Citation Intent Classification Through Weakly Supervised Knowledge Graphs

Xinwei Du[1,2], Kian Ahrabian[1,2,*], Arun Baalaaji Sankar Ananthan[1,2], Richard Delwin Myloth[1,2] and Jay Pujara[1,2]

[1]Information Sciences Institute, Marina del Ray, CA, USA
[2]University of Southern California, Los Angeles, CA, USA

Abstract

Citations are scientists' tools for grounding their innovations and findings in the existing collective knowledge. They are used for semantically distinct purposes as scientists utilize them at different parts of their work to convey specific information. As a result, a crucial aspect of scientific document understanding is recognizing the authorial intent associated with citations. Current state-of-the-art methods rely on contextual sentences surrounding each citation to classify the intent. However, in the absence of textual content, these approaches become unusable. In this work, we propose a text-free citation intent classification method built on relational information among scholarly works in this work. To this end, we introduce a large-scale knowledge graph built from the publications in the SciCite dataset and their multi-hop neighborhood extracted from The Semantic Scholar Open Research Corpus (S2ORC). We also augment this knowledge graph by adding weakly-labeled links based on the intent information available in the S2ORC. Finally, we cast the intent classification task as a link prediction problem on the newly created knowledge graph. 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 features and relations derived from this 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. Moreover, by combining our method with the state-of-the-art NLP-based model, we achieve improvements across all metrics.

Keywords

Citation Intent Classification, Knowledge Graphs, Graph Neural Networks, Large Language Models, Weakly supervised learning

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 when adopting or adapting a previously-used methodology. 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. Current state-of-the-art citation intent classification models [17, 1, 4] rely heavily on textual information, e.g., the sentences surrounding the citation. However, such information is expensive to obtain and in some scenarios inaccessible altogether. Consequently, we need models that could operate without having access to textual information. Previous works [3, 26, 6] have shown the importance of relational and structural information available in links among publications for various tasks. 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 [10], scientific impact analysis [19], scientific document summarization [5], and retrieving related scientific works [16]. The main three categories of citations are “Result,” “Method,” and “Background” [4]. 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.

Previous works have proposed classifying citation intent through feature engineering-based [10] and representation learning-based [1] methods. However, most of these methods depend on textual information. As a result, they require a complex multi-stage pipeline of parsing documents, identifying citation contexts, and

The Third AAAI Workshop on Scientific Document Understanding 2023, February 14th, 2023, Washington, DC, USA

[*] Corresponding author.

xinweidu@usc.edu (X. Du); ahrabian@usc.edu (K. Ahrabian); arunbaal@usc.edu (A. B. S. Ananthan); myloth@usc.edu (R. D. Myloth); jpujara@usc.edu (J. Pujara)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

predicting citation intent [13]. 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. We extend the existing SciCite dataset with 2-hop neighborhoods extracted from The Semantic Scholar Open Research Corpus (S2ORC). To further enrich the graph, we utilize the intent information provided in the S2ORC to create a weakly supervised knowledge graph (KG) consisting of the publications and the relations that match the provided intents. Our main idea is to use contextualized relational patterns to make predictions, obviating the need for textual context. Given the newly built KG, we cast the intent classification problem into the common link prediction problem on KGs. Specifically, we train a model to learn representations for entities and relations. Using these representations, we run the following query on the KG: $(s, ?, o)$, where s cites o.

Converting this problem into a link prediction task 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. Moreover, our experiments with combining the NLP-based and graph-based methods show slight improvements over the current state-of-the-art model. These findings further signify the importance of relational patterns for citation intent classification.

The contributions of this work are as follows:

1. Extending the SciCite dataset using the S2ORC dataset to generate a large-scale weakly supervised KG.
2. Introducing a novel graph-based approach for citation intent classification built on top of the newly built KG.
3. Presenting benchmarks for both transductive and inductive settings.
4. Presenting analyses on the effect of different parts of the methodology such as weak supervision and feature engineering.

