# Graph-Based Structure Aware Citation Intent Classification

**Xinwei Du, Kian Ahrabian, Arun Baalaaji Sankar Ananthan, Richard Delwin Myloth, Jay Pujara**

University of Southern California, Information Sciences Institute
{xinweidu, ahrabian, arunbaal, myloth, jpujara}@usc.edu

## Abstract

Citations are scientists' tools for grounding their innovations and findings in the existing collective knowledge. However, not all citations are semantically identical. Scientists use citations at different parts of their work to convey precise information. Hence, to understand scientific documents best, it is crucial for machines to recognize the intent behind each citation. Current state-of-the-art methods rely on contextual sentences surrounding each citation to classify the intent. In the absence of the actual content, these approaches become unusable. In this work, we propose a text-free citation intent classification method. The proposed method uses a knowledge graph built on top of the SciCite dataset to extract citation information for publications and learn to predict citation intent. We study this problem in both transductive and inductive settings. Our experimental results show that we can achieve a comparable macro F1 score to word embedding content-based methods by only relying on a knowledge graph. Specifically, we achieve macro F1 scores of 62.16 and 59.81 in the transductive and inductive settings, respectively, on the link-level SciCite dataset.

## 1 Introduction

Citations are the primary way of identifying past contributions and connecting progress in new publications to existing literature. Nevertheless, not all citations indicate the same meaning. Authors use citations sparingly with specific intent behind them. For example, some papers are cited for providing background information in a domain, while others are cited for using their methodologies. There are also scenarios where the same paper is used as background information and methodology use-case in different contexts simultaneously. Understanding citation intent is crucial to studying scholarly works, given the universality of using citations. In this work, we propose a general citation intent classification method that relies purely on structural information.

Besides helping researchers better understand the relationship among publications, citation intent analysis has been used for studying various other aspects of scientific works such as research domain evolution (Jurgens et al. 2018), scientific impact analysis (Small 2018), scientific document summarization (Cohan and Goharian 2017), and retrieving related scientific works (Ritchie 2009).

The main three categories of citations are "Background," "Method," and "Result" (Cohan et al. 2019). These categories describe the reasons behind making a scientific connection, referencing a publication in another publication. Classifying citations into these groups has traditionally required a high level of expertise in the respective scientific domains. This constraint, combined with the high cost of expert human labor, has resulted in highly scarce datasets, which makes the task even more difficult.

Many feature engineering-based (Jurgens et al. 2018) and representation learning-based (Beltagy, Lo, and Cohan 2019) methods have been proposed to classify citation intent. However, most of these methods rely heavily on textual information. As a result, they require a complex multi-stage pipeline of parsing documents, identifying citation contexts, and predicting citation intent (Lo et al. 2020). Besides being prone to error propagation from various pipeline stages, the use of these models is limited to situations where the full text is available in a proper format.

This work introduces a pure graph-based approach to classifying citation intent. To this end, we extend the existing SciCite dataset with multi-hop neighborhoods extracted from the Semantic Scholar corpus. Our main idea is to use contextual, relational patterns to make predictions, relinquishing the need for textual context. Given this newly built knowledge graph (KG), we cast the intent classification problem into the common link prediction problem on KGs. This conversion allows us to adapt and extend widely used KG embedding models to this problem. We study the link prediction problem in both transductive and inductive settings. Our experimental results show that although our KG-based method underperforms compared to the large language model-based approaches, it is comparable or even superior to the word embedding-based methods. This finding further signifies the importance of relational patterns for citation intent classification.

Our contributions can be described as follows:

1. Extending the SciCite dataset using the Semantic Scholar corpus to generate a large-scale citation graph.
2. Gathering weakly labeled data to create a large-scale knowledge graph.
3. Introducing a novel graph-based approach to citation intent classification.

4. Presenting benchmarks for both transductive and inductive settings.

## 2 Related Work

**Citation Function/Intent Schemes:** Many prior works have studied the problem of creating categorical schemes for citation intent which in some works is referred to as citation function (Hernández-Alvarez and Gomez 2016). Earlier works were focused on creating more fine-grained categories, going as far as defining 35 (Garzone 1998) and 12 (Teufel, Siddharthan, and Tidhar 2006) fine-grained schemes for scientific arguments. The more recent works however have focused on creating more concise categories. For example, ACL-ARC (Jurgens et al. 2018) proposes a 6-class intent categorization scheme: Background, Motivation, Uses, Extension, Comparison or Contrast, and Future. SciCite (Cohan et al. 2019) is even more restrictive and drops or combines small fine-grained classes to provide a more concise 3-class annotation scheme: Background, Method, and Result.

**Citation Intent Classification Methods:** Before the explosion of deep learning approaches, most methods relied on a combination of hand-crafted features and classic machine learning models. For example, in one instance (Valenzuela, Ha, and Etzioni 2015), authors propose 12 different features, including citation count, PageRank value, and author overlap, and use classic machine learning models such as SVM and Random Forest for classification. In another instance (Jurgens et al. 2018), authors define pattern-based, topic-based, and prototypical argument features and use SVM to make predictions.

