GraphCite: Citation Intent Classification in Scientific Publications via Graph Embeddings

Dan Berrebbi*
Nicolas Huynh*
dan.berrebbi@polytechnique.edu
nicolas.huynh@polytechnique.edu
Ecole Polytechnique, Institut Polytechnique de Paris
Palaiseau, France

Oana Balalau
oana.balalau@inria.fr
Inria, Institut Polytechnique de Paris
Palaiseau, France

ABSTRACT

Citations are crucial in scientific works as they help position a new publication. Each citation carries a particular intent, for example, to highlight the importance of a problem or to compare against results provided by another method. The authors' intent when making a new citation has been studied to understand the evolution of a field over time or to make recommendations for further citations. In this work, we address the task of citation intent prediction from a new perspective. In addition to textual clues present in the citation phrase, we also consider the citation graph, leveraging high-level information of citation patterns. In this novel setting, we perform a thorough experimental evaluation of graph-based models for intent prediction. We show that our model, GraphCite, improves significantly upon models that take into consideration only the citation phrase. Our code is available online[1].

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Natural language processing**.

KEYWORDS

citation intent classification, graph neural network

1 INTRODUCTION

Scientific articles are pivotal in sharing knowledge within a scientific community, but also with nonexperts. A 2019 report of the National Science Foundation found that the number of publications in science and engineering had grown constantly over the years, from 1.8 million articles in 2008 to 2.6 million articles in 2018.[2]

---

*Both authors contributed equally to this research.
[1]https://github.com/nyxpho/graphcite
[2]https://ncses.nsf.gov/pubs/nsb20206/

Given this profusion of information, citations play a crucial role in understanding the lineage and evolution of a field. Authors cite other works with different intentions, such as highlighting the importance of a problem or comparing against results provided by another method. Citation intent has been used to show the evolution of a scientific field [12], or to facilitate the recommendation of further citations [13]. Interestingly, in [12] the authors show that citations can be correlated to a shift to rapid discovery science [5], that is a period in which a field reaches a consensus on the research topics, methods and technologies and strives to improve upon them.
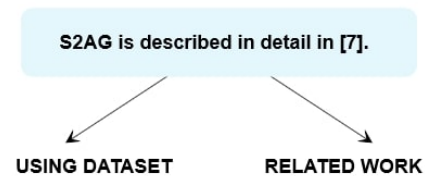

Figure 1: The intent is not clear from the citing phrase, however the majority of papers cite S2AG because they use it as a dataset, a pattern that can be observed in a citation graph.

Different intent taxonomies and several methods for automatic classification of intent have been proposed over the years. In [12], the authors propose six intent classes and use a large number of features together with a random forest classifier for prediction. The features include lexical, grammatical, and morphological information, field-related metrics such as the citation count and the Pagerank of the reference paper, patterns in the citation phrase, and topic model features. Cohan et al. [4] propose three intent classes and a neural network trained on multi-task classification. Citation intent is learned jointly by predicting the section title in which the citation occurs and predicting whether a sentence needs a citation. State of the art results on the datasets proposed in [12] and [4] are obtained by SciBert [3], a pretrained language model based on Bert [7] and trained on a large corpus of scientific text.

In this work, in addition to textual clues present in the citation phrase, we also leverage semantic information present in a citation graph. Differently from [12], we are not retrieving the importance of a particular article via its degree or Pagerank, but its representation based on the citation graph. The neighborhood of an article in a citation graph should bring more context to the classification task and help for ambiguous phrases, such as shown in Figure 1. For

example, when considering other neighboring papers that cited the referenced paper, the most frequent intent used could be a good indication of the intent in a new citation. When publishing in the same area, an author will frequently reuse references with the same citation intent. Moreover, we can incorporate textual information such as titles or abstracts, which can help predict citation intents. For example, in [4], the textual information surrounding the citation phrase was used, but the authors claimed that it led to a degradation of the algorithm's performance. We were motivated to look at the macro-level, i.e., using a citation graph. The citation graph has been used with success for a personalized recommendation for researchers [1] or context-aware citation recommendation [11], but not for citation intent prediction, to the best of our knowledge.

