Scientometrics

Contextualised Segment-Wise Citation Function Classification

--Manuscript Draft--

Manuscript Number:	
Full Title:	Contextualised Segment-Wise Citation Function Classification
Article Type:	Manuscript
Keywords:	Citation context analysis; citation function classification; Deep Learning; SciBERT; ensemble
Corresponding Author:	Xiaorui Jiang Coventry University Birmingham, West Midlands UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Coventry University
Corresponding Author's Secondary Institution:	
First Author:	Xiaorui Jiang
First Author Secondary Information:	
Order of Authors:	Xiaorui Jiang
	Jingqiang Chen
Order of Authors Secondary Information:	
Funding Information:	National Social Science Fund of China (18ZDA238) — Dr. Xiaorui Jiang
Abstract:	Much effort has been made in the past decades to citation function classification. Noteworthy issues exist. Annotation difficulty made existing datasets quite limited in size, especially for minority classes, and quite limited in the representativeness of a scientific domain. Different annotation schemes made existing studies not easily mappable and comparable. 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. In addition to achieving the new state of the art of citation function classification, this study focused on deeper performance analysis of to answer several research questions about the effective ways of performing citation function classification, more specifically, the necessity of modelling in-text citations in context and doing citation function classification at citation (segment) level. A particular emphasis was placed on in-depth per-class performance analysis for the purpose of understanding whether citation function classification is robust enough for scientometric applications, what implications can be derived for the applicability of citation function classification to different scientometric analysis tasks, and what further efforts are required to meet such analytic needs.

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. Noteworthy issues exist. Annotation difficulty made existing datasets quite limited in size, especially for minority classes, and quite limited in the representativeness of a scientific domain. Different annotation schemes made existing studies not easily mappable and comparable. 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. In addition to achieving the new state of the art of citation function classification, this study focused on deeper performance analysis of to answer several research questions about the effective ways of performing citation function classification, more specifically, the necessity of modelling in-text citations in context and doing citation function classification at citation (segment) level. A particular emphasis was placed on in-depth per-class performance analysis for the purpose of understanding whether citation function classification is robust enough for scientometric applications, what implications can be derived for the applicability of citation function classification to different scientometric analysis tasks, and what further efforts are required to meet such analytic needs.

Keywords

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

1. Introduction

Citation context analysis (Zhang et al., 2013) is an important task in scientific text understanding with rich downstream applications envisioned in Ding et al., (2014). In addition to the cited content of a referenced paper, citation context also reveals the citing authors' motivation to cite a paper, i.e. citation function (Teufel et al., 2006b), a.k.a. citation role, intent, or motivation etc. For instance, in Example 1 in Figure 1, the first citation "Prince et al., 1993" describes the weakness ("Weak") of the cited work by Prince et al., while the second citation "Kisseberth, 1970" is simply a neutral citation, here meaning no specific intellectual relations with the citing paper but merely a mention or acknowledgement of an existing past study.

The past two decades have witnessed rich machine learning algorithms for citation function classification (CFC; Teufel et al., 2006b; Agarwal et al., 2010; Dong & Schäfer, 2011; Jochim & Schütze, 2012; Abu-Jbara et al., 2013; Iorio et al., 2013; Li et al., 2013; Jha et al., 2016; Hernández-Alvarez et al., 2017; Meng et al., 2017; Jurgens et al, 2018). See Iqbal et al. (2021) for a comprehensive review. The state-of-the-art (SOTA) of this research topic has been significantly advanced by deep learning in term of classification accuracy (Cohan et al., 2019; Beltagy et al., 2019). On a dataset with 6-class annotation scheme, the CFC performance has been improved from 54.9% macro F1 by the state-of-the-art feature engineering approach (Jurgens et al., 2018) to 67.9% by Cohan et al. (2019)[] and 70.98% by Beltagy et al. (2019). CFC performances on specific academic entities like algorithm (Turado et al., 2021) or resource (Zhao et al, 2019; Zheng et al., 2021) can be higher, but they are not the focus of this paper.

Example 1: "Weak(ness)" and "Neut(ral)" citations appear in the same 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...)

Example 2: "PSim" (similarity) and "Neut" citations appear in the same citance. Context sentence S-2 is needed to infer the functions of the first two citations in the citance S-1 (forming a citation segment and having the same function).
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.

Figure 1: Examples of Citation Function Classification where Multiple In-Text Citations Have Different Functions. (Examples Taken from Teufel's Annotation Guideline).

1.1. Issues with Citation Function Annotations and Datasets

Several noteworthy issues exist in past studies, which inspired the research questions to be answered by the current paper. Firstly, almost each study in the past used a different citation function annotation scheme, ranging from Teufel et al.'s most comprehensive 12 classes (Teufel et al., 2006a) to the drastically reduced scheme consisting of only three classes – Background, Method, and Result of the popular `scicite` dataset (Cohan et al., 2019; Beltagy et al., 2019) and its extension in Zhang et al. (2022). There was little work discussing and experimenting on these different annotation schemes (detailed in Sect. 2.1 and 3.2). We argue that it is time and important to understand how well CFC performs and how well CFC serves different scientometric analysis tasks. Usually, existing datasets are also limited in data size due to annotation difficulty. Minority classes typically have only a few dozens of samples. For example, in Teufel et al. (2006a, 2006b), the two most important classes, "PModi" (technical modification of cited work) and "PBas" (ideationally based on cited work), both only have 60 instances. In Jurgens et al. (2018), the "Extends" class, equivalent to "PBas" and "PModi" combined, only has 73 instances. More extreme cases are the "hed" class (criticism via hedging) in Hernández-Alvarez et al. (2017) and the "Weak" class (weakness of cited

paper) in Su et al. (2019), which have only 40 and 30 annotations respectively. This makes these datasets less feasible for training large deep learning models. Aljohani et al. (2021a) merged the datasets of Teufel et al. and Jurgens et al., however their dataset is not easy to map against other annotation schemes (further elaborated in Sect. 2.1 and 3.2). Because of this, existing studies mainly worked on their own datasets with bespoke annotation schemes. From the discussions above, our first research question is stated below. The answer to this question may allow us to create a larger and more comprehensive citation function dataset.

RQ1. What are the relationships between different citation function annotation schemes and the mappings between existing citation function datasets?