With the advent of deep learning models and the emergence of large language models in recent years, representation learning-based methods have outperformed the hand-crafted methods achieving a higher accuracy by considering the textual information. Recent works have proposed the use of structural scaffolds (Cohan et al. 2019), BERT-based models trained on the scientific corpus (SciBERT) (Beltagy, Lo, and Cohan 2019), word embedding-based approaches (Roman et al. 2021), and creating a heterogeneous context graph based on an academic network (Yu et al. 2020)

**Knowledge Graph Embedding Models:** KGs are structured information repositories consisting of a set of nodes representing entities and a set of typed edges representing relations. Since, in most cases, the KG nodes and edges are not attributed, KG embedding (KGE) models aim to learn low-dimensional representations for all entities and relations. The most common traditional shallow KGE methods are TransE (Bordes et al. 2013), ComplEx (Trouillon et al. 2016), and RotatE (Sun et al. 2019). More recent GNN-based KGE methods leverage the message-passing scheme of GNNs, enabling more complex multi-hop reasoning. Examples of these methods are GCN (Kipf and Welling 2016), which leverages the spectral information for information propagation but is limited to mono-relational KGs, R-GCN (Schlichtkrull et al. 2018), which extends GCN to support multi-relational KGs, and GraphSAGE (Hamilton, Ying, and Leskovec 2017) which introduces an inductive framework to handle unseen nodes.

## 3 Dataset

The SciCite dataset focuses on individual citation links and ignores the significance of broader relational connections and features. To overcome this issue, we construct a knowledge graph by mapping each entity in the SciCite dataset to the Semantic Scholar corpus and adding their multi-hop citation neighborhoods. Moreover, the SciCite dataset is tailored for sentence classification methods. Hence, to adapt it to our link prediction setting, we reconstruct two datasets: SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$.

### 3.1 Entity Mapping

We first map each paper in the SciCite dataset to the Semantic Scholar corpus by matching SciCite's IDs to Semantic Scholar's SHA IDs. Since a publication could have many SHA IDs and only one Corpus ID, we then map each SHA ID to the unique Corpus ID to extract unique entities. From the 13,080 papers with unique IDs in SciCite, we successfully map 13,019 of them to valid SHA IDs in semantic scholar, while the remaining 61 papers do not have any corresponding records. We believe this is due to publication removals, as the SciCite dataset was created from the Semantic Scholar corpus in 2019. After converting SHA IDs to Corpus IDs, we end up with 13,011 unique entities and 8 duplicate entities.

### 3.2 Dataset Splitting

The original SciCite dataset contains 11,020 human-labeled samples. These samples are labeled based on citation sentences by human experts. We map sentence-based samples onto a link prediction task and create two datasets: 1) SciCite$_{\text{origin}}$ which uses the same split as SciCite, and 2) SciCite$_{\text{resplit}}$ which creates a new split from the whole dataset. Table 1 showcases the statistic of these datasets.

| Dataset | SciCite | SciCite$_{\text{origin}}$ | SciCite$_{\text{resplit}}$ |
|---|---|---|---|
| Level | Sentence | Link | Link |
| # Samples | 11,020 | 10,379 | 5,766 |
| # Train | 8,243 | 7,602 | 4,122 |
| # Validation | 916 | 916 | 822 |
| # Test | 1,861 | 1,861 | 822 |

Table 1: The statistic of the SciCite dataset and reconstructed datasets.

**SciCite$_{\text{origin}}$:** To make methods comparable, we use the same validation and test sets as SciCite for this dataset and try to keep the training set as close as possible. We convert each publication in the SciCite dataset to a Semantic Scholar entity using the mapped Corpus IDs and drop the contextual sentence-level information. We assign a random unique ID to publications without a Corpus ID. After this procedure, we end up with a set of links for our link prediction task.

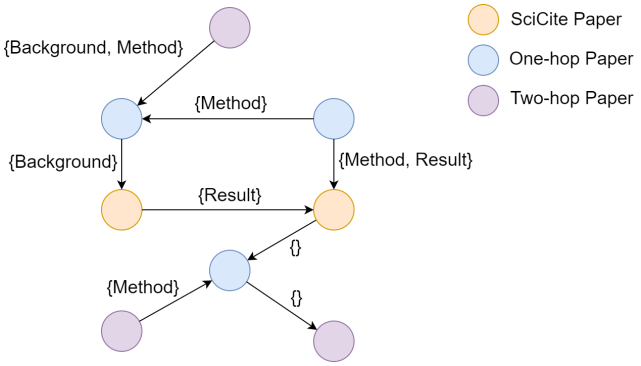


Figure 1: Overview of the extracted multi-hop KG. The set of 0-hop nodes $\mathcal{V}_0$ includes all the orange nodes. The set of 1-hop nodes $\mathcal{V}_1$ includes all the orange and blue nodes. Similarly, the graph could be expanded to include $k$-hop nodes $\mathcal{V}_k$. The annotated set on each edge represents that specific link's intent. Specifically, the empty set denotes that the citation link has no intent label.

Due to the removal of the contextual information, some of the training links appear exactly the same in the test set. Hence, we remove 641 training set samples that also appear in the test set to prevent data leakage.

