classification Research Workshop*, 117-134. http://dx.doi.org/10.7152/acro.v11i1.12774

Nicholson, J.M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N.P., Grabitz, P., & Rife, S.C. (2021). scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitiative Science Studies*, *2*(3), 882–898.

Pride, D., & Knoth, P. (2017). Incidental or influential? - challenges in automatically detecting citation importance using publication full texts. In: J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.) *Lecture Notes in Computer Science: Vol 10450. Research and Advanced Technology for Digital Libraries. TPDL 2017* (pp. 572–578). https://doi.org/10.1007/978-3-319-67008-9_48

Sesmero, M.P., Iglesias, J.A., Magán, E., Ledezma, A., & Sanchis, A. (2021). Impact of the learners diversity and combination method on thegeneration of heterogeneous classifier ensembles. *Applied Soft Computing*, *111*, page 1076689. https://doi.org/10.1016/j.asoc.2021.107689

Su, X., Prasad, A., Kan, M.-Y., & Sugiyama, K. (2019). Neural Multi-task Learning for Citation Function and Provenance. In *Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL'19)*, 394–395. https://doi.org/10.1109/JCDL.2019.00122

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*, 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, 103–110. https://aclanthology.org/W06-1613

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (SIGdial'06)*, 80–87. https://aclanthology.org/W06-1312

Teufel, S. (2010). The Structure of Scientific Articles: Applications to Citation Indexing and Summarization. Centre for the Study of Language & Information.

Teufel, S. (2017). Do "Future Work" sections have a purpose? Citation links and entailment for global scientometric questions. In *Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) co-located with the 40th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. http://ceur-ws.org/Vol-1888/paper1.pdf

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying Meaningful Citations. In Proceedings of the Workshops of Scholarly Big Data: AI Perspectives, Challenges, and Ideas at the 29th AAAI Conference on Artificial Intelligence. https://allenai.org/data/meaningful-citations

Varanasi, K.K., Ghosal, T., Tiwary, P., & Singh, M. (2021). IITP-CUNI@3C: Supervised Approaches for Citation Classification (Task A) and Citation Significance Detection (Task B). In *Proceedings of the Second Workshop on Scholarly Document Processing (SDP'21)*, 140-145. https://aclanthology.org/2021.sdp-1.19

Wan, X., & Liu, F. (2014). Are all literature citations equally Important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, *65*(9), 1929-1938. https://doi.org/10.1002/asi.23083

Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., & Chen, G. (2020). Important citation identification by exploiting the syntactic and contextual information of citations. *Scientometrics*, *125*, 2109–2129. https://doi.org/10.1007/s11192-020-03677-1

Wang, Y., & Zhang, C. (2020) Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics*, *14*(4), 101091. https://doi.org/10.1016/j.joi.2020.101091

Wang, Y., Zhang, C., & Li, K. (2022). A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics*, *127*, 2479–2520. https://doi.org/10.1007/s11192-022-04332-7

Yin, D., Tam, W. L., Ding, M., & Tang, J. (2021). MRT: Tracing the Evolution of Scientific Publications. *IEEE Transactions on Knwoeldge and Data Engineering*. Early Access. https://doi.org/10.1109/TKDE.2021.3088139

Yousif, A., Niu, Z., Chambua, J., & YounasKhana, Z. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, *335*, 195–205. https://doi.org/10.1016/j.neucom.2019.01.021

Zha, H., Chen, W., Li, K., & Yan, X. (2019). Mining Algorithm Roadmap in Scientific Publications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*, 1083–1092. https://doi.org/10.1145/3292500.3330913

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503. https://doi.org/10.1002/asi.22850

Zhang, Y., Wang, Y., Sheng, Q.Z., Mahmood, A., Zhang, W.E., & Zhao, R. (2021). TDM-CFC: Towards Document-Level Multi-label Citation Function Classification. In: W. Zhang, L. Zou, Z. Maamar, & L. Chen (Eds.) *Lecture Notes in Computer Science*: *Vol 13081. Web Information Systems Engineering – WISE 2021* (pp. 363–376). Springer, Cham. https://doi.org/10.1007/978-3-030-91560-5_26

Zhang, Y., Zhao, R., Wang, Y., Chen, H., Mahmood, A., Zaib, M., Zhang W. E., & Sheng, Q. Z. (2022). Towards employing native information in citation function classification. *Scientometrics*,*127*, 6557–6577. https://doi.org/10.1007/s11192-021-04242-0

Zhao, H., Luo, Z., Feng, C., Zheng, A., & Liu, X. (2019). A Context-based Framework for Modelling the Role and Function of On-line Resource Citations in Scientific Literature. 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'19)*, 5206–5215. https://aclanthology.org/D19-1524.

Zheng, A., Zhao, H., Luo, Z., Feng, C., Liu, X., & Ye, Y. (2021). Improving On-line Scientific Resource Profiling by Exploiting Resource Citation Information in the Literature. *Information Processing & Management*, *58*(5), 102638. https://doi.org/10.1016/j.ipm.2021.102638

SUPPLEMENTARY MATERIALS

A. Details of the Re-annotation Process

Our dataset, named `Jiang2021`, was created between Dec 2020 and May 2021. According to Figure 3, the dataset creation pipeline included three steps: dataset preparation, re-annotation and post-processing.