1.2. Issues with Citation Function Classification Algorithms

Concerning algorithmic classification, CFC was typically done on individual in-text citations, called citation-level CFC, although consecutive citations must have the same function. Being less discussed, most deep learning (DL) approaches generated a feature vector for the whole citation sentence (abbr. *citance*) or context, called citance-level CFC or context-level CFC, rather than modelling individual in-text citations (hereafter citations), including those reporting SOTA performances such as Cohan et al. (2019) and Beltagy et al. (2019). Early DL methods used Convolutional Neural Network (CNN; Lauscher et al., 2017; Bakhti et al., 2018; Su et al., 2019) or Bidirectional Long-Short Term Memory (BiLSTM; Munkhdalai et al., 2016; Cohan et al., 2019) as the encoder. These methods could only generate a feature vector for the whole citance or context, rather than individual citations. Recently, the SciBERT model reported impressively strong new SOTAs on a wide range of scientific text classification tasks (Beltagy et al., 2019). However, their CFC experiments were based on the sequence-level classification symbol "[CLS]", i.e., at citance or context level. In practice, it is common to see multiple citations of different functions in the same citance. Figure 1 shows two examples of such case. The above methods would assign the same citation function to these citations. Unfortunately, this is conceptually flawed. Ideally, each citation should be modelled and classified separately. Therefore, the second research question is stated as follows.

RQ2. Should CFC be performed at (in-text) citation level or at citance or context level: Which choice is empirically supported?

Some publicly available datasets only included citances and thus the DL methods only encoded citances (Cohan et al., 2019; Beltagy et al., 2019), while many datasets included contexts of several sentences and thus the DL methods encoded citations in the context (Su et al., 2019). The CNN and BiLSTM encoders typically use max-pooling and self-attention to pool a summary feature vector for either the citance or the context, while the SciBERT encoder can also rely on sequence separator "[SEP]" and sequence classification symbol "[CLS]" for pooling citance representation and context representation respectively. We can conclude that existing DL models have only explored a very limited design space for the representation learning of in-text citations. What is more, there is no systematic study of what encoding methods are the most effective for citation modelling, including the methods for encoding in-text citation, the enclosing citance, and the surrounding context, as well as whether and how citance and context encodings could help improve CFC performance. The third research question is defined as follows.

RQ3. Should citation modelling be done in its context and what are the most effective methods for encoding and utilizing the representations of citation sentence and citation context for CFC?

1.3. Issues with Using Citation Function Classification Results

Ideally, it is cognitively plausible to apply one CFC model to all the classes in an annotation scheme, as (almost) all existing studies do[1]. However, the reality might be that a CFC model does not work equally well on all citation functions. To understand this, we can think about the challenging class "CoCoXY" (comparison between two cited works) in Teufel et al.'s 12-class scheme (Figure 2). On the one hand, it bears linguistic similarity with other "CoCo" (Comparison or Contrast) classes because both use comparative expressions. A model that is good at detecting comparative expressions for the "CoCo" class may misrecognize many "CoCoXY" instances. On the other hand, a "CoCoXY" instance does not describe any relationship between the cited work(s) with the citing paper, so in this sense it is similar to "Neut" (neutral class for cases that cannot fall into other categories) and may be confused with "Neut". This is why Jurgens et al. merged it into "Background" (their neutral class). In fact, while Cohan et al.'s second best model reported very good performance on "Background" on the ACL-ARC dataset (Jurgens et al., 2019), its performance on "CompareOrContrast" was poor (Cohan et al., 2019, Table 5). Their best model was greatly improved for "CompareOrContrast" at the cost of worse performances on "Extension" (cited work is based on or extended by citing paper) and "Motivation" (cited work motivates the citing paper). From the perspective just discussed, there is a need to explore various modelling options to find not only the best CFC model in term of overall classification performance, but also the best models for different citation functions, because there are a range of scientometic tasks that work with a specific citation function.

> *RQ4. How well can a general-purpose citation function classification model suit different types of scientometric analysis tasks and what implications can we derive for the real-world application of citation function classification?*

Example 3. "CoCoXY" needs a meta-statement of comparison or contrast like S-1 below, similar for all "CoCo" (Comparison or Contrast) classes.
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.

Example 4: "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.

Figure 2: Confusing Examples of "CoCoXY": Examples from the Teufel et al.'s Annotations.

1.4. Summary of This Paper

This paper tries to answer, at least partially, the questions raised before. We also hope to provide a good benchmark of citation function classification and a set of strong baseline models which achieve new SOTAs for the purpose of facilitating the scientific community in furthering research in citation context analysis and semantics-driven scientometric and bibliometric analysis based on citation context analysis. For answers for the research question RQ1, we did a critical review of the existing citation function datasets and their annotation schemes and the important nuances in them, starting from which, we will show how different annotation schemes are partially mappable (Sect. 3.1). To deal with the data size issue, we aimed at enlarging as much as possible the minority classes. To this end, Sect. 3.2. will show how we were able to create a new citation context dataset with different citation function annotation schemes, ranging from Teufel et al.'s most cognitively plausible but most challenging 12-class scheme to Jurgens et al.'s most popular and most computationally feasible 6-class scheme. Consecutive

[1] Lauscher et al., (2021) could be said the only exception, where multi-label CDC was the focus, which essentially built, or can be seen as building, one classifier per citation function.

citation strings were merged into citation segments. To answer research questions RQ2 to RQ3, we designed a series of strong deep learning models based on SciBERT (Beltagy et al., 2019) by extensively exploring the options of encoding citation context and/or citance into the citation feature representation (Sect. 4). Experimental results were reported and analysed to answer the questions about the best ways of modelling citations and performing CFC, including whether CFC should be performed at citation level, whether citations should be better modelled in context, and what are the best ways of encoding and combining citation, citance and context (Sect 5). Finally, in Sect. 6 we will make a deeper analysis of the CFC performances. By looking at the per-class performances and further analysing a special class "PSup"/"Support" (knowledge claims support each other or approaches computationally compatible/plug-in-able to each other, according to Teufel et al.'s definition), we will discuss what implications can be derived from both the CFC experiments and the answers for RQ2 and RQ3 for various types of downstream scientometric and bibliometric analysis tasks (RQ4). Finally, preliminary experiments of an ensemble CFC approach will be presented to further improve the overall CFC performance as well as the CFC performances for each class.

2. Related Work

Citation context analysis has a long history dated back to the 1980s. Initially, citation function analysis was mainly focused on the motivations for authors to cite references in scientific writing. Many citation function schemes were proposed by these social science studies. As these studies are not the focus of this paper and the line of research in algorithmic citation function classification, interested readers are referred to a good survey on the tasks about citation context analysis by Hernandez et al. (2016). Lyu et al. (2021) provided a good meta-synthesis approach to understanding the nature and classification of citation motivations. Kunnath et al. (2021) not only presented a comprehensive survey of the annotation schemes for citation function, but also covered the preprocessing steps and feature engineering approaches when applying machine learning algorithms for citation function classification (CFC), as well as a comparative analysis of the existing datasets/benchmarks and the state of the art of machine learning and deep learning methods on this task.