2. Related Work

2.1. 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 [9]. Earlier works were focused on creating more fine-grained categories, going as far as defining 35 [7] and 12 [21] fine-grained schemes for scientific arguments. The more recent works however have focused on creating more concise categories. For example, ACL-ARC [10] proposes a 6-class intent categorization scheme: Background, Motivation, Uses, Extension, Comparison or Contrast, and Future. SciCite [4] 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.

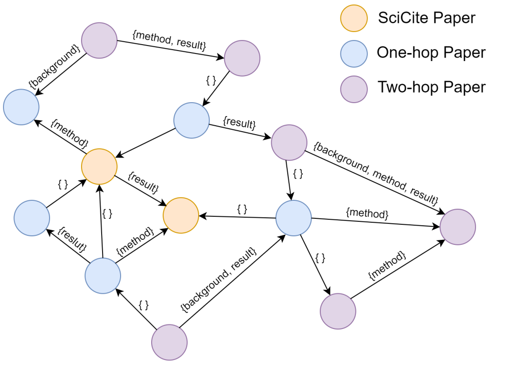


Figure 1: Overview of the extracted multi-hop KG. The set of 0-hop nodes $\mathcal{V}_0$ includes all the orange nodes. The set of 1-hop nodes $\mathcal{V}_1$ includes all the orange and blue nodes. Similarly, the graph could be expanded to include k-hop nodes $\mathcal{V}_k$. The annotated set on each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

2.2. 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 [23], 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 [10], 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 [4], BERT-based models trained on the scientific corpus (SciBERT) [1], word embedding-based approaches [17], and creating a heterogeneous context graph based on an academic

Table 1
The statistic of the SciCite dataset and reconstructed 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

network [26]

2.3. 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 [2], ComplEx [22], and RotatE [20]. 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 [11], which leverages the spectral information for information propagation but is limited to mono-relational KGs, R-GCN [18], which extends GCN to support multi-relational KGs, and GraphSAGE [8] 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 S2ORC and adding their 2-hop citation neighborhoods. The S2ROC contains more than 206 million publications and 2.49 billion citation links. 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

Moreover, the SciCite dataset is tailored for sentence classification methods, where input features are textual excerpts and the output labels are citation intents. We reformulate this task as link prediction on KGs, where the input features are a representation of the source (citing) paper and the target (cited) paper, and the output is the label of a citation link between the source and target. We release all our datasets under a CC-BY-SA license at TBD

3.1. Entity Mapping

We first map each paper in the SciCite dataset to the S2ORC 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 S2ORC 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. Hence, to adapt it to our link prediction setting, we reconstruct two datasets: SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$. SciCite$_{\text{origin}}$ adheres to the same benchmarks reported in prior works but is modified to remove overlapping citation links in the training and test sets. To maximize usage of the training data while removing artifacts, we create SciCite$_{\text{resplit}}$ that performs additional cleaning, provides a stronger separation of training and test sets, and avoids multi-intent citations. Table 1 showcases the statistic of these 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.

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.

Table 2
Statistics of the extracted KGs along with the original S2ORC dataset.

Dataset	# Nodes	# Citation Links	# Background	# Method	# Result	Weak Labels
Zero-Hop ($\mathcal{G}_0$)	13,011	10,733	5,479	4,403	1,335	79.04%
One-Hop ($\mathcal{G}_1$)	5,862,261	119,776,090	39,202,086	16,830,665	16,830,665	43.18%
Two-Hop ($\mathcal{G}_2$)	57,535,880	1,621,293,902	467,860,523	121,877,053	35,283,718	34.41%
S2ORC	206,159,629	2,495,513,737	643,955,457	169,472,164	45,779,793	31.90%

Multi-label methods may be a promising future extension of our work.

SciCite$_{\text{resplit}}$:

Even though we convert the SciCite dataset to the 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.

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. Weak Supervision

In order to enrich our data and provide more information to the models, we extract the set of intents provided in the S2ORC dataset for each citation link. The intent labels in S2ORC are extracted using the structural scaffolds model [4] at a sentence level. In this scenario, we implicitly use the existing data derived from the content for bootstrapping our approach. We refer to these links 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.