A.1. Preparation

The whole ACL Anthology[19]. was crawled. Full texts and citation contexts of each paper were extracted using Allen AI's s2orc-doc2json tool[20], which postprocessed and transformed the output of Grobid[21] to JSON format. To ease re-annotation, two left and three right context sentences were extracted, together with the citance, to form a citation context. All sentences of all source papers were indexed using Lucene[22]. For each source citation instance, we used it to query the source sentence index to get the matched citance and its context. Then the source citation strings were matched against the citation strings in the matched citance. Both the matched citation context and citation strings were manually checked during re-annotation.

A.2. Re-annotation

Three postgraduate research students in natural language processing were recruited for re-annotation. The first author was the main annotator (C_4) who lead the annotation guideline development and refinement, designed the re-annotation protocols, and guided the three annotators (C_1–C_3) in the whole process. We re-annotated most non-neutral citation instances (excluding "Neut(ral)", "Background", "ack") from the six datasets according to Teufel et al.'s 12-class scheme (Teufel et al., 2006a; Teufel, 2010) plus a "Future" class for future work (Jurgens et al., 2018). The instances of all "CoCo" and "Weak" classes of `Teufel2010` were directly copied. This was because C_4 had already done this part of re-annotation before the whole project started using the partial annotation guidelines Teufel publicised in her monograph (Teufel, 2010), and got almost 100% consistent results with the latter. Neutral instances of the final dataset consisted of "Neut" instances from `Teufel2010` and instances from all six datasets that were re-annotated to "Neut". This way implicitly down-sampled the biggest class "Neut". The final function for each sample was agreed by consensus among all four annotators. Difficult cases were discussed by all annotators and adjudicated. The re-annotation was done in three stages.

Stage 1 (guideline development). In the training process, we re-annotated the "PSup", "PSim", "PUse", "PMot", "PModi" and "PBas" instances from `Teufel2010`. Our own annotation guidelines were based on and extended from Teufel's full annotation guidelines about citation function classification (CFC) and argumentative zoning (AZ-II) that were kindly shared by the latter at the beginning of this project. Description of both schemes can be found in Teufel's monograph (Teufel, 2010). We started from Teufel's description of these functions, reached consensus among all four annotators, and drafted our annotation guidelines for these functions. We found our re-annotations were highly consistent with `Teufel2010`'s original annotation (Table A1 and Sect. A.4). This gave us confidence in the overall quality of our guidelines and re-annotations.

Stage 2 (guideline refinement). The focus was put on cognitively more difficult classes. According to Teufel (2010), there is a blurred border between some functions, such as "PUse" v.s. "PBas" when we see expressions like "following" or "based on", "PUse" v.s. "PSim" when we see expressions like "similar to" or "in the same way as". The annotation guidelines for "PSup", "PSim", "PUse", "PMot", "PModi" and "PBas" were refined at this stage. To further enrich or refine our own

[19] Association for Computational Linguistics (ACL) maintains an open repository of computational linguistics (CL) papers published in ACL-sponsored venues, called ACL Anthology: https://aclanthology.org
[20] https://github.com/allenai/s2orc-doc2json
[21] https://github.com/kermitt2/grobid
[22] https://lucene.apache.org/core/

annotation guidelines (Stage 2: guideline refinement), we further worked on Jha et al.'s "Substantiating" class to gain more insights from collectively re-annotating more examples. We chose this class because it was the most ill-defined class in the six datasets we used. Re-annotation of this "complex" class gave us quite a comprehensive coverage of different classes to help us further clear the boundaries between confusing classes in the difficult cases.

Stage 3 (re-annotation & adjudication): The three co-authors re-annotated (a large part of) the remaining non-Neutral samples. This included "Fundamental" (_Idea or _Basis) and "Background+Neg" (_MRelared, _GRelated or _SRelated) from `Dong2011`, "Criticising", "Uses" and "Basis" from `Jha2016`, "use", "bas", "wea" and "hed" from `Alvarez2017`, "Uses" (not used as "PUse" was already a large enough category so was less needed to be expand) and "Extends" from `Jurgens2018` (Future was directly copied because of 100% agreement of re-annotation done by C_4), and "CoCo", "Weak" and "Pos" from `Su2019`. The main author re-annotated all instances and reached consensus with each co-annotator. A large portion of these samples were re-annotated to a semantically different function according to Teufel's 12-class scheme, which implied the incomparability of the results in different CFC papers. However, after careful guideline refinement and more collective practice, the four annotators greatly improved the re-annotation stability in the last stage (Table A2 and Sect. A.4).

A.3. Post-processing

After re-annotation, we merged consecutive citation strings in each citance into a citation segment, represented by a pseudoword "CITSEG". For example, the citance "SHRDLU (Winogard, 1973) was intended to address this problem." would be tokenized and rewritten to "["SHRDLU", "(", "CITSEG", ")", "was", "intended", "to", "address", "this", "problem", "."]". We performed segment-wise CFC for each CITSEG because citations in the same CITSEG must have the same function. As a result, our dataset `Jiang2021` gathered in total 3356 citation contexts, 4784 in-text citations, and 3854 CITSEGs. Note that only `Teufel2010` annotated implicit citations represented by author names, which we left as future work.