Our salient contributions are: *i*) we investigate the creation of citation graphs for existing intent datasets; *ii*) we propose a new intent classification model that takes into account both the citation graph and the citation phrase; *iii*) in a thorough experimental evaluation, we show that using citation graphs as additional context for citation prediction can significantly improve upon state-of-the-art results.

2 PROBLEM STATEMENT

Given a phrase of the form "S2AG is described in detail in [7]" contained in the paper Anonymous et al. (2022), we denote as the **citing paper** *Anonymous et al. (2022)*, **cited paper** *[7]*, and **citation phrase** *S2AG is described in detail in [7]*. A **citation intent dataset** contains pairs of *(citation phrase, citation intent)*.

Citation graph. We define a **citation graph** as a graph $G = (V, E, \theta)$ with **auxiliary information on nodes**, given by the function $\theta : V \mapsto R^d$, where d is an input dimension. Nodes can be papers, authors, or conferences, and auxiliary information for nodes can be the paper title or abstract, an author's short bio, and conference tracks. Edges are between papers when there is a citation, between papers and authors for authorship relations, and between papers and conferences for publications. The auxiliary information is transformed from text to a dense vector via an embedding technique. When no auxiliary information is present, the dense vectors are initialized randomly.

Intent classification with auxiliary information. Given a dataset of citation intents, we construct a citation graph to facilitate the task of citation intent prediction. The citation graph initially contains only the citing and cited papers. However, one can add additional nodes and edges by considering inbound and outbound citations of the original papers, the authors, and the conferences.

Problem statement Given a citation graph $G = (V, E, \theta)$ and a citation intent dataset, predict the intent of unseen citation phrases.

3 DATASETS

Existing citation intent datasets [4, 12] consist of information regarding the citation phrase, its intent, and metadata about the cited and citing papers. However, often this metadata is incomplete. Below we describe the datasets and how we construct citation graphs.
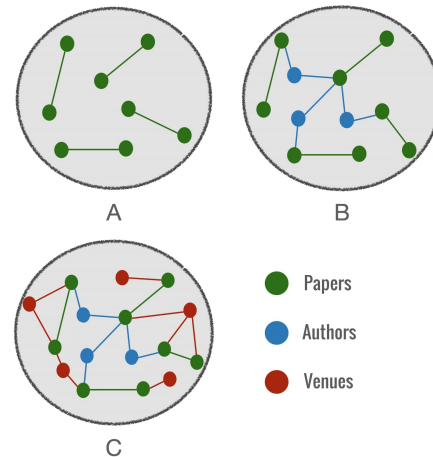

Figure 2: Creating a citation graph.

S2AG. Ammar et al. [2] introduced the Semantic Scholar Academic Graph (S2AG), a dataset accessible via an API[3], that provides metadata on research papers published in all fields. The metadata information includes the paper title, abstract, authors, citations, publication venues. We used this dataset to construct citation graphs when context information was missing in the citation datasets.

ACL-ARC. Jurgens et al. [12] created a dataset of $2K$ citations annotated for their intent for 186 papers in the field of natural language processing. The authors considered 6 intents: background (51% of citations), uses (19%), compare/contrast (18%), motivation (5%), extends (4%), future work (4%). Each entry in the dataset contains several fields, however, we only keep the citation intent, citing phrase, citing and cited paper title and ID. We tried to use the paper's ID to obtain more information on the papers, such as their authors and venues. However, 65% of citations have at least one paper from outside the ACL Anthology Corpus[4], indicated by a prefix "External" in the ID, and we cannot match them to any existing source. We retrieved their authors and publication venue from the S2AG dataset for the rest of the papers.

