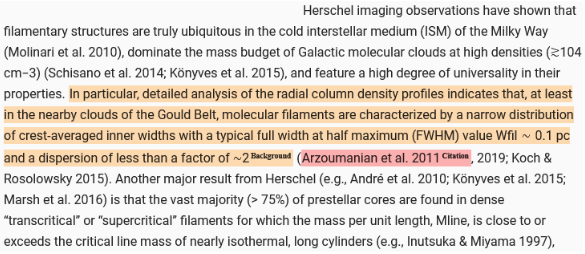
Function of Citation in Astrophysics Literature (FOCAL): Findings of the Shared Task

Felix Grezes[1], **Thomas Allen**[1], **Tirthankar Ghosal**[2], **Sergi Blanco-Cuaresma**[1 3]
[1]Center for Astrophysics, Harvard & Smithsonian, USA
[2]Oak Ridge National Laboratory, USA
[3]Laboratoire de Recherche en Neuroimagerie, University Hospital (CHUV)
and University of Lausanne (UNIL), Lausanne, Switzerland
[1](felix.grezes, thomas.allen, sblancocuaresma)@cfa.harvard.edu
[2]ghosalt@ornl.gov

Abstract

In this article, we describe the overview of our shared task: Function of Citation in Astrophysics Literature (FOCAL). The FOCAL shared task was part of the Workshop on Information Extraction from Scientific Publications (WIESP)[1] in IJCNLP-AACL 2023. Information extraction from scientific publications is critical in several downstream tasks such as identification of critical entities, article summarization, citation classification, etc. In particular, the citation graph is an essential tool for helping researchers find relevant literature. To further empower discovery, the motivation of this shared task was to develop a community-wide effort to label the edges of the graph with the function of the citation: e.g. is the cited work necessary background knowledge, or is it used as a comparison, to the citing work? We propose a shared task of automatically labeling citations with a function based on the textual context of the citation, and analyze the systems, performances, and findings of FOCAL participants.



Herschel imaging observations have shown that filamentary structures are truly ubiquitous in the cold interstellar medium (ISM) of the Milky Way (Molinari et al. 2010), dominate the mass budget of Galactic molecular clouds at high densities (≳104 cm−3) (Schisano et al. 2014; Könyves et al. 2015), and feature a high degree of universality in their properties. In particular, detailed analysis of the radial column density profiles indicates that, at least in the nearby clouds of the Gould Belt, molecular filaments are characterized by a narrow distribution of crest-averaged inner widths with a typical full width at half maximum (FWHM) value Wfil ∼ 0.1 pc and a dispersion of less than a factor of ∼2$^{\text{Background}}$ (Arzoumanian et al. 2011$^{\text{Citation}}$, 2019; Koch & Rosolowsky 2015). Another major result from Herschel (e.g., André et al. 2010; Könyves et al. 2015; Marsh et al. 2016) is that the vast majority (> 75%) of prestellar cores are found in dense “transcritical” or “supercritical” filaments for which the mass per unit length, Mline, is close to or exceeds the critical line mass of nearly isothermal, long cylinders (e.g., Inutsuka & Miyama 1997),

Figure 1: Sample annotation. The citation *Arzoumanian et al. 2011* is used as *Background* by the authors of this paragraph.

1 Introduction

In addition to its archival mission, the NASA Astrophysics Data System (Kurtz et al., 2000) aims to empower astrophysics researchers in their work. One powerful tool at their disposal is access to the citation graph, allowing them to find papers related to, and quantify the impact of, their research. By enriching the edges of the citation graph with labels that explain why a citation was made, and the relevant textual context to understand the citation, researchers can more rapidly assess the literature, and gain more granularity into impact metrics. For example a researcher who is already familiar with the *Background* of a topic may primarily be interested in citations that *Compare / Contrast* with other works. Further, by augmenting impact metrics, such as citation counts, with metrics pertaining to citation function, researchers can gain finer grained insight into the impact of their work, e.g. if they provide the *Motivation* for the citing work or the *Background*. Large scale labeling of the citation graph requires automated methods. In our FOCAL@WIESP2023 shared task, we instigate a community initiative to design such methods.

2 Task

2.1 Definition

The shared task *Function of Citation in Astrophysics Literature (FOCAL)* (Grezes et al., 2023) consists of automatically labeling citations with a function based on the textual context of the citation.