A.4. Re-Annotation Reliability

Although all annotations were agreed by consensus among the four annotations, we still report the inter-annotator agreement (IAA) among the guideline development and final annotation stages. By doing this, we did not purpose to prove the quality of our re-annotation. Instead, we think the emphasis should be more put on the quality of the annotation guidelines, especially Teufel's comprehensive annotation guidelines (Teufel, 2010) that were later extensively enriched by us. The first author, as the main annotator (C_4), lead the annotation guideline development and refinement, designed the re-annotation protocols, and guided the three annotators (C_1–C_3) in the whole process.

In the annotation guideline development stage, all four annotators were required to annotate all samples and collaborated in a flat structure. Four annotators each re-annotated all samples independently and sat together to discuss if either C_4 disagreed with others or there was no majority-based "consensus" with C_4 (i.e., three annotators, inclusive of C_4, agreed on the instance). Even there was such "automatic consensus" based on majority voting, each disputed case was still carefully checked. Table A1 shows that the four annotators reached very high agreement on the selected classes of `Teufel2010` (see the "Stage 1" column). We must be very cautious when interpreting the high IAA in this stage. We believe it mainly reflected the high quality of Teufel's annotation guidelines and manual annotations (Teufel et al., 2006a; Teufel, 2010). Meanwhile, it also reflected that consensual understanding of Teufel's annotation guidelines were established through the collective learning process. Table A2 shows the pair-wise agreement between the four annotators (see the "Stage 1" column). The high agreements show that all three coders C_1–C_3 were able to establish good understandings of the verified/re-annotated samples with C_4, and at the same time brought a lot of value insights for annotation development and refinement to complement each other.

In the re-annotation and adjudication stage, the four annotators worked in a two-level structure. Thus, there was two-level consensus to ensure quality and accelerate annotation. The samples were split into three folds. For each fold, one of C_1–C_3 was

assigned as the responsible (secobd) annotator. Because annotators performed much better in verifying annotations (proved in the previous stages), the remaining two annotators were asked to reach consensus with the responsible annotator of the fold. This was Level-1 consensus. The main annotator of the project C_4 annotated all samples. After that, whenever there was a dispute on a sample, consensus was reached by discussion among all four annotators. This was Level-2 consensus. Table A2 (the "Stage 3" column) shows the pair-wise agreements between annotators C_1–C_3 and C_4. Because annotations made by C_1–C_3 were already adjudicated with the other two before being compared with annotations made by C_4, each sample was in effect annotated by all four annotators. In this sense, we somehow treated C_1–C_3 as one single coder, denoted as $C_{1\text{-}3}$ (Table A2) and calculated its agreement with the main coder C_4. Again, we feel obliged to warn that the results presented here are not comparable to other studies because this study focused on re-annotating the instances of existing datasets that we believe are partially mappable to Teufel et al.'s annotation scheme. We argue that our re-annotation task should be easier than annotating a dataset from scratch. In addition, we aimed at agreeing on each disputed instance, so IAA was not the focus of our design of re-annotation protocol.

Table A1. Inter-Annotator Agreement.

Stage 1: Guideline Development				Stage 3: Re-annotation & Adjudication			
Category	κ	Category	κ	Category	κ	Category	κ
13-class	0.8252	9-class	0.8529	13-class	0.6358	9-class	0.6614
Neut	0.2162	Neutral	0.9433	Neut	0.5718	Neutral	0.6268
CoCoXY	0.7521			CoCoXY	0.5593		
Future	0.1512	Future	0.1512	Weak	0.6406	Weakness	0.6406
Weak	x	Weakness	x	Future	1.0000	Future	1.0000
PSup	0.7508	Support	0.9702	PSup	0.6692	Support	0.6692
PSim	0.9063	Similar	0.9063	PSim	0.7176	Similar	0.7176
PUse	0.8552	Usage	0.8552	PUse	0.7735	Usage	0.7735
PMot	0.9577	Motivation	0.9553	PMot	0.5073	Motivation	0.5073
PModi	0.9122	Basis	0.7688	PModi	0.4631	Basis	0.6318
PBas	0.5186			PBas	0.6534		
CoCo-	x	Comparison	x	CoCo-	0.4792	Comparison	0.6306
CoCoR0	--			CoCoR0	0.4174		
CoCoGM	--			CoCoGM	0.4124		

Note: "x" means no report due to extremely small sample size, and "--" means no sample at all for reporting

Table A2. Pair-Wise Inter-Annotator Agreement.

	Stage 1: Annotation Guideline Development			Stage 3: Re-annotation & Adjudication			
	C_1 v.s. C_4	C_2 v.s. C_4	C_3 v.s. C_4	C_1 v.s. C_4	C_2 v.s. C_4	C_3 v.s. C_4	$C_{1\text{-}3}$ v.s. C_4
13-class	0.7959	0.8194	0.8332	0.6344	0.6413	0.5107	0.6358
9-class	0.8480	0.8204	0.8391	0.6758	0.6714	0.5350	0.6614

Note: $C_{1\text{-}3}$ is the pseudo-coder by treating coders C_1-C_3 as a single second coder and merging all three splits.