The early attempt to building an automated citation function classifier was rule-based where a so-called *pragmatic grammar* was manually created, which describes a number of handcrafted *cue lists* and certain syntactic constraints and relations on the cue words (Garzone and Mercer, 2000). Similarly, by designing a set of 160 cue phrased-based rules, Nanba et al. (2000) developed a three-type citation classifier for theoretical basis, gap or weakness, and other. In 2006, Teufel et al. (2006a) provided the first comprehensive and, at the same time, operationalizable 12-class annotation scheme and a dataset suitable for machine learning algorithms . The 12-class scheme follows a four-way distinction between the 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 from Teufel et al.'s scheme and more or less mappable to it (Li et al., 2013; Abu-Jbara et al., 2013; Jha et al., 2017; Hernández-Alvarez et al., 2017; Jurgens et al., 2018; Su et al., 2019). See Sect. 3.1-3.2 for more in-depth discussions.

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). Amongst this line, the work by Jochim and Schütze (2012) was special as they defined a four-aspect annotation scheme: including conceptual v.s. operational (akin to "Fundamental_Idea" v.s. "Technical_Basis" in Dong and Schäfer (2011)), organic v.s. perfunctory (equivalent to important v.s. incidental), evolutionary v.s. juxtapositional (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. This study also concluded on the significance of named entity features in CFC. The problem

with these 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, including "Background", "Future", "CompareOrContrast", "Motivation", "Uses", and "Extends". 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 shared tasks (Kunnath et al., 2020, 2021).

Recently deep learning approaches have been introduced to the CFC task. Earlier works applied CNN (Convolutional Neural Network) (Lauscher et al., 2017; Bakhti et al., 2018), BiLSTM (Bidirectional Long-Short Term Memory) (Munkhdalai et al., 2016) or CNN stacked over BiLSTM (Yousif et al., 2018) to encode the citation sentence or context and pool a feature representation of it, which was then fed into a linear or MLP (Multiple-Layer Perceptron) classifier. 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. A recent trend was to incorporate semantically related tasks into modelling the CFC task by use of multi-task learning. These supplementary 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, as will be demonstrated and discussed in more detail in Sect. 5, is that they typically model the whole citation sentence or context, which we believe is flawed. This study explored various ways of modeling citations.

A very closely related task, though being not our focus, is about identifying important or significant citations. Wan & Liu (2014), Zhu et al. (2014), and Valenzuela et al. (2015) were the seminar studies embarking on the topic of citation importance classification, after which a lot of studies were presented in this line (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 a further reduced annotation scheme because citation importance in essence has been defined based on citation function (Lu et al., 2014; Valenzuela et al. 2015). There difference is that CFC is done per each in-text citation, but existing citation importance classification is done 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 suffered the same problem as in CFC (Yousif et al., 2019; Aljohani et al., 2021b; Maheshwari et al., 2021).

3. Dataset and Annotation

3.1. Citation Function Datasets and Annotation Schemes: A Critical Review

In the past two decades, many citation context datasets were proposed. Table 1 gives a comprehensive but by no means exhaustive review of existing datasets. The table contains three parts: 1) general-purpose citation function datasets (the majority part), 2) special-purpose citation function datasets for a subset of citation functions or citation functions on specific scientific entities, and 3) a special type of datasets about citation importance (note that they are not about citation contexts but 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* sentences to the *left* and *r* sentences to the *right* sides of the citance, while a "?" indicates the information is unclear from the paper. The value "variable" means the context can be of a variable length according to user needs, usually thanks to the fact that full-texts of the articles are parsed and annotated (i.e., parse). "OA" stands for "open accessible". The last column "Authoritative" indicates whether the annotations of the involved papers were done by the authors of papers (marked by "•").

Most datasets provided citation contexts of certain lengths. Teufel et al. (2006a, 2006b) and Hernández-Alvarez et al. (2017) annotated all citations in their full contexts. Dong and Schäfer (2011) instead annotated all citations only in their citances

and replaced each citation string with a pair of empty parentheses. 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[2]. It is valuable to provide full texts in the dataset so that users have the flexibility to define citation context according to application needs. As we observed in our own annotation process, sometimes context could be very large, far beyond 2-3 sentences at both sides of the citance. Lauscher et al. (2021) has made a similar claim. A representative case is Teufel's General Rule 27 about *meta-statements* of the "CoCo" (Comparison or Contrast) class (Teufel, 2010). The meta-statement may appear at the beginning of a paragraph to qualify all subsequent citations as "CoCo", but some "CoCo" citations may be far from the meta-statement, as Example 5 in Figure 3 shows. "PMot" often requires larger context too. 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 full text availability to be the first prerequisite for creating a citation context dataset.

Example 5: Meta-statement of comparison and contrast. (Teufel, 2010, pp. 434).

We will outline here the main parallels and differences between our method and previous work. 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. 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. Then the bigram estimate derived by cooccurrence smoothing is given by EQN. In addition, the cooccurrence smoothing method sums over all words in the lexicon. [Miller et al] (CoCoGM) suggest a similar method... They do...

Figure 3. Example of Meta-statement of Comparison and Contrast.

We can observe that very different aspects were annotated for biomedicine (BM) and computational linguistics (CL) or computer science (CS) domains, which are non-ignorable nuances to mapping and merging 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 categories are of specific 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 not specialised in the area of citation context analysis. Therefore, our second prerequisite was that the annotation schemes of source datasets should be at least partially mappable.

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

Table 1. Survey of Existing Citation Function Datasets

Dataset	Fields*	Size	Annotation Scheme	Fulltext	Context	OA	Authoritative
Teufel et al. (2006a, 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

3.2. Partial Mappability between Different Annotation Schemes

In the past two decades, many citation context datasets were proposed for automatic citation function classification (see Table 1 for a critical review). Refer to Kunnath et al. (2021), Lyu et al. (2021), and Hernández-Alvarez and Gómez (2016) for complete surveys of the datasets, tasks and methods. A possible way of creating a large citation context dataset is to merge and re-annotate existing datasets. Although biomedicine (BM) papers are freely available through PubMed Central[3], there are only few datasets. 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. Therefore, most existing datasets were annotated on computational linguistics (CL) papers[4]. There was no publicly available BM dataset, but we were able to obtain six publicly available citation function datasets of CL papers for re-annotation, namely `Teufel2010`[5] (Teufel et al., 2006a; Teufel et al., 2006b; Teufel, 2010), `Dong2011`[6] (Dong & Schäfer, 2011), `Jha2016`[7] (Abu-Jbara et al., 2013; Jha et al., 2016), `Alvarez2017`[8] (Hernández-Alvarez et al., 2017), `Jurgens2018`[9] (Jurgens et al., 2018), and `Su2019`[10] (Su et al., 2019).