More precisely, given a paragraph of text from the astrophysics literature, and the `start` and `end` position of a citation in the paragraph, the FOCAL participants are tasked with building a model that outputs why it was cited (the function) and the associated span of text in the paragraph (the context). Figure 1 shows a sample annotation.

2.2 Evaluation

For evaluation, submissions were first tokenized into words using the default spaCy tokenizer (Honnibal et al., 2020); references and predictions were converted into IOB2 style labels; and finally scored

[1] https://ui.adsabs.harvard.edu/WIESP/

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 143–147
Nov 1, 2023. ©2023 Association for Computational Linguistics

by three metrics derived from the CoNLL-2000 shared task seqeval (Nakayama, 2018):

- Full Seqeval: the full seqeval score and main evaluation metric. This metrics check that the functions of the citation were placed correctly in the paragraph along with the correct function labels.

- Generic Label Seqeval: a seqeval score with a generic label instead of functions. This metric checks that the parts of the paragraph that explain the functions of the citation were correctly found, without checking if the reason(s) a given citation was made (the function labels) were correctly predicted.

- Labels Only F1: an F1-score on the function labels only. This metric checks that the reason(s) a given citation was made were correctly predicted, without checking if the parts of the paragraph that explain the function of the citation were correctly found.

All reported scores use micro-averaging.

3 Dataset Description

3.1 Data Collection and Creation

The dataset consists of paragraph sized text fragments that were curated from over 25,000 astronomy articles, from the Astrophysical literature. The journals that the text fragments were obtained from are the Astrophysical Journal, Astronomy & Astrophysics, and the Monthly Notices of the Royal Astronomical Society. All text fragments are from recent publications, between the years of 2015 and 2023. From this set of articles, over 2 million citations and their context were harvested. Further, only citations with context sizes between 2,000 and 10,000 characters are selected. A domain area expert manually examined these text fragments to determine the citation function as well as label the relevant context.

We are considering a set of eight potential citation functions. These are:

- Background: The cited work provides background information needed to understand the citing work

- Motivation: The cited work is motivating the citing work

Split Function	Train	Val	Test
Background	1607	390	438
Uses	877	230	530
Compare/Contrast	615	178	140
Similarities	279	50	72
Motivation	233	70	56
Differences	125	24	40
Future Work	40	9	4
Extends	9	5	2
Totals	3785	956	1282

Table 1: Counts of function labels in the dataset. Note that totals are larger than dataset sizes because some samples have multiple function labels associated.

- Uses: The citing work used a result from the cited work

- Extends: The citing work extends a result from the cited work.

- Similarities: Results from the cited work are similar to results from the citing (or another) work.

- Differences: Results from the cited work are different to results from the citing (or another) work.

- Compare/Contrast: Results are being compared in a neutral manner between the cited and the citing (or another) work.

- Future Work: Citing work contains implications for future research that are beyond the scope of the citing work.

These citation functions were selected because of their similarity to the classification scheme used in Pride and Knoth (2020), see table 3 in the appendix for a full description with examples.

3.2 Data Segmentation for Shared Task

The FOCAL dataset consists of 3 components, the training dataset consisting of 2421 samples, the validation dataset consisting of 606 samples, and the testing dataset consisting of 821 samples. Table 1 shows the counting statistics of the function labels for each component.

4 Participant Systems

Ikoma and Matsubara (2023) proposed a SciBERT-based sequence labelling system that outputs IOB2

Model	Baseline		(Ikoma and Matsubara, 2023)		Veeramani et al. (2023)	
Split Metric	val	test	val	test	val	test
Full Seqeval	23.68	20.94	54.08	**51.87**	23.79	30.17
Generic Label Seqeval	59.86	54.55	79.92	**73.00**	43.96	47.65
Labels Only F1	42.87	35.99	65.94	**69.44**	42.61	57.51

Table 2: Main FOCAL@WIESP 2023 shared task results. All scores computed using micro-averaging.

tags, and uses statistical insights on which sentences (preceding, citing, following) contain function labels to limit the range of the input text to what the language model can handle. The authors explore the performance of multiple BERT-based models.

Veeramani et al. (2023) proposed a system that leverages state-of-the-art BERT-based language models and combines paraphrasing and question-answering techniques. Paraphrasing is used in the pipeline to reduce the text input length to 512 tokens, allowing for sequence classification models to be applied, which provide the function label of the citation. To find the boundaries of the function, the authors apply BERT-based Question Answering techniques.