B. Re-arranged Views of Citation Function Classification Performances

Table B1. Citation Function Classification Performances Re-Arranged to Investigate the Impact of Sentence Encoder

Model options						Macro F1 (%)													
Model	citseg	ctx_type	Encoding methods			11-class				9-class				7-class				6-class			
			citance	context	sentence	best	avg	std		best	avg	std		best	avg	std		best	avg	std	
hie-03	O	hierarchical	max_pool	max_pool	SEP	63.30	63.30	1.12		65.39	63.24	1.40	♣	69.18	67.35	1.50		71.71	69.60	1.36	
hie-07	O	hierarchical	max_pool	max_pool	max_pool	62.63	62.16	0.51		62.37	61.25	1.00		70.76	68.71	1.60		70.22	67.94	1.38	
hie-09	O	hierarchical	max_pool	max_pool	self_attnd	63.38	62.45	0.59	↑	64.69	62.92	1.16	↑	69.38	67.66	1.49	↓	72.11	70.07	1.8	↑
hie-04	O	hierarchical	max_pool	self_attnd	SEP	63.79	63.79	1.71		63.12	61.95	1.60		70.00	67.76	1.73	♣	72.10	70.25	1.69	♣
hie-08	O	hierarchical	max_pool	self_attnd	max_pool	65.02	62.10	2.24		**67.49**	64.51	1.97		69.53	67.47	1.73		69.77	68.24	1.33	
hie-10	O	hierarchical	max_pool	self_attnd	self_attnd	63.31	62.44	0.89	↓	65.45	63.11	2.21	↓	69.45	68.75	0.41	↓↑	71.40	70.02	1.03	↑
hie-05	O	hierarchical	self_attnd	max_pool	SEP	63.69	63.69	2.21		64.96	62.95	1.50		67.77	66.39	0.84		70.09	67.83	1.74	
hie-11	O	hierarchical	self_attnd	max_pool	max_pool	64.46	62.17	1.99		64.00	62.80	1.62		68.76	67.09	1.50		72.38	69.33	3.07	
hie-13	O	hierarchical	self_attnd	max_pool	self_attnd	64.99	63.56	1.15	↑	64.97	63.97	0.80	↑	70.10	67.99	1.88	↑	71.49	69.52	1.66	↓
hie-06	O	hierarchical	self_attnd	self_attnd	SEP	63.79	63.79	1.71	♣	63.12	61.95	1.60		70.00	67.76	1.73		72.10	70.25	1.69	♣
hie-12	O	hierarchical	self_attnd	self_attnd	max_pool	63.43	62.26	0.83		65.36	64.28	0.97		69.66	68.27	1.60		70.78	69.56	1.57	
hie-14	O	hierarchical	self_attnd	self_attnd	self_attnd	63.16	62.09	1.02	↓	65.69	64.44	1.29	↑	**72.81**	69.47	2.64	↑	71.32	68.35	2.22	↑
hie-15	O	hierarchical	X	max_pool	SEP	61.17	59.98	1.14		65.53	63.07	1.66		68.71	67.12	1.45		73.24	70.19	2.41	♣
hie-17	O	hierarchical	X	max_pool	max_pool	64.56	64.16	0.39		65.96	62.81	2.29		69.35	67.96	1.31		70.90	70.04	0.94	
hie-19	O	hierarchical	X	max_pool	self_attnd	62.62	61.61	1.18	↓	65.35	64.16	1.08	↓↑	**72.39**	68.40	2.47	↑	71.89	70.48	1.04	↑
hie-16	O	hierarchical	X	self_attnd	SEP	63.22	62.25	0.89		65.24	63.79	1.09		69.57	67.97	1.90		71.56	70.40	1.18	
hie-18	O	hierarchical	X	self_attnd	max_pool	64.95	62.82	1.64		66.07	63.76	1.56		70.05	68.87	0.97		72.09	69.35	2.11	
hie-20	O	hierarchical	X	self_attnd	self_attnd	63.15	62.39	0.60	↓	66.25	63.79	1.97	↑	70.88	69.54	1.10	↑	70.72	69.75	1.1	↓

Table B2. Citation Function Classification Performances Re-Arranged to Investigate the Impact of Citance Encoder