SciCite. Cohan et al. [4] created a dataset of $11K$ manually annotated citation intents for papers in computer science and medicine, covering $6.6K$ papers. The authors consider 3 citation intents: background information (58% of citations), method (29%), and result comparison (13%). Each entry in the dataset contains the citation intent, citing phrase, name of the section where the citation occurs, citing and cited paper ID corresponding to Semantic Scholar IDs. We use S2AG to retrieve the title, authors and venues.

Citation graphs. We created several graphs by leveraging citations in the datasets of citation intents. The citations labeled with intents are used to create an initial graph, with nodes representing citing or cited papers. To capture more context about the papers, we add their authors and the venues of publication as nodes, as shown in Figure 2. This additional information also creates better-connected graphs, where we should observe citation patterns. We give the size of these graphs in Table 1.

---

[3]https://api.semanticscholar.org/graph/v1
[4]https://www.aclweb.org/anthology/

Graph	Nodes	Edges
SciCite (citations)	13K	11K
SciCite (citations, authors)	49.4K	49K
SciCite (citations, authors, venues)	51K	55K
ACL-ARC (citations)	1.3K	1.9K
ACL-ARC (citations, authors)	2.3K	3.4K
ACL-ARC (citations, authors, venues)	2.4K	4K

Table 1: Citation graphs

4 CITATION INTENT PREDICTION

4.1 Graph-based Intent Classifier

The Graph Attention Network (GAT)[16] is an inductive method
for computing node embeddings in graphs with node attributes. An
inductive approach will be able to provide embeddings for dynamic
graphs and hence it is appropriate for citations graphs that are up-
dated to reflect new publications. GAT computes node embeddings
by aggregating information from their neighbourhood but differs
from earlier techniques [10], as it uses masked self-attentional lay-
ers to specify different weights for nodes in a neighborhood.
SciBert[3] is a pretrained language model based on Bert [7],
that is trained on a large corpus of scientific text, from a variety
of scientific domains. SciBert has shown improved performance
compared to Bert on many NLP tasks. SciBert is generally fine-
tuned on the downstream task to achieve optimal performance. In
[3], the authors show that on the citation intent prediction task the
fine-tuned architecture had a better performance than the frozen
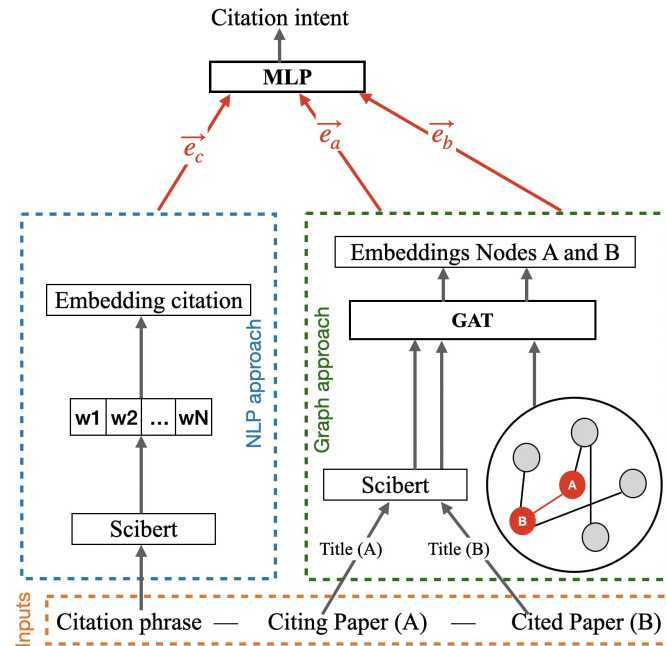architecture with an additional prediction layer.