Moreover, since only one link in the test set has multiple intents, we treat the link prediction problem as a multi-class task rather than a multi-label task. In this scenario, the multi-intent links are represented as separate samples with the same inputs and different outputs. The multi-label variation is left to explore in future works.

**SciCite$_{\text{resplit}}$:** Even though we convert the SciCite dataset to SciCite$_{\text{origin}}$, problems, such as duplicate citations and multi-label links, still exist. Therefore, we further tailor the SciCite dataset to create a better link prediction dataset for graph-based models. First, we remove all the entities, and their related samples, that do not have a mapped Corpus ID. Then, similar to SciCite$_{\text{origin}}$, we convert the remaining samples to a set of links. Following this, we drop all duplicate samples. Among the remaining 6,458 unique links, 5,886 only have one intent, 489 have two intents, and 83 have all three intents. We remove all the multi-intent links and resplit the dataset with ratios of 70%/15%/15% for training, validation, and test sets, respectively.

### 3.3 Knowledge Graph Construction

The Semantic Scholar corpus contains more than 206 million publications and 2.49 billion citation links. Moreover, apart from the regular citation links, this corpus provides partial intent labels for citations using a 3-class scheme as follows:

1. **Background**: Describe a problem, topic, or concept
2. **Method**: Provide a method, tool, or dataset
3. **Result**: To make a comparison

At a sentence level, these intent labels are extracted using the structural scaffolds model (Cohan et al. 2019). We refer to this data as weakly labeled due to being labeled by a noisy model rather than a human expert. Since the intent labels are partial at a sentence level, citation links could have zero intent in the absence of text or several intents in an abundance of use cases.

We expand the SciCite dataset using the mapped entities to construct a KG containing multi-hop neighborhoods of those entities. Figure 1 illustrates an overview of the expanded KG. This work uses the 2022-09-13 version of the corpus downloaded from the bulk API. Formally, given the set of mapped entities $\mathcal{V}_0$, the set of $k$-hop nodes $\mathcal{V}_k$ is defined as

$$\mathcal{V}_k = \mathcal{V}_{k-1} \cup \{y \mid \exists x \in \mathcal{V}_{k-1} : y \in \mathcal{N}_x\} \tag{1}$$

where for a given entity $x$, $\mathcal{N}_x$ denotes all the entities that cite or are cited by $x$, i.e., the set of neighboring entities. Given the sets of unlabeled links $\mathcal{U}$ and weakly labeled links $\mathcal{L}$, the set of $k$-hop edges $\mathcal{E}_k$ is defined as

$$\mathcal{E}_k^{\mathcal{U}} = \{(x, y, \text{UNK}) \mid x, y \in \mathcal{V}_k, (x, y) \in \mathcal{U}\} \tag{2}$$

$$\mathcal{E}_k^{\mathcal{L}} = \cup_r \{(x, y, r) \mid x, y \in \mathcal{V}_k, (x, y) \in \mathcal{L}_r\} \tag{3}$$

$$\mathcal{E}_k = \mathcal{E}_k^{\mathcal{U}} \cup \mathcal{E}_k^{\mathcal{L}} \tag{4}$$

where $r \in \{\text{Background, Method, Result}\}$ and $\mathcal{L}_r$ denotes the set of all weakly labeled links with label $r$. Consequently, given the sets of $k$-hop nodes $\mathcal{V}_k$ and edges $\mathcal{E}_k$, the extracted $k$-hop KG, $\mathcal{G}_k$, is defined as

$$\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k) \tag{5}$$

The specific statistics of the extracted KG and the original semantic scholar corpus are reported in Table 2. Although we extract $\mathcal{G}_2$, given its scale, we opt to run our current experiment only on $\mathcal{G}_1$ and leave the larger-scale experiments for future works.

## 4 Method

Throughout the rest of this work, for simplicity, we use the term **publication** to denote all types of academic publications, e.g., books and papers. Moreover, we use the terms **citation** and **reference** to denote incoming and outgoing links, respectively.

### 4.1 Features Engineering

We propose representing publications through their references, citations, and graph-based features. More specifically, we extract the in-degrees and out-degrees of citations (or references), background links, method links, and result links. As a result, each paper is represented with an 8-dimensional feature vector, 4 for each in-degree and out-degree feature. For the publications where the content is unavailable, the out-degree intent-based features will be zero since those features are based on the noisy sentence-level model that the Semantic Scholar uses. However, the in-degree features may not be zero as long as the citing paper's content is available. For the new publications, i.e., unseen nodes in the inductive setting, the only known non-zero feature is the reference count.

| Dataset | # Nodes | # Citation Links | # Background | # Method | # Result |
|---|---|---|---|---|---|
| Zero-Hop ($\mathcal{G}_0$) | 13,011 | 10,733 | 5,479 | 4,403 | 1,335 |
| One-Hop ($\mathcal{G}_1$) | 5,862,261 | 119,776,090 | 39,202,086 | 16,830,665 | 16,830,665 |
| Two-Hop ($\mathcal{G}_2$) | 57,535,880 | 1,621,293,902 | 467,860,523 | 121,877,053 | 35,283,718 |
| Semantic Scholar | 206,159,629 | 2,495,513,737 | 643,955,457 | 169,472,164 | 45,779,793 |

Table 2: Statistics of the extracted KGs along with the original Semantic Scholar corpus

We normalize the reference and citation features by a biased log factor defined as

$$\bar{h}_x = \log_{10}(h_x + 1 + \alpha) \tag{6}$$

where $\alpha$ is a bias hyperparameter. We specifically set $\alpha = -0.9$ to get a normalized value of $-1$ for zero-reference and zero-citation situations.