Model options						Macro F1 (%)													
Model	citseg	ctx_type	Encoding methods			11-class				9-class				7-class				6-class			
			citance	context	sentence	best	avg	std		best	avg	std		best	avg	std		best	avg	std	
seq-01	O	sequential	max_pool	CLS	N/A	63.93	62.72	1.11		66.53	63.89	1.94		70.70	69.03	1.45		**74.03**	70.88	1.87	
seq-04↓	O	sequential	self_attnd	CLS	N/A	63.12	62.07	1.00	↓	65.16	63.86	1.03	↓	68.56	67.54	1.46	↓	69.96	68.22	1.58	↓
seq-02	O	sequential	max_pool	max_pool	N/A	63.21	62.61	0.45		64.84	63.60	1.08		71.39	68.13	1.89		70.23	68.25	1.60	
seq-05	O	sequential	self_attnd	max_pool	N/A	64.12	62.82	1.20	↑	64.69	64.19	0.47	↓	68.86	66.80	1.62	↑	71.56	69.05	1.85	↑
seq-03	O	sequential	max_pool	self_attnd	N/A	64.26	62.82	1.04		65.61	63.66	1.91		70.19	69.24	0.64		70.99	68.86	1.71	
seq-06	O	sequential	self_attnd	self_attnd	N/A	**65.12**	63.05	1.60	↑	64.84	62.52	1.48	↓	70.63	69.16	1.43	↑	72.19	69.81	1.37	↑
seq-10	O	sequential	max_pool	X	N/A	63.93	62.72	1.11		66.19	63.72	2.74		69.16	67.87	1.85		71.89	70.18	1.77	
seq-11	O	sequential	self_attnd	X	N/A	64.42	63.01	0.89	↑	66.92	64.58	1.45	↑	68.83	67.22	1.75	↓	71.32	69.69	1.01	↓
seq-x10	X	sequential	max_pool	X	N/A	64.09	62.23	1.70		65.04	63.62	1.6		68.68	67.85	0.62		***73.52***	69.31	3.12	
seq-x11	X	sequential	self_attnd	X	N/A	64.38	62.46	1.13	↑	***67.08***	64.21	2.38	↑	69.34	67.31	1.90	↑	69.48	68.85	0.59	↓
cita-01	X	citance	CLS	N/A	N/A	58.16	56.20	1.64		60.30	58.75	1.38	♣	60.30	58.75	1.38	♣	63.58	62.39	1.16	♣
cita-02	X	citance	max_pool	N/A	N/A	57.47	55.77	1.36		59.07	58.00	1.06		59.07	58.00	1.06		63.88	61.81	1.58	
cita-03	X	citance	self_attnd	N/A	N/A	59.49	58.13	1.11	↑	56.99	56.01	1.17	↓	56.99	56.01	1.17	↓	62.54	61.51	0.95	↓
hie-01	O	hierarchical	SEP	max_pool	SEP	62.78	61.76	0.89		65.39	63.24	1.40	♣	69.18	67.35	1.50	♣	69.39	68.42	1.25	
hie-03	O	hierarchical	max_pool	max_pool	SEP	63.30	63.30	1.12		65.39	63.24	1.40		69.18	67.35	1.50		71.71	69.60	1.36	
hie-05	O	hierarchical	self_attnd	max_pool	SEP	63.69	63.69	2.21	↑	64.96	62.95	1.50	↓	67.77	66.39	0.84	↓	70.09	67.83	1.74	↓
hie-02	O	hierarchical	SEP	self_attnd	SEP	61.42	61.42	0.96		63.12	61.95	1.60		70.00	67.76	1.73		71.08	69.87	1.51	
hie-04	O	hierarchical	max_pool	self_attnd	SEP	63.79	63.79	1.71		63.12	61.95	1.60		70.00	67.76	1.73		72.10	70.25	1.69	
hie-06	O	hierarchical	self_attnd	self_attnd	SEP	63.79	63.79	1.71	↑↓	63.12	61.95	1.60	↑↓	70.00	67.76	1.73	↑↓	72.10	70.25	1.69	↑↓
hie-07	O	hierarchical	max_pool	max_pool	max_pool	62.63	62.16	0.51		62.37	61.25	1.00		70.76	68.71	1.60		70.22	67.94	1.38	
hie-11	O	hierarchical	self_attnd	max_pool	max_pool	64.46	62.17	1.99	↑	64.00	62.80	1.62	↑	68.76	67.09	1.50	↓	72.38	69.33	3.07	↑
hie-09	O	hierarchical	max_pool	max_pool	self_attnd	63.38	62.45	0.59		64.69	62.92	1.16		69.38	67.66	1.49		72.11	70.07	1.8	
hie-13	O	hierarchical	self_attnd	max_pool	self_attnd	64.99	63.56	1.15	↑	64.97	63.97	0.80	↑	70.10	67.99	1.88	↑	71.49	69.52	1.66	↓
hie-08	O	hierarchical	max_pool	self_attnd	max_pool	65.02	62.10	2.24		**67.49**	64.51	1.97		69.53	67.47	1.73		69.77	68.24	1.33	
hie-12	O	hierarchical	self_attnd	self_attnd	max_pool	63.43	62.26	0.83	↓	65.36	64.28	0.97	↓	69.66	68.27	1.60	↑	70.78	69.56	1.57	↑
hie-10	O	hierarchical	max_pool	self_attnd	self_attnd	63.31	62.44	0.89		65.45	63.11	2.21		69.45	68.75	0.41		71.40	70.02	1.03	
hie-14	O	hierarchical	self_attnd	self_attnd	self_attnd	63.16	62.09	1.02	↓	65.69	64.44	1.29	↑	**72.81**	69.47	2.64	↑	71.32	68.35	2.22	↓
hie-21	O	hierarchical	SEP	X	N/A	63.48	61.27	1.39	♣	65.36	64.11	0.96		69.81	68.19	1.04	♣	72.81	70.96	1.32	♣
hie-22	O	hierarchical	max_pool	X	N/A	63.48	61.27	1.39		66.88	63.38	2.06		69.47	67.89	1.93		72.81	70.96	1.32	
hie-23	O	hierarchical	self_attnd	X	N/A	62.55	61.09	1.05	↓	64.60	61.89	1.56	↓	68.82	66.55	2.23	↓	70.38	69.28	1.19	↓

C. Further Discussions about The Support Class

According to Teufel's annotation rules, "Support"/"PSup" has two meanings: "mutual compatibility" between knowledge statements (or viewed as conceptual compatibility) and "computational plug-in-ability" of approaches to each other (or viewed as technical compatibility). This might be the potential cause for its low recognition rate. We did a few additional experiments by re-annotating this class into other categories. The majority fell into "Similar" (for "mutual compatibility") and "Neutral" (for "computational plug-in-ability"; not into "Usage" because not actually used). This resulted in a 10-class scheme (Table C1). Table C2 presents the CFC performances. The best F1 was significantly improved over the 11-class scheme to 68.93% after re-annotating "Support", even higher than the 9-class scheme. We guess that the significant performance gap between the 9-class and 10-class schemes was due to two factors: (i) The poor performance of the "Support" class on the 9-class scheme; and (ii) the better performance of the comparison functions on the 10-class scheme (mean F1 of "CoCoGM" and "CoCoRes") compared to the 11-class and 9-class schemes (refer to the per-class performances in Table C3).