RQ1. What are the relationships between different citation function annotation schemes and the mappings between existing citation function datasets? The six datasets' citation function annotation schemes are partially mappable (summarised in Table 2). 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 less notable benefit is the wide time span of the combined dataset, ranging from early 1990s to late 2010s. We believe that the merged dataset can better reflect authors' patterns in placing citations and exhibit richer language expressions around citations.

3.3. Annotation Schemes and Dataset Reannotation

Our dataset, named `Jiang2021`, was created in three steps: dataset preparation, re-annotation and post-processing (Figure 4). Due to space limit, the details of the whole pipeline were moved to the Appendix A. Three postgraduate research students in natural language processing were recruited for re-annotation. The four annotators, including the first author of this paper, re-annotated all 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) plus a "Future" class for future work. The final function for each sample in dispute was adjudicated by consensus among the four annotators. Therefore, no inter-annotator agreement was reported. 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", "."]". As a result, our dataset `Jiang2021` gathered 3356 citation contexts, 4784 in-text citations, and 3854 CITSEGs in total (Table 3). Because "PModi" and "PBas" were still too small, although much bigger than past datasets, we decided to merge them into "Basis" (equivalent to "Extends"). "CoCo-" was split and re-annotated into "CoCoGM" and "CoCoR0" due to its small size. These

[3] https://pubmed.ncbi.nlm.nih.gov/
[4] 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/
[5] https://www.cl.cam.ac.uk/~sht25/CFC.html
[6] https://aclbib.opendfki.de/repos/trunk/citation_classification_dataset/
[7] https://github.com/ivder/University-Project/tree/master/ (The "Citation Sentiment_Purpose Analyser/citation_sentiment_umich/" subfolder)
[8] http://rua.ua.es/dspace/handle/10045/47416
[9] https://github.com/davidjurgens/citation-function
[10] https://github.com/WING-NUS/citation_func_n_prov (We combined the "func" and "prov" portions of this dataset)

treatments resulted in our own 11-class citation function annotation scheme, which was mapped to 9-class, 7-class and 6-class schemes.

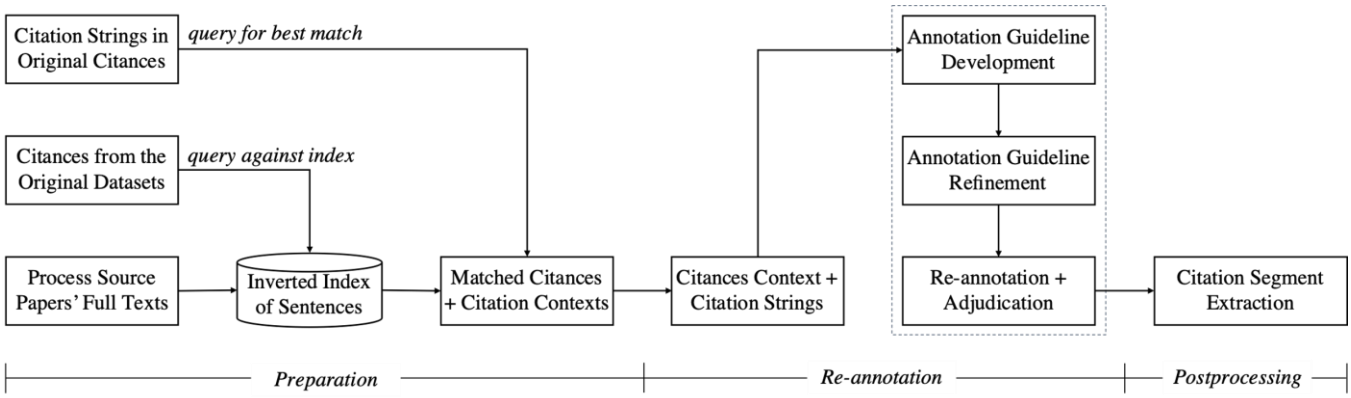


Figure 4. The Dataset Re-annotation Pipeline.

3.4. Discussions

Context matters a lot for annotation. "Weak" instances 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. An issue related to multi-sentence context is that a citation may have multiple functions depending on how we deem it in the context, such as "Neutral" by only looking at the citance and "Weak" by looking at the context. From an application point of view, the stronger class overwrites or subsumes the weaker (Teufel, 2010). Therefore, we defined overwriting rules following Teufel (2010), e.g., "Weak" overwrites "Neutral", and marked each citation with the strongest function based on its context. For another overwriting example, "Motivation" subsumes "Usage" if the *plausible usage* of something (qualifying "Usage") is justified by a *positive statement* (qualifying "Motivation"), because one prerequisite of "Motivation" according to Teufel's guidelines is that the citing study uses something from the cited. Multi-label annotation (Lauscher et al., 2021) is also a reasonable choice in such scenarios. We left it as a potential future direction to explore.

Table 2. 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]	Neu	953	55.15	**Substantiate??** *Unmappable!!*		126	6.01	Background			**ComOrCon**	-	-	**Neut**	993	70.83
Neut Neutral description, or not fit into other classes	2398	59.6	**- GRelated** General **- SRelated** Specifics: method, parameter, … **- MRelated** Methods that may be usable				**Neutral**		1283	61.15	**cor**roborate	0	0	**Future**	69	3.53			
											acknowledge	982	32.59	**Background**	999	51.13			
CoCoXY Contrast between 2 cited methods	125	3.11		Pos *PMot?*	149	8.62					**deb**ate (0)	0	0						
							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) Idea** *+PSim*		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					**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	

Table 3. 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%	Comparison	479	12.43%	ComOrCon	479	12.43%	ComOrCon	944	24.49%
CoCo-	108	80	2.08%												
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	

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 5. 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 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.