Moreover, we normalize the non-zero in-degree intent-based features into a $[0, 1]$ probability distribution as follows:

$$\bar{h}_x = \frac{h_x}{h_{\text{Background}} + h_{\text{Method}} + h_{\text{Result}}} \tag{7}$$

The same normalization step is used for out-degree features separately.

### 4.2 Multi-Hop Link Prediction (MHLR)

Transductive and inductive settings are the most common link prediction evaluating schemes for KGs. The main difference between these two settings is having a fixed set of nodes in both the training and evaluation phases (transductive) versus allowing the addition of unseen nodes in the evaluation phase (inductive). This work refers to citation intent prediction on unseen publications as the inductive setting, whereas the transductive setting refers to citation intent prediction on already seen publications.

We propose an adaptable graph-based model for citation intent prediction in both the transductive and inductive settings. The primary basis of this approach is that a node, i.e., publication, could be represented as a combination of the neighboring nodes' representations. Let $h_x^{(0)}$ be the extracted feature vector for any arbitrary node $x$. We calculate the representation of an arbitrary node $v$ at layer $l{+}1$ of a multilayer model as

$$h_{\mathcal{N}_v}^{(l+1)} = \frac{1}{|\mathcal{N}_v|} \sum_{u \in \mathcal{N}_v} h_u^{(l)} \tag{8}$$

$$h_v^{(l+1)} = \sigma(W^{(l+1)}[h_v^{(l)} \| h_{\mathcal{N}_v}^{(l+1)}]) \tag{9}$$

where $\sigma$ is a non-linear function. Throughout our experiments, we specifically use ReLU to introduce non-linearity. Given the node representation from a $L$-layer model and a link $(u, v)$, we calculate the logit values as

$$p = \text{MLP}([h_u^{(L)} \| h_v^{(L)}]) \tag{10}$$

where $p \in \mathbb{R}^{\mathcal{C}}$ contains the unnormalized logits for each class and $\mathcal{C}$ is the set of all classes. The predicted class $c$ is then calculated as

$$\text{argmax}_c \ \text{sigmoid}(p). \tag{11}$$

The main disadvantage of the inductive settings is that the unseen nodes only have one available feature, i.e., reference count. This absence of information makes the task extremely difficult, as the feature vectors are highly sparse. However, our model tries to diminish this effect by using the message-passing scheme, as defined in Equation 9, to aggregate information through connected entities, i.e., cited papers, creating a denser representation for the unseen nodes.

All our models are trained using the cross-entropy loss defined as

$$l_n = -\log \frac{\exp(p_{y_n})}{\sum_{i=1}^{|\mathcal{C}|} \exp(p_i)} \tag{12}$$

where and $p_x$ is the logit value for class $x$ given the prediction vector $p$.

### 4.3 Baselines

**Knowledge Graph Embedding Models**: Traditional KGE models consist of two shallow embeddings as entity and relation encoders and a score function as a decoder to predict the likelihood of a link. These models are trained in a contrastive way by masking either one of the entities in a given triplet (head, relation, tail) and sampling a set of negative entities, contrasting the positive entity.

Since the traditional KGE methods rely on shallow embeddings for encoding entities and relations, they can only be used in the transductive setting and cannot operate on unseen nodes. For our experiments, we use the available implementations of TransE, ComplEx, and RotatE in the DGL-KE toolkit (Zheng et al. 2020). In the evaluation phase, we calculate the likelihood of all different relation types for each link and consider the highest likelihood as the model's intent prediction.

**Hybrid Models**: To increase the reasoning power of the traditional KGE models, we devise a two-stage approach based on multilayer perceptron (MLP). We first use the traditional KGE models to learn embeddings for entities and relations. Then, instead of relying on the produced likelihood scores, we concatenate the vectors of two entities and pass that through an MLP to get logit values. Formally, given a link $(u, v)$ and their respective learned representation $(z_u, z_v)$, we calculate the logit values as

$$p = \text{MLP}([z_u \| z_v]) \tag{13}$$

where $p \in \mathbb{R}^{\mathcal{C}}$ contains the unnormalized logits for each class. The predicted class $c$ is then calculated as