The 10-class scheme was further reduced to 8-class by merging "CoCoGM" and "CoCoRes" into "Comparison" and merging "CoCoXY" into "Neutral". By comparing the 10-class (resp. 8-class) against 11-class (resp. 9-class) schemes, we see the former improved the overall performance over the latter by a large margin. One conclusion we can make is that "**Support" should better be re-annotated if it is not the focus** of the downstream application. At the same time, however, we also observe performance drop for the "Similar" class on the 8-class scheme compared to on the 9-class scheme. On the other hand, the "mutual compatibility" meaning of "Support" is an important relationship between knowledge claims of biomedical papers (Li et al., 2013; Meyers, 2013). The same applies to the contradiction relationship, e.g., "Conflict" in Agarwal, et al. (2010) and "Anti-Support" in Teufel (2010). To reflect this, our second conclusion is that **if "Support" is the focus of study, we must develop a bespoke citation function classification model for it and focus on its "mutual compatibility" meaning**. Recently, Nicholson et al., (2022) made a significant contribution to the annotation and classification of "supporting" v.s. "contrasting" relationships. Unfortunately, their proprietary dataset is not publicly accessible. We leave the annotation and recognition of these two functions to future work. Note that, "Support" (at its "mutual compatibility between scientific claims" meaning) and "Anti-Support" are very small classes, so we expect to apply semi-supervised learning and few-shot learning techniques to developing efficient machine learning CFC models for them in our future work.

Table C1. Citation function scheme mapping and CITSEG-level statistics after Re-annotating "PSup"/"Support"

Teufel2010+ (12+1 class)				Jiang2021 (11-class)			Jiang2021 (10-class)			Jiang2021 (8-class)			Jurgens2018 (6-class)		
label	size		ratio	label	size	ratio	label	size	ratio	label	size	ratio	label	size	ratio
	citstr	citseg	citseg												
Future	97	85	2.21%	Future	85	2.21%	Future	89*	2.31%	Future	89	2.31%	Future	89	2.31%
CoCoXY	200	152	3.94%	CoCoXY	152	3.94%	CoCoXY	153	3.97%	Background	1673	43.41%	Background	1670	43.38%
Neut	1924	1463	37.96%	Neutral	1463	37.96%	Neutral	1520	39.44%						
PSup	123	100	2.59%	Support	100	2.59%									
PSim	247	207	5.37%	Similar	207	5.37%	Similar	235	6.10%	Similar	235	6.10%	ComOrCon	877	22.78%
Weak	223	158	4.10%	Weakness	158	4.10%	Weakness	158	4.10%	Weakness	158	4.10%			
CoCoGM	390	299	7.76%	CoCoGM	328	8.51%	CoCoGM	328	8.51%	Comparison	485	12.58%			
CoCo-	108	80	2.08%												
CoCoR0	107	100	2.59%	CoCoRes	151	3.92%	CoCoRes	157	4.07%						
PMot	365	288	7.47%	Motivation	288	7.47%	Motivation	289	7.50%	Motivation	289	7.50%	Motivation	289	7.52%
PUse	794	755	19.59%	Usage	755	19.59%	Usage	758	19.67%	Usage	758	19.67%	Usage	755	19.59%
PModi	72	65	1.69%	Basis	167	4.33%	Basis	167	4.33%	Basis	167	4.33%	Basis	167	4.33%
PBas	134	102	2.65%												
Total	4784	3854		3854			3854			3854			3854		

* A small number of "Support" instances were reannotated to classes other than "Neutral" or "Similar", e.g., Future" for potential "computational plug-in-ability".

Table C2. Citation Function Classification Performances with after Re-annotating "PSup"/"Support"