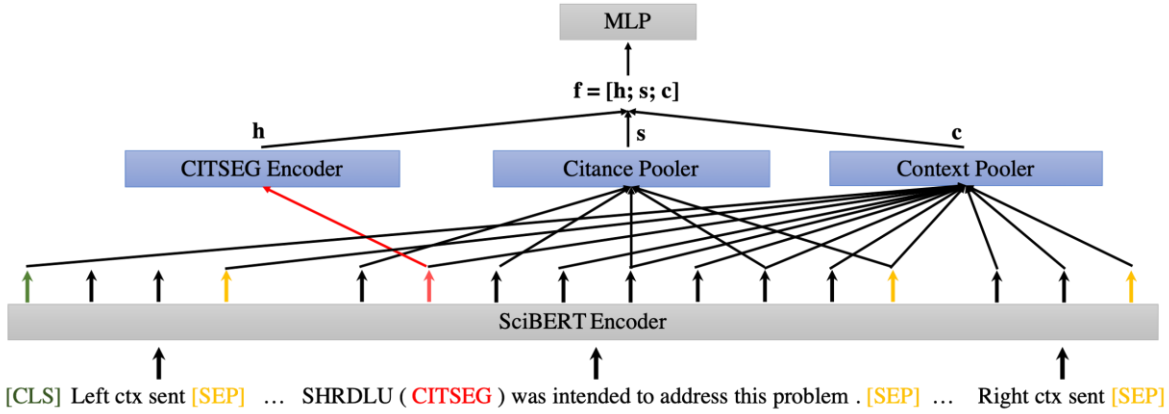


Figure 5. SciBERT-based Citation Semantics Analysis Model (Demonstrated Using a Hierarchical Context).

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 4 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_attend" 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 (discussed in the Introduction section).

5. Results

5.1. Experimental Implementation

The models were implemented using HuggingFace's Transformers library[11] (version 4.2.2). The pretrained SciBERT model was downloaded from the official website[12] 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., 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 feature vector (**f**) size. The AdamW optimizer was used with most parameters set to default. The initial learning rates for the parameters of SciBERT and MLP were initialised to 5e-5 (lr_pret: learning rate for the pretrained model, i.e., the SciBERT part) 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 testing (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 modelling options, the best F1, average F1 and the standard deviation across 5 runs were reported.

5.2. An Additional Dataset and Baseline

For a fair comparison (explained below), we also experimented on an additional dataset extended from `scicite` and compared with the most recent SciBERT-based contextualised CFC approach, which is also the only known SciBERT-based contexualised CFC method to the best of our knowledge. Both the dataset and the baseline methods were proposed by Zhang et al. (2022). The dataset was named `NI-Cite` (Native Information enhanced Citation dataset) by us. The `NI-Cite` dataset was extended from `scicite` by (i) including all instances from the latter and introducing a few thousand more instances from ACL papers, (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[13], the titles, DOIs and Web URLs of the citing paper and cited paper. The original dataset has 11195 citation contexts, each citation context 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 using our own in-house scripts and removed as many errors as possible by both programmatical and manual check. Finally, there were 9645 citation contexts remained in the cleaned `NI-Cite` dataset[14]. Note that this dataset only has three functions: Background, Method and Results. The 3-class

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

annotation scheme is used in Semantic Scholar[15], 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 too simplified and cognitively incomplete, which severely limits its real-world scientometric use.

The two baselines were compared to were both from Zhang et al. (2022). Because 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 unreliable to use such information. In addition, the `Jiang2021` dataset only contains section titles but no manually annotated section 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 4 and the last two columns in Figure 6 in Sect. 5.6). 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 hyperparameter ranges: 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 4 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 CFC experiments 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)[16]. In addition, we also ran five more models which encoded each citation's context or each citation's enclosing citance in its context but uses the context or citance representation alone for classification, 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 with cita-01 is that the former encoded the whole context while the latter only encoded the citance. The difference between seq-x08 (resp. seq-x09) and cita-02 (resp. cita-03) is similar. 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 better be encoded in their contexts, 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-20) in term of best macro F1 were highlighted 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 already promising. We also tested all models on the 9-class, 7-class, and 6-class schemes. An observable consistent trend from all models in Table 4 was 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, an 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 around 70.9% on the 6-class scheme, an approximately 7.40% absolute improvement. These performance results could be deemed rather strong compared to two recent SOTAs: 67.9% by (Cohan et

[15] https://www.semanticscholar.org/

[16] 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 $\mathbf{f} = \mathbf{h}$, or $\mathbf{f} = [\mathbf{h}; \mathbf{s}]$ (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.

al. (2019) and 70.98% by Beltagy et al., (2019). Concerning the latter SciBERT baseline, we will present more experimental results in Sect. 5.6. Note that the 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, not all citations in `Jurgens2018` were annotated and even the citations in the same citance were not all annotated. 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 models seq-x07/08/09 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, however, model seq-x07 simulated Beltagy et al.'s approach, so we are rather confident to conclude the superiority of our contextualised models

5.4. Citation-Level v.s. Citance-Level

In this section, we try to answer *RQ2. Should CFC be performed at (in-text) citation level or at citance or context level: Which choice is empirically supported?* 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 are considered. This is how Beltagy et al. (2019) reported their results on the `scicite` dataset (Cohan et al., 2019). On the contrary, models seq-x07/x08/x09 encoded the surrounding context and used the context representation alone for classification. This is how Beltagy et al. reported their results on the `Jurgens2018` dataset. We can observe a consistent phenomenon across all four annotation schemes that the performances of the citance-level CFC models, i.e., the CITSEG-agnostic models seq-x07/08/09, 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 models seq-x07/08/09 justified our statement that it is conceptually flawed to use the summarised context representation for CFC. The citance-level CFC models cita-01/02/03 got even worse performance than all the citance-level 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, this trend seemed to disappear when comparing the models which used the pooled citance representation to enhance citation representation against the citance-level CITSEG-agnostic counterparts, i.e., models seq-09/10 against models seq-x09/x10. Generally, it looked like that the CITSEG-aware models and CITSEG-agnostic models performed on par. They both could win in some scenarios and the performance figures could be said close. It seemed that in certain cases, citance alone could provide strong enough signals for CFC. This partially explains why Beltagy et al. (2019), the first SciBERT baseline, performed extremely well on the `scicite` dataset, 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; its avg F1 is not very competitive implying 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 4. Citation Function Classification Performances on Different Annotation Schemes

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-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-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-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-05	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-08	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

In this section, we will look at what answers can be derived for *RQ3. Should citation modelling be done in its context and what are the most effective methods for encoding and utilizing the representations of citation sentence and citation context for CFC?* To answer whether citation modelling should be done in its context, we should look at several aspects. Firstly, considering the citance-only CITSEG-agnostic models, i.e., cita-x01 to cita-x03, the highest F1 was only around 59%, which demonstrated the necessity of 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. These performances were much worse than the models which encode citance in its context, i.e., seq-x07 to seq-x11. We also ran the CITSEG-aware counterparts of cita-x01to cita-x03, which encoded the citance and used citance representation to enhance citation representation, i.e., **f** = [**h**; **s**]. Their performances were very close to cita-x01 to cita-x03, thus much worse than their contextualised counterparts seq-07 to seq-12. 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 its context**.