4.2. Knowledge Graph Construction

Given the S2ORC dataset, we expand the SciCite dataset using the mapped entities to construct a KG containing 2-hop neighborhoods of the publications. 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. Since not every link has weakly labeled intent, this table also provides the percentage of weakly labeled links for each corresponding graph. 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.

4.3. Feature Engineering

Since none of the publications in our KGs have any features or pre-defined representation, we propose to represent them through their references, citations, and graph-based features. More specifically, from S2ROC 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.

Table 3
Intent classification results on $\text{SciCite}_{\text{origin}}$ and $\text{SciCite}_{\text{resplit}}$ datasets. All the metrics are macro averaged. Bold values represent the highest performance within the metric and dataset scope.

		$\text{SciCite}_{\text{origin}}$				$\text{SciCite}_{\text{resplit}}$			
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
Random + MLP	Transductive	49.60	30.58	35.17	32.42	45.35	30.26	35.83	32.78
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
Infersent-KMeans	Universal	-	58	64	60	-	-	-	-
Infersent-HDBSCAN	Universal	-	57	63	58	-	-	-	-
Glove-KMeans	Universal	-	51	56	51	-	-	-	-
Glove-HDBSCAN	Universal	-	52	57	52	-	-	-	-
MHLP (Ours)	Transductive	66.20	62.18	56.13	57.88	66.10	63.69	61.33	62.16
MHLP (Ours)	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	86.94	85.30	85.92	85.58	86.39	85.51	85.14	85.28
SciBERT + MHLP	Universal	**87.53**	**85.56**	**87.07**	**86.25**	**86.85**	**86.80**	**85.96**	**86.35**

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.

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 α 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.

4.4. 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 [27]. 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

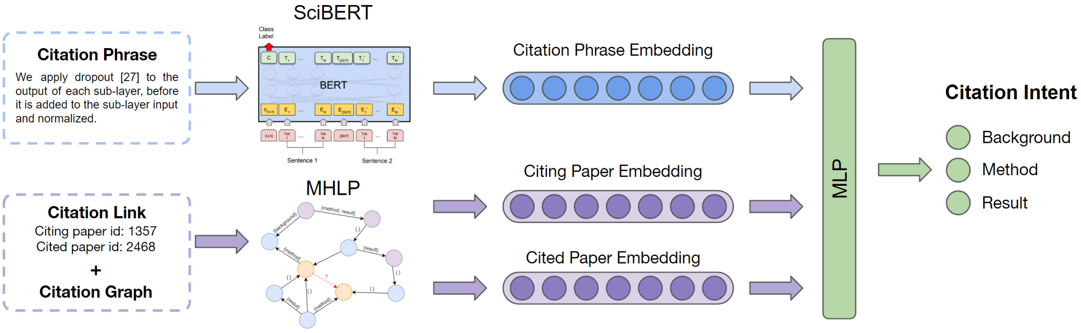


Figure 2: Overview of the composite model. The model consists of two encoders for the citation phrase and the citation graph around the citation link. During the training phase, we freeze the SciBERT model in the first two epochs as a warm-up step for the graph encoder; then, we jointly train both encoders along with the final prediction module.

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

$$p = \text{MLP}([z_u \| z_v]) \tag{8}$$

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

$$\text{argmax}_c \ \text{sigmoid}(p). \tag{9}$$

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 [17], BiLSTM-based method Structural Scaffolds [4], and large language model-based method SciBERT [1]. Moreover, we report the results of fine-tuning a pre-trained SciBERT model on both datasets. All these methods use textural information and are evaluated on the SciCite dataset.

4.5. Multi-Hop Link Prediction (MHLP)

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

$$h_{\mathcal{N}_v}^{(l+1)} = \frac{1}{|\mathcal{N}_v|} \sum_{u \in \mathcal{N}_v} h_u^{(l)} \tag{10}$$

$$h_v^{(l+1)} = \sigma(W^{(l+1)}[h_v^{(l)} \| h_{\mathcal{N}_v}^{(l+1)}]) \tag{11}$$