Figure 3: GraphCite

GraphCite. GAT allows us to encode information about the cita-
tion graph, however the citation phrase should also be considered.
To account for this, we construct the architecture GraphCite based
on the state-of-the-art architecture SciBert and GAT, as shown
in Figure 3. We use two types of embeddings. For the NLP part,
SciBert embeds the citation via the embedding of the [CLS] token.
The node embeddings are the output of GAT, which aggregates
information in the neighborhood of each node. We note that the
GAT architecture receives as node features for the nodes of type
paper, the SciBert embeddings of the titles of the papers. We do not
provide node features for the nodes of type author and venue. We
aggregate the two node embeddings using the Hadamard product,
which has been shown to produce the best edge embeddings [15].
We concatenate these two types of embeddings (citation phrase em-
bedding and edge embedding), passing the result to the MLP. The
loss is jointly backpropagated through the GAT and the SciBert
models. Hence, SciBert is fined tuned on the task while the GAT
network is trained. The loss we use is the multiclass classification
loss over the classes. Our intuition is that a unique loss backpropa-
gated on the two parts of the model would enable learning semantic
features and graph-hereditary features.

4.2 Experimental Evaluation

We compare against several baselines:
Random. We generate predictions by respecting the class distribu-
tions in the training set.
Rule-based classifier. We predict the intent based on the sec-
tion containing the citation. In the SciCite dataset, each citation
has one of five sections tags: introduction, methods, results, back-
ground, discussion. For citations appearing in introduction and
background we predict the intent background, for the section meth-
ods we predict the intent method, while for results and discussion
we predict the intent result. These rules were chosen as they ob-
tained the best results. In the ACL-ARC dataset, the sections are
the following: introduction, related work, method, background, ex-
periments, conclusion. We map the intents as follows: background
for sections introduction and related work, uses for section method,
compare/contrast for experiments and finally future for section
conclusion.
SciBert. We consider the SciBert architecture as described in [3].
This model receives as input only the citation phrase, and it is fine
tuned on citation intent classification.
Pre-Training Variants. We also report results of [9] and [6] on
the ACL-ARC classification task. Those two papers study task-
aware continued pre-training methods and evaluate their systems
on several downstream tasks including citation intent prediction in
the ACL-ARC dataset. These recent methods have high F1-scores,
[9] being current **state-of-the-art** on the ACL-ARC dataset, hence
they are relevant comparisons for our proposed approach. We name
TARTAN and **DAPT** respectively the best methods of [6] and [9].
We note that we do not have the precision and recall results of these
methods, or the results on SciCite.
GAT. Finally, we consider both the structural information of the
citation network and titles in order to compute paper embeddings.
We compute the citing and cited paper's node embeddings and pass

them through a trainable MLP network to predict a citation intent. This approach does not consider the citation phrase.

Settings. We train for 10 epochs, with a learning rate of $2e - 5$ as suggested in the SciBert paper, and a batch of 16. We averaged the macro precision, recall and F1 score over five seeds. We implemented our models using PyTorch [14] and PyTorch Geometric [8]. The SciBert model has embeddings of size 768, while the GAT model has embeddings of size 128. The code is available online[5].

Model		Precision	Recall	F1
Random		15.72	15.92	15.74
Rule-based		29.46	36.78	31.51
SciBert		73.28	72.12	71.70
TARTAN		-	-	70.48
DAPT (SOTA)		-	-	75.40
GAT	citations	15.95	3.65	5.56
	citations, authors	17.86	10.71	13.04
	citations, authors, venues	12.18	5.25	7.18
GraphCite	citations	74.91	73.18	72.99
	citations, authors	**78.65**	**78.79**	**77.34**
	citations, authors, venues	69.96	65.18	66.67

Table 2: Macro results (P,R,F1) on the ACL-ARC dataset.

Model		Precision	Recall	F1
Random		33.29	33.28	33.18
Rule-based		55.62	58.19	56.61
SciBert (SOTA)		87.86	84.95	86.15
GAT	citations	33.33	13.81	18.85
	citations, authors	25.00	20.99	22.82
	citations, authors, venues	20.00	14.74	16.96
GraphCite	citations	86.24	**88.06**	86.24
	citations, authors	**89.90**	87.66	**88.54**
	citations, authors, venues	67.54	72.25	69.61

Table 3: Macro results (P,R,F1) on the SciCite dataset.