Secondly, there was a confirmative fact that, for all annotation schemes, the best models all encoded the whole context and used context representation and/or citance representation to enhance citation representation, i.e., the representation of "CITSEG". 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 used "[CLS]" as the pooled representation to enhance citation representation. For the 7-class scheme, hie-14 was the best model, which used both citance representation and context representation that were, respectively, pooled from the citance words that were encoded in the whole context and 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 its context**. This conclusion can be further supported by the unexpected strong performances of models seq-x10/x11, which pooled citance representation from the SciBERT-encoded context for citance-level CFC, i.e., without using citation representation, while the context-agnostic counterparts cita-x02/x03 performed very poor.

Concerning the second part of the question, *what are the most effective methods for encoding and utilizing the representations of citation sentence and citation context for CFC*, it is 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 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 is a coincidence or a certain level of regularity. However, in general, it is 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-x10/11.

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 4 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 4. See Table B1 in Appendix B, where we used upward or downward arrows 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 trèfles 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 "[SEP]" could be further pre-trained to be a better sentence encoder, for example, by following the pre-training paradigm used for long document extractive summarisation (Xu et al., 2020).

Concerning citation pooler, still we are unable to answer which is better, "self_attnd" or "max_pool". To better see this, 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 could find is again 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 mix behaviours with hierarchical context. The only regularity occurred when `context` = "X", where "max_pool" outperformed "self_attend" 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. The only two exceptions were seq-12 on the 9-class and 6class schemes. Anyway, even for these two exceptions where only citance representation was used to enhance citation modelling, the citance tokens were contextually encoded too. This re-iterates the importance and usefulness of **encoding citation in its context**.

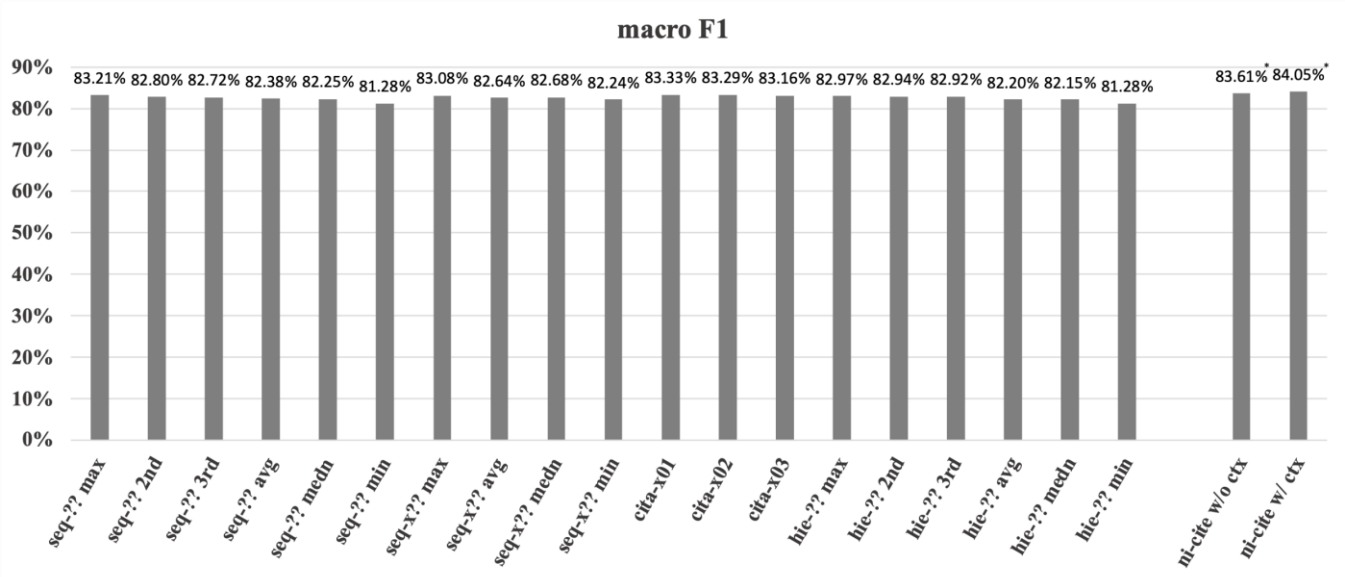
5.6. Additional Experiments on NI-Cite

For the purpose of demonstrating the necessity of modelling citations in their contexts, this section presents our additional experimental results on the `NI-Cite` dataset (Figure 6). We report 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 of all CITSEG-aware models with a hierarchical context (with prefix "hie-??"). We also report the max, mean and median performances of all five CITSEG-agnostic models with a sequential context (with prefix "seq-x??"), the performances of the three CITSEG-agnostic models on single citance (cita-01/02/03). For comparison purpose, the rightmost two columns report 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/ context" column in Figure 6). 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., models seq-x07 to seq-x11. These results implies that the `NI-Cite` dataset (Zhang et al., 2022), as well as the `scicite` dataset (Cohan et al., 2019) it extends on, is "problematic" in the sense that the citations actually could be recognised using the citance alone. Indeed, the `scicite` dataset, as we discussed in Sect. 3.1, does not contain citation context information. We argue 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 the citance-level model slightly outperformed the citation-level models on `NI-Cite`, although the latter, i.e., the citation-level models, proved to be much stronger on the `Jiang2021` dataset (described in the next paragraph). Note that, our contextualised CFC models were not hyperparameter-tuned; due to high computational overloads, the same learning parameters as in the experiments on `Jiang2021` were used for all model variants, 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 the Zhang et al.'s method 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 citation context often may be quite large. Again, we can claim the **importance that citations should be encoded in its context** (answer for *RQ3*), and that **citation function classification should be done at citation level rather than per citance** (answer for *RQ2*). 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. A promising ideal is to determine a "dynamic" context, i.e., the minimal context around a citance which can provide 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 6. CFC Performances on the NI-Cite Dataset.

6. Analysis and Discussions

This section will present a more in-depth analysis of the CFC results, with our thoughts focused on the last research question: *RQ4. How well can a general-purpose citation function classification model suit different types of scientometric analysis tasks and what implications can we derive for the real-world application of citation function classification?*

6.1. Per-Class Performance Analysis

Table 5-8 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"/"Background" and "Usage"/"Uses".

Teufel et al. (2006b) and Cohan et al. (2019) made similar observations. For example, the best models for neutral citations achieved 78.85% F1 with the 11-class scheme (model hie-08 in Table 5), also the highest per-class performance 78.85%; achieved 79.66% F1 with the 9-class scheme (model hie-14 in Table 6); achieved 82.64% F1 with the 7-class scheme (model seq-06 in Table 7), already close to the highest per-class performance 83.45%; and achieved 79.88% F1 with the 6-class scheme. "Future" was a small and easy class. Indeed, 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).