where σ 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

$$p = \text{MLP}([h_u^{(L)} \| h_v^{(L)}]) \tag{12}$$

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

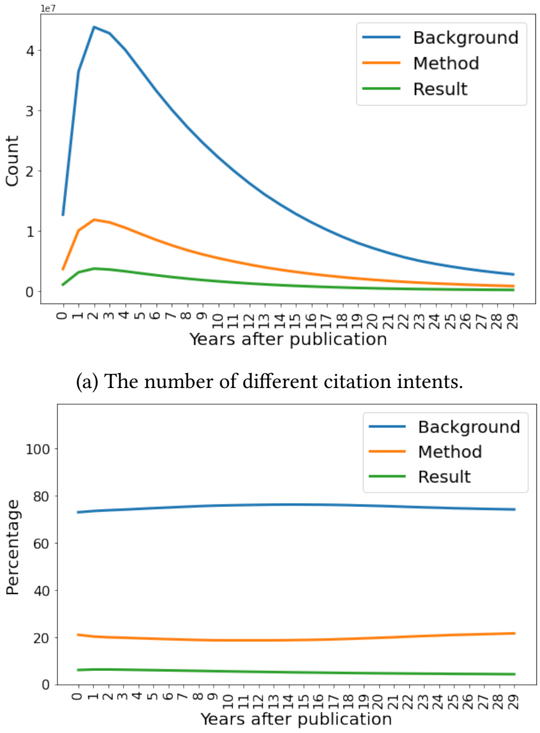
$$\text{argmax}_c \ \text{sigmoid}(p). \tag{13}$$

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 11, 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

$$l_n = -\log \frac{\exp(p_{y_n})}{\sum_{i=1}^{|\mathcal{C}|} \exp(p_i)} \tag{14}$$

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



(a) The number of different citation intents.

(b) The percentage of different citation intents.

Figure 3: The statistic of citation intent for all publications in the Semantic Scholar corpus. The temporal trends stay steady over time, suggesting a lack of information in the elapsed time from the time of publication to the time of citing.

Composite Model:

To further test the capabilities of our proposed model and use both structural and textual information, we devise a multi-modal model comprising encoders for both the graph structure and the citation context. Specifically, we use a pre-trained SciBERT model for encoding the citation phrase text and our MHLP model for encoding the citation graph around the citation link. Figure 2 illustrates an overview of the composite model.

5. Experiments

In this section, we report our experimental results on both of the SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$ datasets. All the graph-based experiments are carried out on the $\mathcal{G}_1$ KG. For the traditional KGE methods, we tune their hyperparameters as described in Appendix A.1 and train them using the hyperparameters showcased in Table 4. For the 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. For the MHLP-based methods, in both transductive and inductive settings, we use a 1-layer variation on top of the normalized features extracted as described in Section 4.3. Moreover, we tune their hyperparameters and train them as described in Appendix A.3. For the SciBERT method, we freeze the pre-trained model and add an MLP module on top of the 768-dimensional [CLS] token output. Similar to the other models, the MLP module is tuned using the parameters described in A.2. For the composite model, during the training phase, we freeze the SciBERT model in the first two epochs as a warm-up step for the graph encoder; then, we jointly train both encoders along with the final prediction module.

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 previously reported results [17] 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. All the models are implemented using PyTorch [14] and trained on a machine with a single Quadro RTX 8000 GPU, 72 CPU cores, and 768GB of RAM. Implementations are available under a CC-BY-SA license at TBD.

5.1. Results

Table 3 illustrates our experimental results on both 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. Interestingly, both ComplEx and RotatE perform worse than TransE on both datasets. This finding is surprising as both ComplEx and RotatE are more expressive than TransE [20]. However, when combined with MLP models, all exhibit significant performance boost, up to more than 100% in the case of RotatE. After this addition, we can see the same expressivity trend in the model results, i.e., the more powerful the model, the better the result. 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. Altogether, 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.