5 DISCUSSION AND CONCLUSION

In this work, we investigate if more contextual information, e.g. a citation graph, can help to predict citation intent. Our intuition is that a citation graph might reveal patterns such as a specific paper being cited often with the same intent or an author reusing citations in new papers with the same intent. To investigate this hypothesis, we propose a new architecture that accounts for both the citation phrase and the citation network of the citing and cited papers. Because the graph constructed from citations labeled with intent can be very sparse, we investigate different strategies for increasing its density. We note that this might not always be possible, when we lack the required metadata.

We present the results in Table 2 and Table 3. For SciCite, we observe that a simple rule-based method can give very good results.

---

[5]https://github.com/nyxpho/graphcite

This is not the case for ACL-ARC, partly because we have more citation intents than section types. Our method, GraphCite gives better results that the previous state-of-the-art when considering both the citation graphs containing the papers, and citation graphs that contain both the papers and the authors. For the latter configuration, the results are substantially better than the SOTA, with an increase of 1.96 in macro $F1$ score for ACL-ARC[6], and an increase of 2.36 for SciCite. We note that the improvement is both in precision and recall. However, adding the venues leads to a drop in performance. This can be explained because venues are not important in predicting the intent, and venues nodes in the citation graph add noise to the task. We also observe that the GAT model is insufficient to have a good performance, and the additional information carried by citation phrase has to be incorporated. Finally, we recall that for the ACL-ARC data we could retrieve the authors' information for only around 35% of the papers, hence it is very likely that we could observe a bigger increase in performance if the complete data was available.

In conclusion, we have showed in this work that citation intent prediction benefits significantly from macro-level information related to the citation graph of a paper. Citation intent classification can be an important tool in assessing the impact of papers or in recommending papers according to the reader's goal, e.g. finding papers that extend a given work, or finding papers that use resources created in a given article.

In the future, we will investigate the use of our method on larger and more complete citation graphs. One interesting direction is to include the papers cited by the *cited paper*, for which we might not know the citation intent, but which could leveraged to obtain a better representation of the cited paper node. The attention mechanism underpinning our method will then learn to select relevant information in the neighborhood of the papers. Adding more textual information, in the form of the paper abstract or authors' research interest is another promising direction.

ACKNOWLEDGEMENTS

REFERENCES

[1] Zafar Ali, Pavlos Kefalas, Khan Muhammad, Bahadar Ali, and Muhammad Imran. 2020. Deep learning in citation recommendation models survey. *Expert Systems with Applications* 162 (2020), 113790. https://doi.org/10.1016/j.eswa.2020.113790

[2] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the Literature Graph in Semantic Scholar. In *NAACL-HLT (3)*.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[4] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3586–3596. https://doi.org/10.18653/v1/N19-1361

---

[6]It is interesting to note that we increased the F1-score by 5.64 from SciBert, so we might have reached an even better score by using DAPT or TARTAN.

[5] Randall Collins. 1994. Why the social sciences won't become high-consensus, rapid-discovery science. In *Sociological forum*, Vol. 9. Springer, 155–177.

[6] Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. 2022. Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative. In *International Conference on Learning Representations*. https://openreview.net/forum?id=2bO2x8NAIMB

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

[9] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[10] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 1025–1035.

[11] Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics* 124, 3 (2020), 1907–1922.

[12] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics* 6 (2018), 391–406. https://doi.org/10.1162/tacl_a_00028

[13] Anita Khadka, Ivan Cantador, and Miriam Fernandez. 2020. Exploiting Citation Knowledge in Personalised Recommendation of Recent Scientific Publications. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 2231–2240.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[15] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. 2018. Verse: Versatile graph embeddings from similarity measures. In *Proceedings of the 2018 world wide web conference*. 539–548.

[16] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).