$$\text{argmax}_c \ \text{sigmoid}(p). \tag{14}$$

|  |  | SciCite$_{\text{origin}}$ |  |  |  | SciCite$_{\text{resplit}}$ |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Method | Setting | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| Random | Universal | 33.05 | 33.05 | 33.83 | 31.22 | 32.99 | 32.88 | 33.85 | 31.89 |
| Most Common | Universal | 53.57 | 17.86 | 33.33 | 23.26 | 42.63 | 14.21 | 33.33 | 19.93 |
| TransE | Transductive | 40.41 | 37.09 | 37.81 | 36.52 | 39.57 | 35.96 | 35.70 | 35.59 |
| ComplEx | Transductive | 49.01 | 44.11 | 37.94 | 33.30 | 40.25 | 41.85 | 35.64 | 28.78 |
| RotatE | Transductive | 23.54 | 32.97 | 32.74 | 22.98 | 28.12 | 36.88 | 36.31 | 27.88 |
| TransE + MLP | Transductive | 54.16 | 45.77 | 45.21 | 45.24 | 51.93 | 45.68 | 44.16 | 43.89 |
| ComplEx + MLP | Transductive | 55.72 | 47.80 | 45.19 | 44.77 | 48.64 | 43.46 | 43.15 | 43.24 |
| RotatE + MLP | Transductive | 56.37 | 48.79 | 46.15 | 46.55 | 51.81 | 46.92 | 45.46 | 45.63 |
| Random + MLP | Transductive | 49.60 | 30.58 | 35.17 | 32.42 | 45.35 | 30.26 | 35.83 | 32.78 |
| Infersent-KMeans | Universal | - | 58 | 64 | 60 | - | - | - | - |
| Infersent-HDBSCAN | Universal | - | 57 | 63 | 58 | - | - | - | - |
| Glove-KMeans | Universal | - | 51 | 56 | 51 | - | - | - | - |
| Glove-HDBSCAN | Universal | - | 52 | 57 | 52 | - | - | - | - |
| Ours (MHLR) | Transductive | 66.20 | 62.18 | 56.13 | 57.88 | 66.10 | 63.69 | 61.33 | 62.16 |
| Ours (MHLR) | Inductive | 63.94 | 58.36 | 55.05 | 56.13 | 64.17 | 59.86 | 59.83 | 59.81 |
| Structural Scaffolds | Universal | - | 84.7 | 83.6 | 84.0 | - | - | - | - |
| SciBERT | Universal | - | - | - | 85.49 | - | - | - | - |

Table 3: Intent classification results on SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$ datasets. All the metrics are macro averaged.

**Natural Language Processing Models**: We include the reported results of several state-of-the-art Natural Language Processing (NLP) methods. Specifically, we include results from the word embedding-based methods such as Infersent-KMeans, Infersent-HDBSCAN, Glove-KMeans, and Glove-HDBSCAN (Roman et al. 2021), BiLSTM-based method Structural Scaffolds (Cohan et al. 2019), and large language model-based method SciBERT (Beltagy, Lo, and Cohan 2019). All these methods make use of textural information.

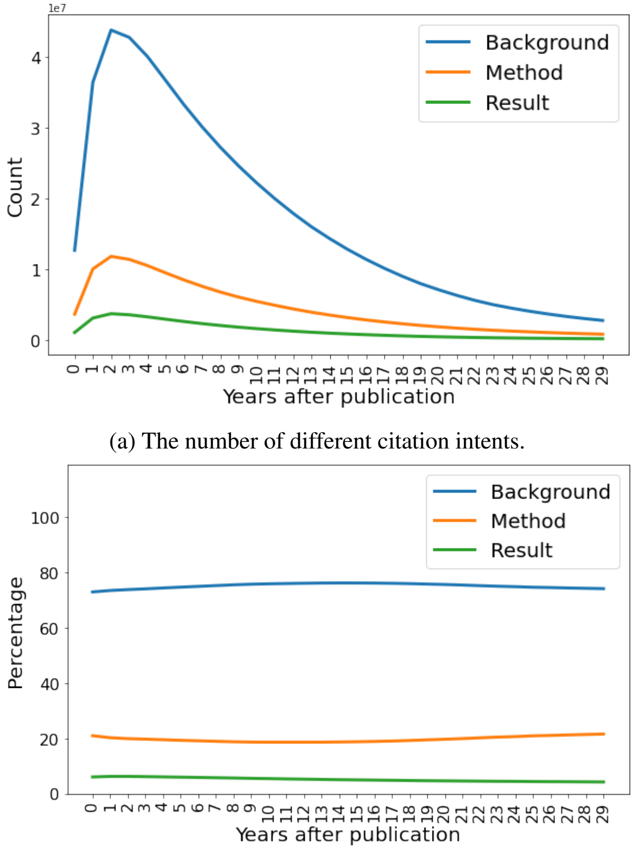
## 5 Experiments

In this section, we report our experimental results on both SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$ datasets. All the graph-based experiments are carried out on the $\mathcal{G}_1$ KG. For the MHLR method, in both transductive and inductive settings, we use a 1-layer variation on top of the normalized features extracted as described in Sections 4.1. For traditional KGE methods, we tune their hyperparameters as described in Appendix A.1 and train them using the hyperparameters showcased in Table 4. For hybrid methods, the KGE component is first trained to generate node features using the hyperparameters described in Table 4. Then, the MLP component is trained using the procedure described in A.2 to predict the citation intent. To control for the effect of the pre-training using traditional KGE models, we also run a variation with randomly initialized node features and designate it as "Random + MLP." For the NLP models, we use the reported results to compare our models on the test set-aligned SciCite$_{\text{origin}}$ dataset. Finally, we also include the results from random and most common class predictions as sanity checks.