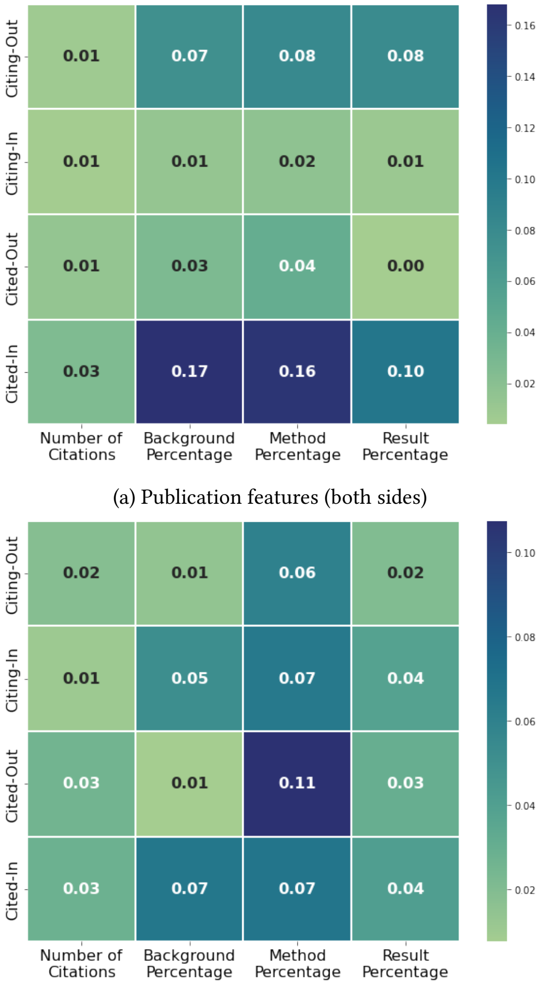
As for the MHLP method, in the transductive setting, it achieves 57.88 and 62.16 macro F1 scores on SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$ datasets, respectively. Moreover, its inductive results showcase the robustness of our approach in an extreme out-of-distribution setting, achieving 56.13 and 59.81 macro F1 scores. Compared to previously reported results [17], our model achieves superior performance to Glove-based models while slightly lagging behind Infersent-based models. Looking into the precision and recall comparison, our method has better precision scores on both transductive and inductive settings compared to all word embedding-based models; however, for recall, it performs better than Glove-based models and worse than the Infersent-based models which might stem from the imbalance in the links as illustrated by Figure 3a. Further experimentation to address the class imbalance problem in future works might help improve the overall performance of MHLP. 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.

Finally, the composite model denoted as *SciBERT + MHLP* in Table 3, achieves the best performance among all models, even beating the fine-tuned SciBERT. When considering MHLP's standalone performance, these results showcase the potential improvements that could be achieved through the use of structural information that is not available in citation phrases. The presented experiments are a stepping stone for better understanding and using the structural information at scale for citation intent classification.

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 3a and 3b 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 are based on the weakly labeled links that we obtained from S2ORC. Consequently, these links are generated by another noisy model that could potentially be biased. Hence, it should not discourage further analysis or studies of temporal information for citation intent classification.



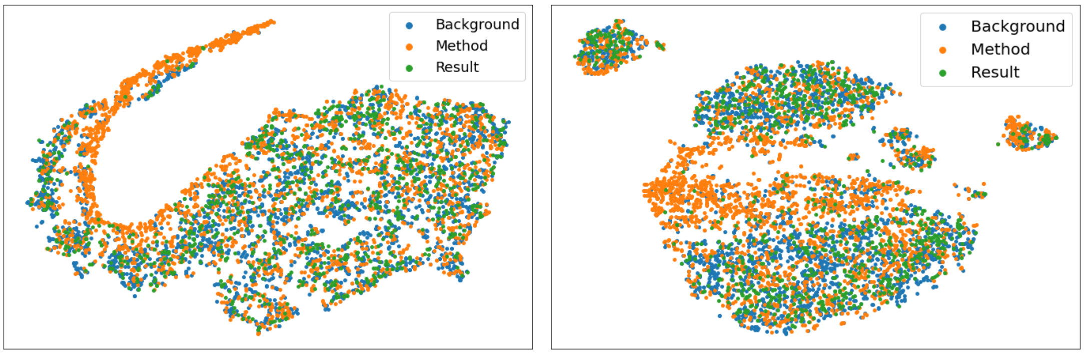
(a) Publication features (both sides)

(b) Averaged neighborhood features (both sides)

Figure 4: The calculated MI values for publication features and averaged neighborhood features. On average, the publication features show stronger connections to the target variable.