In addition to the above papers, two submissions were made to the Codalab platform hosting the shared task[2].

4.1 Baseline

Baseline scores from a simple model are provided as benchmark for the participants. This model is defined as follows:

- the function of the citation is the majority class (i.e. Background).

- the start and end of the function is the sentence that includes citation, as defined by pySBD (Sadvilkar and Neumann, 2020).

5 Results, Analysis, and Findings of FOCAL

We report the results of the participating teams in table 2. Both systems were able to outperform the baseline on the Full Seqeval and Labels-Only metric, but only Ikoma and Matsubara (2023) were able to improve on the Generic Label Seqeval. Upon further analysis, this is likely due to the method used by Veeramani et al. (2023) to label functions, which does not incorporate information specific to the citation given, versus any other that may appear in the paragraph. Indeed this difficulty is central to the task. Models cannot solely rely on the text of the paragraph to make function label predictions, since those will differ from citation to citation present in the text.

Both submissions make extensive use of BERT-based models, highlighting just how generically useful and practical those models have become, even as state-of-the-art architectures have grown much larger (ex: BLOOM, LLAMA2, etc ...).

6 Conclusion and Future Directions

The results of the FOCAL@WIESP2023 shared task show that the task of labelling the citation graph and locating the text relevant to the citation is far from solved. One aspect that future challenges can improve upon is the quantity of labeled data along with inter-annotator agreement statistics, to confirm that the task is sound and well understood. The advent of open-source Large Language Models also may be used as zero-shot systems that can form a more robust and challenging baseline.

References

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of citation in astrophysics literature (focal): Findings of the shared task. In *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Tomoki Ikoma and Shigeki Matsubara. 2023. On the use of language models for function identification of citations in scholarly papers. In *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and

[2] https://codalab.lisn.upsaclay.fr/competitions/15292

Joyce M. Watson. 2000. The NASA Astrophysics Data System: Overview. , 143:41–59.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL 2020)*.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. Automated citation function classification and context extraction in astrophysics: Leveraging paraphrasing and question answering. In *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

A Appendix

Label	Definition	Example
Background	Citation whose purpose is to provide background information so that the reader can understand the problem, or the object.	The AGN in these systems have been shown to deposit vast amounts of energy into the surrounding intracluster medium via heating and (mega-parsec scale) jets both observationally and by means of modelling (e.g. Binney 2004...
Motivation	Citation that is used to justify the current work or problem.	Unlike NGC 3894, for which no observations with Cherenkov telescopes have been performed, M 87 and 3C 84 are also detected at very high energy (VHE, E >100 GeV; Aharonian et al. 2006
Uses	Result or idea from cited work is used in the current work. Could be in the form of using data or an idea to build an argument.	Our data set consists of 4348 hr of data in both the nominal LPF configuration and the "Disturbance Reduction System" (DRS) configuration, in which a NASA-supplied controller and thruster system took over control of the spacecraft (Anderson et al. 2018).
Extends	Citing work is extending the results of the cited work.	In doing so we extend the analysis of Planck Collaboration Int. XXXVIII (2016) and Planck Collaboration Int. XLIV (2016) to sky areas in which the filaments have very little contrast with respect to the diffuse background emission.
Similarities	There are similarities, in results or observations, between the cited and citing works,	All of these galaxies are consistent with the relationship between X-ray luminosity and mid-IR luminosity for starburst galaxies (. . . ; Sell et al. 2014).
Differences	There are differences, in results or observations, between the cited and citing works,	We also remark that the expression from Mishima et al. (1983) would give a penetration depth of 56 m at 2.2 cm, which is an order of magnitude larger than indicated by the laboratory measurements of Paillou et al. (2008)
Compare/Contrast	A neutral comparison between works or ideas	At these early epochs, this difference could be caused by the poor constraints on the GSMFs adopted by Duncan et al. (2019) which result in large uncertainties on their data, making it impossible to draw robust conclusions at z $\sim$5.
Future Work	Used when cited work provides a means to expand the scope of the citing work	The study presented here will also be further extended to explore the effects of different retention fractions of dark remnants (neutron stars and black holes; see, e.g., Giersz et al. 2019

Table 3: Definitions of the FOCAL labels.