### 5.1 Results

Table 3 illustrates our experimental results on both graph-based datasets. As evident from Table 3, traditional KGE methods perform poorly on both datasets, only slightly beating the random baseline on the macro F1 metric. However, when combined with MLP models, all exhibit significant performance boost, up to more than 100% in the case of RotatE. Moreover, the control "Random + MLP" experiment showcases very similar results to the random baseline, indicating the importance of both components for the hybrid model to perform well. Furthermore, it is evident that the reasoning power of shallow traditional KGE models is not enough to capture the complexity of this task, and we require models with more reasoning power.

Our method achieves 57.88 and 62.16 macro F1 scores on SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$ datasets. Furthermore, the inductive results showcase the resilience of our approach in an out-of-distribution setting, losing less than 1% of the F1 score. Compared to previously reported results (Roman et al. 2021), our model achieves superior performance to Glove-based models while slightly lagging behind Infersent-based models. The significance of these results is that we show structural and relational information could be used to achieve relatively high performance without using textual information. Moreover, although our models underperform compared to language model-based approaches such as Structural Scaffolds and SciBERT, we showcase interesting future directions for combining graph-based and NLP-based methods.



(a) The number of different citation intents.

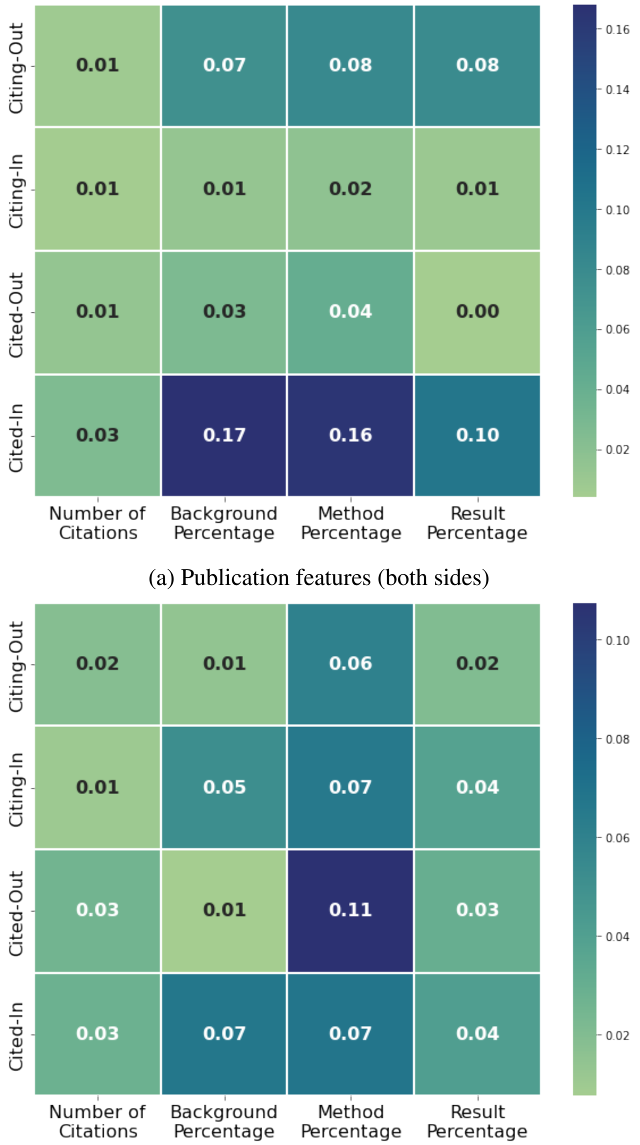
(b) The percentage of different citation intents.

Figure 2: The statistic of citation intent for all publications in the Semantic Scholar corpus.

# 6 Analysis

## 6.1 Temporal Analysis

This analysis studies the relationship between the time that has passed since publication and citation intent. We hypothesize that a paper is more likely to be cited as “Result” or “Method” right after its publication, and as time passes, it will be more likely to be cited as “Background.” If this is proven accurate, we could get a relatively strong signal from the temporal information for each citation. We plotted the years after publication against intent counts and ratios for all papers in the semantic scholar corpus to test our hypothesis. Figure 2a and 2b illustrate the results of our analysis. As evident from these figures and contrary to our original hypothesis, we find out that the ratio of intent classes almost stays the same as time passes with insignificant fluctuations. As a result, using temporal information in our models is unlikely to provide any significant improvement. Note that these results should be taken in with a grain of salt as all these extracted weak labels are generated by another model that could potentially be biased. Hence, it should not discourage further analysis or studies of temporal information for citation intent classification.