6.2. Mutual Information Analysis

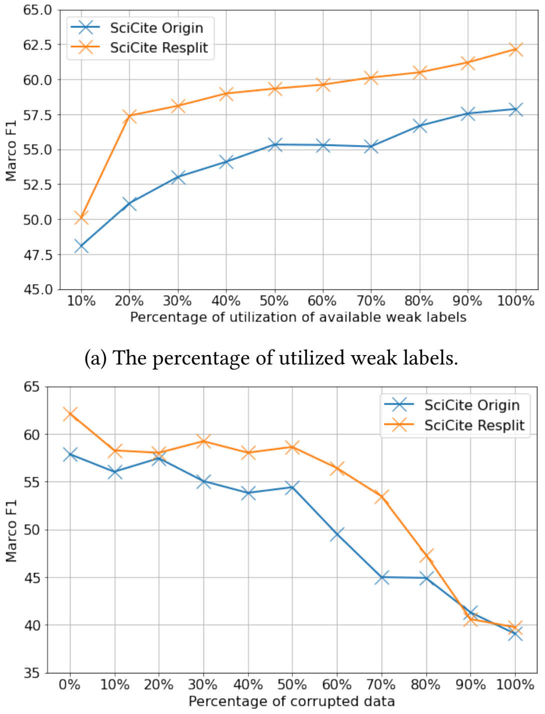
In this analysis, we study the quality of the engineered features as described in Section 4.3 concerning the weakly labeled intent classes. To this end, we use the well-known mutual information (MI) [12] measurement to quantify the importance of each feature. Formally, the MI between



(a) Features before normalization

(b) Features after normalization

Figure 5: The t-SNE visualizations for the unnormalized and normalized features.



(a) The percentage of utilized weak labels.

(b) The percentage of corrupted data.

Figure 6: The macro F1 score of MHLP (Transductive) on SciCite$_{\text{origin}}$ and SciCite$_{\text{resplit}}$ dataset

two discrete random variables X and Y is defined as

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

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 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 4a and 4b present the results of our experiments. As evident from these results, while the publication-averaged features generally show stronger connections to the target variable, the neighborhood-averaged features seem to show complementary connections, further emphasizing the importance of using both sets of features.

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 [24]. Figure 5a and 5b illustrate the projected space for the unnormalized and normalized features, respectively. As evident from Figure 5a, it is challenging to distinguish different intent types in the unnormalized space. However, after normalization, as evident from Figure 5b, 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.

6.4. Robustness Analysis

In this analysis, we focus on studying the robustness of our proposed graph-based method. To this end, we devise two ablation studies. In the first study, we randomly corrupt a percentage of the weak labels by replacing the correct label with a random label. This study aims to understand the model's resilience to noise better. In the second study, we randomly remove a percentage of the weak labels. This study's idea is to understand better the effect of weak supervision on the model's performance. These studies are carried out by running the MHLP method in the transductive setting on both $\text{SciCite}_{\text{origin}}$ and $\text{SciCite}_{\text{resplit}}$ datasets.

The feature vectors for the publications are calculated by counting the number of citations and intents. These vectors are normalized then using Equation 6 and 7. To analyze the relationship between the model's performance and the amount of available data, we create ten variations of the dataset by only using a portion of the available weak labels, varying from using all the available weak labels to only using 10% of them. Figure 6a presents the result of this study.

As evident from Figure 6a, the more weakly labeled links are available, the better our method performs. The other significant observation is the robustness of the model, even in the extreme scenario of having access to only 10% of the labels. Note that only 31.90% of links in the S2ROC have at least one weakly labeled intent, which means, even if the utilization percentage is 100%, only 31.90% citation links are weakly labeled.