There were a few difficult classes too. "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" ($\kappa = 0.41$) and "PModi" ($\kappa = 0.55$)[17]. Concerning algorithmic classification, good models could achieve 66.67% F1 on the 11-class scheme (model seq-08 in Table 5). Note that the best performance 69.70% F1 was obtained by a citation-agnostic model seq-x10. This is probably because usually there is only one approach, i.e., one CITSEG, in a citation sentence that is "based on" by the citing paper. On the 7-class scheme, we are excited to see an obvious performance improvement to 70% F1 (model seq-01 in Table 7, "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 8). 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, which is an important first step for **semantic analysis of scientific knowledge flows** (Jiang et al., 2022; Ghosal et al., 2022).

Comparison or contrast functions often recorded good performances. With the 11-class scheme (Table 5), 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 6) and 77.84% (model hie-21, the "best of all models" column in Table 7) respectively. The 7-class scheme merged "Weakness" into "Background". We can see that the "Background" performance on the 7-class scheme was greatly improved to 83.45%. Empirically it seems better to merge "Weakness" into "Background"; meanwhile we also feel it cognitively more plausible to do so compared to merging it into "ComOrCon". On the contrary, the 6-class scheme (Jurgens et al., 2018) merged both "Weakness" and "Similar" (including "Support") into the comparison classes. However, the performance for "ComOrCon" was not worsened on the 6-class scheme. This is promising if our analysis does not distinguish similarity versus comparison, contrast, and difference between two studies. Cognitively, both "Similar" ("Support") and "ComOrCon" imply high topical or technical relatedness between studies, thus these results are very useful for building **academic recommendation systems** for many downstream applications, such as identifying related studies for 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 7 shows that simply absorbing "Support" into the "Similar" class made it more difficult to correctly recognise the similarity class. According to Teufel's annotation guidelines, there are two distinct

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

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. Sect. 6.3 will present the 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. Example 3 in Figure 2 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" (="Neutral") should be applied if no meta-statement of comparison exists. Although in certain cases we can infer the comparisons from juxtaposed citations, the annotation guideline says that only explicit comparison expressions qualify "CoCoXY". Example 4 in Figure 2 illustrates the latter case.

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 function categories or on all annotation schemes. This phenomenon is especially obvious on the 6-class scheme (Table 8), 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 for *RQ4* that **no single CFC model is robust enough for performing scientometric 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, e.g., Lu et al. (2014) and Zhang et al. (2021), which considers not only why something is cited but also what specific scientific entity is cited. To analyse **scientific research lineage** or **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., 2015; Valenzuella et al., 2015), there is large room of improvement and demand of further research. In Sect. 7, 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 even 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**. We leave this line of research for future work. Recall that the best performances reported on the `Ni-Cite` dataset were about 83-84%, on par with the best "Background" model we obtained. In the 3-class annotation scheme, only the "Usage" class corresponds to significant citations. 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 et al. (2022) presented the initial attempt of this idea.

Table 5. 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 6. 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 7. 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 8. 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**

6.3. 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 9). Table 10 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 11).

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 9. 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 10. 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 11. 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**

Table 12. 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.

7. Ensemble

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 beame 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 beared 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 7 illustrates the idea of the *naïve ensemble* classifier. Refer to Zhou (2014) for more details of ensemble learning.

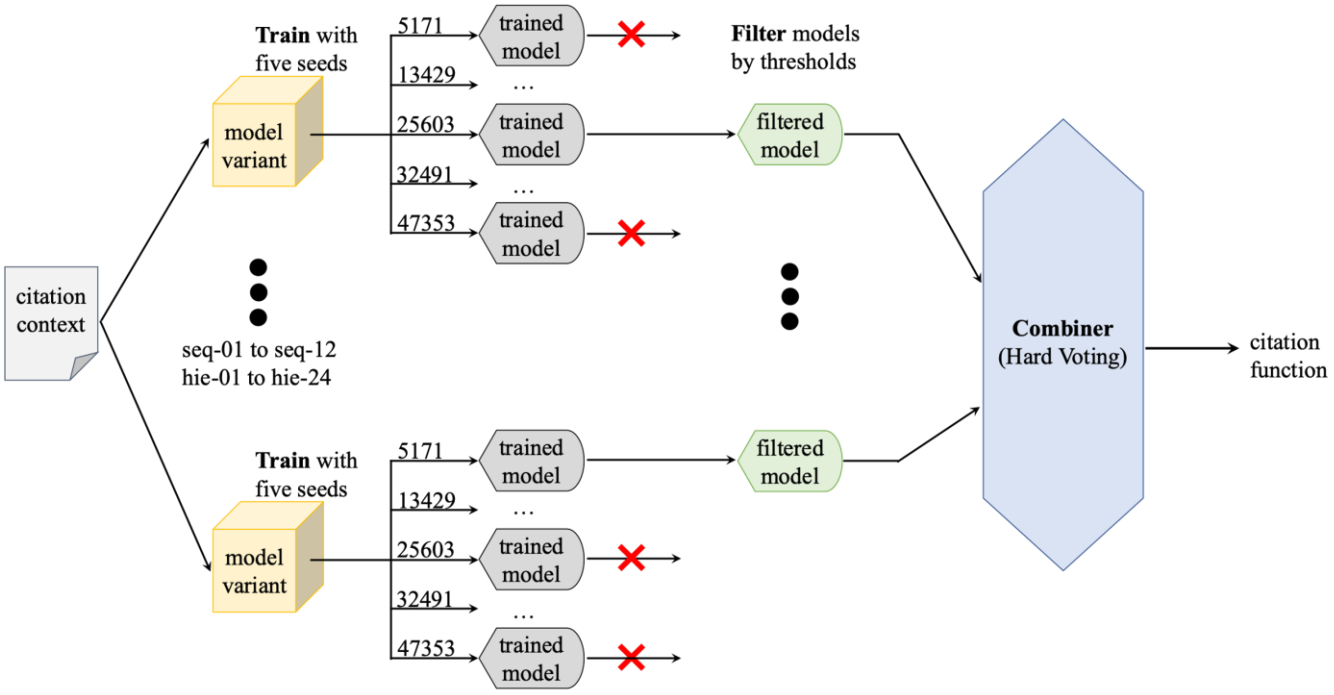


Figure 7. Naïve Ensemble Model for Citation Function Classification

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.

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 12 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 scientometric analysis tasks that were discussed in Sect. 6.2. This would be one very important and interesting future direction.