| | Number of Citations | Background Percentage | Method Percentage | Result Percentage |
|---|---|---|---|---|
| Citing-Out | 0.01 | 0.07 | 0.08 | 0.08 |
| Citing-In | 0.01 | 0.01 | 0.02 | 0.01 |
| Cited-Out | 0.01 | 0.03 | 0.04 | 0.00 |
| Cited-In | 0.03 | 0.17 | 0.16 | 0.10 |

(a) Publication features (both sides)

| | Number of Citations | Background Percentage | Method Percentage | Result Percentage |
|---|---|---|---|---|
| Citing-Out | 0.02 | 0.01 | 0.06 | 0.02 |
| Citing-In | 0.01 | 0.05 | 0.07 | 0.04 |
| Cited-Out | 0.03 | 0.01 | 0.11 | 0.03 |
| Cited-In | 0.03 | 0.07 | 0.07 | 0.04 |

(b) Averaged neighborhood features (both sides)

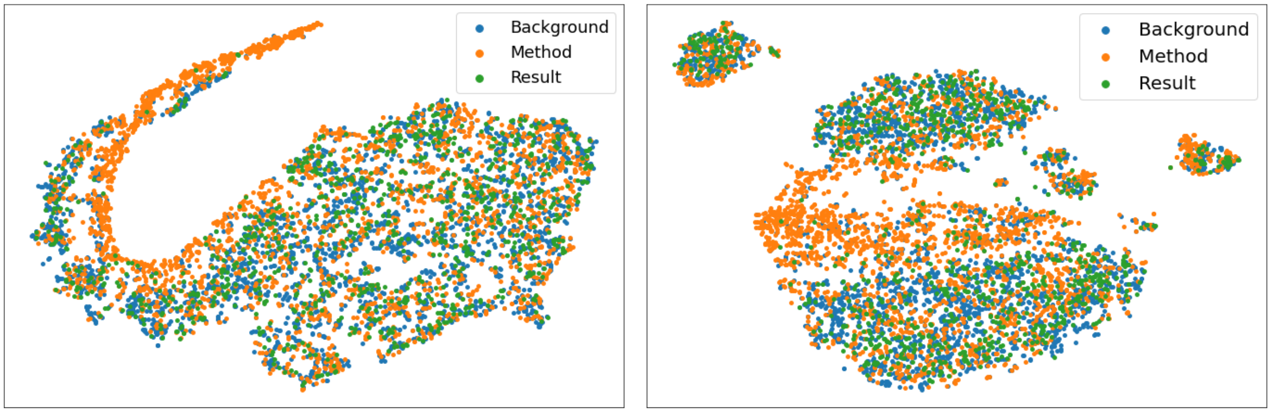
Figure 3: The calculated MI values for publication features and averaged neighborhood features.

## 6.2 Mutual Information Analysis

In this analysis, we study the quality of the engineered features as described in Section 4.1 concerning the weakly labeled intent classes. To this end, we use the well-known mutual information (MI) (Kraskov, Stögbauer, and Grassberger 2004) measurement to quantify the importance of each feature. Formally, the MI between two discrete random variables $X$ and $Y$ is defined as

$$I(X,Y)=\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}P_{X,Y}(x,y)\log(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}) \quad (15)$$

where $\mathcal{Y}$ is the value space for $Y$, $\mathcal{X}$ is the value space for $X$, $P_{X,Y}$ is the joint probability distribution, and $P_X$ and $P_Y$



(a) Features before normalization

(b) Features after normalization

Figure 4: The t-SNE visualizations for the unnormalized and normalized features.

are the marginal probability distributions. Note that MI is a non-negative value, and higher values indicate more correlation between the two random variables. For our analysis, we calculate MI for both sides of the 5,886 unique citation links in the SciCite$_{\text{resplit}}$ dataset. Moreover, to study these features in the graph context, we also calculate MI for the average of these features over the neighborhood of each publication, i.e., all citing and cited publications, from both sides of the citation links. Figures 3a and 3b present the results of our experiments. As evident from these results, the neighborhood-averaged features generally show stronger connections to the target variable, which is aligned with better results that we achieved using our proposed approach.

### 6.3 Feature Quality Analysis

In this analysis, we study the effect of normalization as described in Equations 6 and 7. To this end, we project the extracted features of the 5,886 unique citation links in the SciCite$_{\text{resplit}}$ dataset to a 2-dimensional space using t-SNE (Van der Maaten and Hinton 2008). Figure 4a and 4b illustrate the projected space for the unnormalized and normalized features, respectively. As evident from Figure 4a, it is challenging to distinguish different intent types in the unnormalized space. However, after normalization, as evident from Figure 4b, we can see that the "Method" intention more or less creates a distinguishable cluster. This result shows that the use of normalization is potentially helpful for the model. Further studies on different types of normalization and their effects are left for future work.