Model options						Macro F1 (%)														
Model	citseg	ctx_type	Encoding methods			11-class w/ "Support"			10-class w/o "Support"			9-class w/ "Support"			8-class w/o "Support"			7-class "Support" → "Similar"		
			citance	context	sentence	best	avg	std	best	avg	std	best	avg	std	best	avg	std	best	avg	std
seq-01	O	sequential	max_pool	CLS	N/A	63.93	62.72	1.11	67.01	65.24	1.43	66.53	63.89	1.94	67.98	66.20	1.41	70.70	69.03	1.45
seq-02	O	sequential	max_pool	max_pool	N/A	63.21	62.61	0.45	66.86	65.47	1.40	64.84	63.60	1.08	68.23	66.66	1.15	71.39	68.13	1.89
seq-03	O	sequential	max_pool	self_attnd	N/A	64.26	62.82	1.04	**68.93**	66.42	2.20	65.61	63.66	1.91	**70.14**	67.05	2.50	70.19	69.24	0.64
seq-04	O	sequential	self_attnd	CLS	N/A	63.12	62.07	1.00	65.53	64.76	0.48	65.16	63.86	1.03	67.18	66.30	1.03	68.56	67.54	1.46
seq-05	O	sequential	self_attnd	max_pool	N/A	64.12	62.82	1.20	67.63	66.39	0.82	64.69	64.19	0.47	68.59	67.26	1.47	68.86	66.80	1.62
seq-06	O	sequential	self_attnd	self_attnd	N/A	**65.12**	63.05	1.60	66.77	65.88	0.94	64.84	62.52	1.48	66.75	65.72	0.89	70.63	69.16	1.43
seq-07	O	sequential	X	CLS	N/A	64.65	61.01	2.21	66.96	65.10	1.64	65.38	62.20	1.78	68.75	66.03	1.89	70.35	68.28	1.33
seq-08	O	sequential	X	max_pool	N/A	**66.16**	63.53	1.55	68.23	65.33	2.02	66.03	62.98	2.05	**70.27**	67.78	2.24	69.89	67.98	1.90
seq-09	O	sequential	X	self_attnd	N/A	63.92	62.80	0.89	**68.40**	65.60	1.61	65.41	64.18	0.75	67.23	65.11	2.38	70.80	69.78	0.85
seq-10	O	sequential	max_pool	X	N/A	63.93	62.72	1.11	67.36	65.65	1.05	66.19	63.72	2.74	69.23	67.06	1.94	69.16	67.87	1.85
seq-11	O	sequential	self_attnd	X	N/A	64.42	63.01	0.89	67.64	66.45	1.19	66.92	64.58	1.45	66.75	66.02	0.63	68.83	67.22	1.75
seq-12	O	sequential	X	X	N/A	64.93	63.50	1.04	67.47	66.24	0.98	**67.78**	64.74	1.88	68.05	67.16	1.05	70.65	69.28	1.30
seq-x10	X	sequential	max_pool	X	N/A	64.09	62.23	1.70	66.80	65.16	1.19	65.04	63.62	1.6	68.04	66.38	1.09	68.68	67.85	0.62
seq-x11	X	sequential	self_attnd	X	N/A	64.38	62.46	1.13	66.32	64.37	1.39	***67.08***	64.21	2.38	67.52	65.63	1.73	69.34	67.31	1.90
hie-03	O	hierarchical	max_pool	max_pool	SEP	63.30	63.30	1.12	65.84	64.26	1.23	65.39	63.24	1.40	67.35	66.01	1.52	69.18	67.35	1.50
hie-04	O	hierarchical	max_pool	self_attnd	SEP	63.79	63.79	1.71	66.34	64.67	1.28	63.12	61.95	1.60	68.41	65.79	1.65	70.00	67.76	1.73
hie-07	O	hierarchical	max_pool	max_pool	max_pool	62.63	62.16	0.51	**68.41**	65.65	1.73	62.37	61.25	1.00	65.83	65.00	0.90	70.76	68.71	1.60
hie-08	O	hierarchical	max_pool	self_attnd	max_pool	65.02	62.10	2.24	67.41	65.87	1.34	**67.49**	64.51	1.97	68.61	67.94	0.70	69.53	67.47	1.73
hie-09	O	hierarchical	max_pool	max_pool	self_attnd	63.38	62.45	0.59	64.71	63.22	1.76	64.69	62.92	1.16	67.66	65.39	2.42	69.38	67.66	1.49
hie-10	O	hierarchical	max_pool	self_attnd	self_attnd	63.31	62.44	0.89	66.91	64.59	2.06	65.45	63.11	2.21	66.71	66.24	0.57	69.45	68.75	0.41
hie-13	O	hierarchical	self_attnd	max_pool	self_attnd	64.99	63.56	1.15	67.09	65.38	1.59	64.97	63.97	0.80	68.44	66.98	1.39	70.10	67.99	1.88
hie-14	O	hierarchical	self_attnd	self_attnd	self_attnd	63.16	62.09	1.02	66.63	64.77	1.54	65.69	64.44	1.29	67.93	66.09	1.75	**72.81**	69.47	2.64
hie-15	O	hierarchical	X	max_pool	SEP	61.17	59.98	1.14	66.62	64.77	1.13	65.53	63.07	1.66	68.45	66.22	1.93	68.71	67.12	1.45
hie-16	O	hierarchical	X	self_attnd	SEP	63.22	62.25	0.89	65.67	64.69	0.57	65.24	63.79	1.09	68.07	66.58	1.35	69.57	67.97	1.90
hie-17	O	hierarchical	X	max_pool	max_pool	64.56	64.16	0.39	66.23	65.48	1.32	65.96	62.81	2.29	68.90	66.29	1.54	69.35	67.96	1.31
hie-18	O	hierarchical	X	self_attnd	max_pool	64.95	62.82	1.64	66.78	65.68	1.08	66.07	63.76	1.56	67.89	66.28	1.01	70.05	68.87	0.97
hie-19	O	hierarchical	X	max_pool	self_attnd	62.62	61.61	1.18	64.79	64.12	0.48	65.35	64.16	1.08	68.02	66.45	1.22	**72.39**	68.40	2.47
hie-20	O	hierarchical	X	self_attnd	self_attnd	63.15	62.39	0.60	65.97	64.76	0.97	66.25	63.79	1.97	66.37	65.39	0.60	70.88	69.54	1.10
hie-21	O	hierarchical	SEP	X	N/A	63.48	61.27	1.39	67.81	64.64	1.95	65.36	64.11	0.96	67.98	66.61	1.28	69.81	68.19	1.04
hie-22	O	hierarchical	max_pool	X	N/A	63.48	61.27	1.39	67.81	64.64	1.95	66.88	63.38	2.06	67.98	66.61	1.28	69.47	67.89	1.93
hie-23	O	hierarchical	self_attnd	X	N/A	62.55	61.09	1.05	66.10	64.13	1.69	64.60	61.89	1.56	69.49	66.74	1.71	68.82	66.55	2.23
hie-24	O	hierarchical	X	X	N/A	64.37	62.80	1.51	65.35	64.17	1.19	65.68	64.97	0.73	68.27	67.81	0.32	70.44	69.01	1.29

Models hie-01/02, hie-05/06, and hie-11/12 are removed because no model in either group appeared to be a top-3 model on any annotation scheme.