8. Conclusions Remarks

This paper studied contextualised segment-wise citation function classification and analysed the implications of the results for downstream scientometric applications. Several contributions were made. The first contribution was a new citation context dataset that was created by merging and re-annotating six existing datasets in the computational linguistics domain. The first

research question around dataset and annotation was the relationships and mappings between the different annotation schemes of existing datasets. A comprehensive critical review revealed that re-annotation is possible because their annotation schemes are conceptually related and, at least, partially mappable. What is more, samples of different datasets complement each other. Four annotators collaborated in re-annotation using Teufel et al.'s12-class scheme plus an additional function about future work. Conflict annotations were adjudicated by consensus of the four annotators after discussion. In total, 3356 citation contexts, 4784 in-text citations and 3854 citation segments (consecutive block of in-text citation strings) were annotated.

Secondly, effective citation function classification models were studied at citation (segment) level based on SciBERT – the pretrained scientific language model. Our research questions mainly centered around the necessity of contextualised in-text citation modelling and the effectiveness of different feature representations of in-text citation, citation sentence and citation context. Empirically, we were able to 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 most of the time led to very poor performance and in many cases using citation representation alone produced surprisingly competitive results. Experiments also justified our claim that citations should better be encoded in its context. This also means that often the best performances were only achievable by appropriately modelling citation sentence and/or context into citation representation; different combinations of the representations of in-text citation, citation sentence and citation context worked well on different datasets, with different annotation schemes, and for different citation functions.

The citation function classification models developed in this study produced competitive classification performances such that they are promising to be applied to certain scientometric applications. The macro F1 of the best model was improved from 66.16% on the 11-class scheme to 74.03% on 6-class scheme. An observable trend was that a more concise annotation scheme would result in better overall classification performance. However, this does not mean performance boosts to all classes. An in-depth per-class performance analysis revealed that a general-purpose citation function classification model can NOT suit all kinds of scientometrics analysis tasks. Unfortunately, there was NO single citation function classifier that worked well for all citation functions. As for the real-world scientometric and bibliometric analysis based on citation contexts, we need to either depend on general-purpose models that work well for a specific function or develop bespoke models tailored to the specific function. An encouraging by-product of the versatility of well-performing models was that we were able to build a naïve ensemble citation function classifier to not only improve the overall performance but also significantly improve all difficult classes' recognition rates. We believe that there is much room to explore in this direction and citation function classifiers of more robust performances were anticipatable if more advanced deep learning and machine learning approaches are introduced.

Concerning per-class classification performances, large functions like "Neutral"/"Background" and "Usage"/"Uses" got the best and most stable results. "Future" was the easiest function, reaching 100% precision and higher than 90% F1. "Weakness" was a more difficult class because the weak point of something is often pointed out after a neutral description in the citation context. Merging "Weakness" into "Background" greatly improved the performance. Two citation functions were difficult to recognise: (i) functions whose language expressions overlap other categories like "Similar" v.s. "Usage", or (ii) functions whose definitions embrace two or more distinct meanings such as "Support". Especially, we argue that "Support", in the sense of its "mutual compatibility" meaning, is better re-annotated if it is not the main purpose of scientometic analysis. However, if it is indeed the focus of study, we should develop a bespoke binary citation function classifier for it. We desperately need further work on the annotation and recognition of the supporting and conflicting relationships between scientific studies, i.e., "Support" v.s. "Anti-Support", in the sense of "mutual agreement" v.s. "disagreement".

In summary, we were able to conclude that, although not perfect for all citation functions, existing citation function classification models allow for application to a wide range of scientometric anlaysis tasks. For example, the best models were extremely strong in screening out unimportant, insignificant, 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. Comparison and contrast functions were able

to be rather correctly recognised, which makes it promising to be applied to recommending related studies to facilitate many useful applications such as peer review and systematic review etc. The current citation function classification models also obtained decent performances for the relationships about "being technologically, theoretically or conceptually based on or motivated by". 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 is the analysis of the pattern of scientific entity usage, including dataset, software, algorithm, method and so on. Existing models are already strong enough for such applications. We believe and hope that the methods, models, analysis, conclusions, and implications made in this study will be helpful to scientometricians and bibliometricians in their analysis based on citation context analysis.

9. ACKNOWLEDGMENTS

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. Xiaorui Jiang is partially supported by National Office for Philosophy and Social Sciences of China (18ZDA238).

10. 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

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

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

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. (2022). 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

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

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., 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

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

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). 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., Siddharthan, A., & Tidhar, D. (2006b). 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. (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

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

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 2, the dataset creation pipeline included three steps: dataset preparation, re-annotation and post-processing.

A.1. Preparation

The whole ACL Anthology was crawled. Full texts and citation contexts of each paper were extracted using Allen AI's s2orc-doc2json tool[18], which postprocessed and transformed the output of Grobid[19] 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[20]. 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 four annotators, including the author of this paper, re-annotated all 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. 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", "PModi", "PBas" and "PMot" instances from `Teufel2010`. We started from Teufel's description of these functions, reached consensus among all four annotators, and draft our annotation guidelines for these functions. Our own guidelines were based on and adapted from Teufel's annotation guidelines. We found our re-annotations were highly consistent with `Teufel2010`'s original annotations. Similar observations occurred in re-annotating instances of "Weak" and all "CoCo" classes from `Teufel2010`. This gave us confidence in the overall quality of our guidelines and re-annotations.

Stage 2 (guideline refinement). The samples of conceptually related functions from other datasets were re-annotated, including "Substantiating" instances from `Jha2016`, and "Background" instances with Positive and Negative sentiment from `Dong2011` (suspect for "PMot" and "Weak" respectively) etc. The "CoCo" and "PSim" instances from `Teufel2010` were also re-annotated, so were the comparison functions from other five datasets. 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"

[18] https://github.com/allenai/s2orc-doc2json
[19] https://github.com/kermitt2/grobid
[20] https://lucene.apache.org/core/

v.s. “PSim” when we see expressions like “similar to” or “in the same way as”. The annotation guidelines for “PSup”, “PSim”, “PUse”, “PModi”, “PMot” and “PBas” were refined at this stage. Finally, we also re-annotated citation instances about weakness (“Weak” from `Teufel2010` and `Su2019`) and future work (“Future” from `Jurgens2018`).

Stage 3 (re-annotation & adjudication): The three co-authors re-annotated the remaining non-Neutral samples. This included “Fundamental” and “Background+Neg” from `Dong2011`, “Criticising”, “Uses” and “Basis” from `Jha2016`, “use”, “bas”, “wea” and “hed” from `Alvarez2017`, “Uses” and “Extends” from `Jurgens2018`, 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. This also implied the incomparability of the results in different CFC papers.

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.

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	↓