## 7 Conclusions and Future Work

In this work, we first introduced an expansion to the SciCite dataset by extracting scholarly information from the Semantic Scholar corpus and creating an extended citation graph. Then, we gathered a large-scale weakly labeled dataset to augment the extracted citation graph with citation intents and create a multi-relational knowledge graph. Following this, we adapted the sentence-based intent classification into a citation-based link prediction task on graphs. We then introduced a set of engineered graph-based and citation-based features. Built on top of these features, we introduced a graph-based multi-hop reasoning approach for the newly introduced task. Our approach achieves 62.16 and 59.81 macro F1 scores in the transductive and inductive settings, respectively. The result in the inductive setting showcases the robustness of the proposed approach in the information-deprived out-of-distribution environment. Moreover, compared to NLP-based models, we reached a comparable performance to, and in some cases outperform, the word embedding-based methods that rely on contextual sentences to make predictions. These results further signify the strong signal in relational information and highlight the importance of future analysis and studies in this domain.

For future works, one straightforward idea is to extend the knowledge graph with more scholarly information, such as venues, institutions, and fields of study. There already exist some open repositories such as OpenAlex (Priem, Piwowar, and Orr 2022), and Microsoft Academic Graph (MAG) (Wang et al. 2020) that contain this information. Another direction is further investigation into the temporal signals. Finally, since the link-level intent classification has its natural limitations due to ignoring the contextual information for each citation, a fusion between graph-based and NLP-based methods could prove superior to the current state-of-the-art model, i.e., SciBERT.

## Acknowledgements

This work was funded by the Defense Advanced Research Projects Agency with award W911NF-19-20271 and with support from a Keston Exploratory Research Award.

## References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676.*

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Cohan, A.; Ammar, W.; Van Zuylen, M.; and Cady, F. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Cohan, A.; and Goharian, N. 2017. Scientific article summarization using citation-context and article's discourse structure. *arXiv preprint arXiv:1704.06619*.

Garzone, M. A. 1998. *Automated classification of citations using linguistic semantic grammars.* University of Western Ontario.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hernández-Alvarez, M.; and Gomez, J. M. 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22(3): 327–349.

Jurgens, D.; Kumar, S.; Hoover, R.; McFarland, D.; and Jurafsky, D. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6: 391–406.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical review E*, 69(6): 066138.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. Online: Association for Computational Linguistics.

Priem, J.; Piwowar, H.; and Orr, R. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Ritchie, A. 2009. Citation context analysis for information retrieval. Technical report, University of Cambridge, Computer Laboratory.

Roman, M.; Shahid, A.; Khan, S.; Koubaa, A.; and Yu, L. 2021. Citation intent classification using word embedding. *Ieee Access*, 9: 9982–9995.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, 593–607. Springer.

Small, H. 2018. Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, 12(2): 461–480.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Teufel, S.; Siddharthan, A.; and Tidhar, D. 2006. An annotation scheme for citation function. In *SIGDIAL Workshop*.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, 2071–2080. PMLR.

Valenzuela, M.; Ha, V.; and Etzioni, O. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, K.; Shen, Z.; Huang, C.; Wu, C.-H.; Dong, Y.; and Kanakia, A. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1): 396–413.

Yu, W.; Yu, M.; Zhao, T.; and Jiang, M. 2020. Identifying referential intention with heterogeneous contexts. In *Proceedings of The Web Conference 2020*, 962–972.

Zheng, D.; Song, X.; Ma, C.; Tan, Z.; Ye, Z.; Dong, J.; Xiong, H.; Zhang, Z.; and Karypis, G. 2020. DGL-KE: Training Knowledge Graph Embeddings at Scale. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, 739–748. New York, NY, USA: Association for Computing Machinery.

# A Hyperparameters

## A.1 Knowledge Graph Embedding

We use a randomized search to tune our models and find near-optimal hyperparameters using the following ranges: *embedding dimensions* $\in \{50, 100, 200\}$, *learning rate* $\in \{0.03, 0.1, 0.3\}$, *regularization coefficient* $\in \{0.0, \text{1e-9}, \text{1e-8}, \text{1e-7}, \text{1e-6}, \text{1e-5}\}$, *number of negative samples* $\in \{64, 128, 256, 512, 1024\}$, $\alpha \in \{0.25, 0.5, 1\}$, $\gamma \in \{6, 12, 24\}$. Note that $\alpha$ and $\gamma$ are the adversarial temperature and the margin value (RotatE-only), respectively.

| Hyperparameter | TransE | ComplEx | RotatE |
|---|---|---|---|
| embedding dimension | 100 | 100 | 50 |
| learning rate | 0.1 | 0.3 | 0.1 |
| regularization coefficient | 1e-6 | 1e-6 | 1e-6 |
| negative samples size | 128 | 512 | 64 |
| $\alpha$ | 0 | 0.25 | 1 |
| $\gamma$ | - | - | 6 |

Table 4: Hyperparameters of KGE algorithms.

## A.2 Multilayer Perceptron

To simplify the model tuning process, we find the optimal hyperparameters of "ComplEx + MLP" on SciCite$_{\text{origin}}$ using grid search and reuse them for the rest of our experiments. Specifically, we run a grid search over the following ranges: *number of layers* $\in \{0, 1, 2, 3\}$, *dropout* $\in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, *dimension* $\in \{32, 64, 128\}$, The optimal hyperparameters are as follows: *number of layers* $= 2$, *dropout* $= 0.2$, and *dimension* $= [64, 32]$. We use ReLU as the activation function for all layers.