Table C3. Per-Class Performances of Selected Models after Reannotating "PSup"/"Support"

		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	ID (seed)	seq-08 (5171)			seq-06 (47353)			hie-18 (13249)			hie-08 (32491)			seq-12 (5171)			seq-11 (47353)			best of all models		
11-class	Macro Avg	68.50	65.17	66.16	67.74	64.05	65.12	67.10	64.17	64.95	65.18	65.31	65.02	65.59	65.11	64.93	65.28	64.00	64.42	--	--	--
	CoCoGM	67.11	77.27	**71.83**	70.97	66.67	68.75	72.41	63.64	67.74	63.38	68.18	65.69	62.16	69.70	65.71	67.65	69.70	68.66	67.11	77.27	**71.83**
	CoCoRes	65.62	67.74	**66.67**	62.50	80.65	70.42	81.82	58.06	67.92	63.64	67.74	65.52	75.76	80.65	**78.12**	62.50	80.65	70.42	80.00	77.42	**78.69**
	Similar	67.74	50.00	57.53	60.87	66.67	63.64	60.00	57.14	58.54	68.42	61.90	**65.00**	59.18	69.05	63.74	67.57	59.52	63.29	71.05	64.29	**67.50**
	Support	46.15	30.00	36.36	46.15	30.00	36.36	34.62	45.00	39.13	36.84	35.00	35.90	38.10	40.00	39.02	42.11	40.00	**41.03**	47.53	55.00	**51.16**
	ID (seed)	seq-03 (25603)			hie-04 (25603)			seq-09 (13249)			seq-08 (5171)			seq-11 (5171)			seq-12 (25603)			best of all models		
10-class	Macro Avg	72.00	66.60	68.93	68.90	68.42	68.41	70.28	67.34	68.40	70.76	66.44	68.23	68.48	67.19	67.64	70.78	65.58	67.47	--	--	--
	CoCoGM	58.75	71.21	64.38	69.01	74.24	71.53	76.27	68.18	**72.00**	63.89	69.70	66.67	65.15	65.15	65.15	63.16	72.73	67.61	70.83	77.27	**73.91**
	CoCoRes	80.77	65.62	72.41	77.42	75.00	**76.19**	60.00	75.00	66.67	69.70	71.88	70.77	65.79	78.12	71.43	61.54	75.00	67.61	75.68	87.50	**81.16**
	Similar	68.18	61.22	64.52	59.26	65.31	62.14	71.43	61.22	**65.93**	59.18	59.18	59.18	57.78	53.06	55.32	64.44	59.18	61.70	70.45	63.27	**66.67**
	ID (seed)	seq-12 (47353)			hie-08 (47353)			seq-11 (47353)			hie-18 (13491)			seq-08 (32491)			hie-14 (5171)			best of all models		
9-class	Macro Avg	69.25	67.13	67.78	67.51	67.89	67.49	68.24	66.31	66.92	72.21	62.51	66.07	70.71	63.02	66.03	67.90	64.62	65.69	--	--	--
	Comparison	68.81	80.41	**77.23**	63.25	76.29	69.16	64.55	73.20	68.60	60.00	71.13	65.09	67.31	72.16	69.65	67.77	84.54	75.23	76.47	61.90	**68.42**
	Similar	62.79	64.29	63.53	61.36	64.29	62.79	60.87	66.67	**63.64**	64.86	57.14	60.76	58.54	57.14	57.83	57.78	61.90	59.77	79.17	65.52	**71.70**
	Support	45.45	50.00	47.62	50.00	55.00	52.38	52.94	45.00	**48.65**	30.00	30.00	30.00	61.54	40.00	48.48	46.67	35.00	40.00	66.67	40.00	**50.00**
	ID (seed)	seq-08 (13249)			seq-03 (47353)			hie-23 (5171)			hie-08 (32491)			seq-02 (5171)			seq-12 (25603)			best of all models		
8-class	Macro Avg	71.91	69.21	70.27	72.33	68.51	70.14	71.93	68.35	69.49	67.63	70.19	68.61	70.82	67.61	68.23	69.48	66.77	68.05	--	--	--
	Comparison	67.27	75.51	71.15	72.00	73.47	72.73	63.87	75.55	70.05	66.67	75.51	70.81	68.42	79.59	**73.58**	73.20	72.45	72.82	77.32	76.53	**76.92**
	Similar	70.73	59.18	64.44	62.26	67.35	**64.71**	65.85	55.10	60.00	62.50	61.22	61.86	60.00	48.98	53.93	58.70	55.10	56.84	69.57	65.31	**67.37**