Figure 6b showcases the relationship between the model performance and the percentage of corrupted data. Following our intuition, the model's performance monotonically decreases as we add more noisy labels to the data. However, two interesting observations could be made from this figure. First, the performance of our method only drops less than five macro F1 scores when half (50%) of the weak labels are replaced with randomly assigned noisy labels. This observation shows that the proposed method is exceptionally resilient when faced with mistakes. Second, even when all the labels are replaced with random ones (100%), the model performs better than the random baselines. This observation indicates that the model is learning to make inferences based on purely structural information, which further solidifies our hypothesis regarding the importance of structural information.

7. Conclusions and Future Work

In this work, we first introduced an expansion to the SciCite dataset by extracting scholarly information from the S2ORC dataset 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 experimental results in the inductive setting further showcase the robustness of the proposed approach in the information-deprived out-of-distribution environment. 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. Moreover, with a composite model comprising our method as the graph encoder and the state-of-the-art NLP-based model as the text encoder, we outperformed all the other models we experimented with. These results further signify the strong signal in relational information and highlight the importance of future analysis and studies in this domain. Finally, our presented analyses further support our methodological choices.

For future works, one straightforward idea is to extend the knowledge graph with more scholarly information, such as authors, venues, and fields of study. There already exist some open repositories such as OpenAlex [15] and Microsoft Academic Graph (MAG) [25] that contain this information. Another direction is further investigation into the temporal signals. Last but not least, although we achieved an improved performance through a fusion of textual and structural information, more investigation and analysis could be done in this setting in future works.

Acknowledgments

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

[1] 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

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko.

2013. Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 2787–2795.

[3] Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? A review of studies on citing behavior. *J. Documentation* 64 (2008), 45–80.

[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

[5] Arman Cohan and Nazli Goharian. 2015. Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 390–400. https://doi.org/10.18653/v1/D15-1045

[6] Daniel Cummings and Marcel Nassar. 2020. Structured Citation Trend Prediction Using Graph Neural Networks.. In *ICASSP*. IEEE, Barcelona, Spain, 3897–3901. http://dblp.uni-trier.de/db/conf/icassp/icassp2020.html#CummingsN20

[7] M.A. Garzone. 1997. *Automated Classification of Citations Using Linguistic Semantic Grammars*. Thesis (M.Sc.)–University of Western Ontario, London, Canada. https://books.google.com/books?id=V-bwSgAACAAJ

[8] 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* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

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

[10] 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

[11] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR '17)*. OpenReview.net, Palais des Congrès Neptune, Toulon, France, 14 pages. https://openreview.net/forum?id=SJU4ayYgl

[12] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

[13] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in PyTorch. In *NIPS 2017 Workshop on Autodiff*. OpenReview.net, Long Beach, California, USA, 4 pages. https://openreview.net/forum?id=BJJsrmfCZ

[15] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* abs/2205.01833 (2022), 5 pages.

[16] Anna Ritchie. 2009. *Citation context analysis for information retrieval*. Technical Report. University of Cambridge, Computer Laboratory.

[17] Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubaa, and Lisu Yu. 2021. Citation intent classification using word embedding. *Ieee Access* 9 (2021), 9982–9995.

[18] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*, Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 593–607.

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

[20] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space.. In *ICLR (Poster)*. OpenReview.net, New Orleans, LA, 18 pages. http://dblp.uni-trier.de/db/conf/iclr/iclr2019.html#SunDNT19

[21] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop*

on Discourse and Dialogue. Association for Computational Linguistics, Sydney, Australia, 80–87. https://aclanthology.org/W06-1312

[22] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, New York, NY, USA, 2071–2080.

[23] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying Meaningful Citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop (Technical Report, WS-15-13)*, Cornelia Caragea, C. Lee Giles, Narayan Bhamidipati, Doina Caragea, Sujatha Das Gollapalli, Saurabh Kataria, Huan Liu, and Feng Xia (Eds.). AAAI Press, Menlo Park, CA, 21–26. http://www.aaai.org/Library/Workshops/ws15-13.php

[24] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[25] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.

[26] Wenhao Yu, Mengxia Yu, Tong Zhao, and Meng Jiang. 2020. Identifying Referential Intention with Heterogeneous Contexts. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 962–972. https://doi.org/10.1145/3366423.3380175

[27] Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 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)*. Association for Computing Machinery, New York, NY, USA, 739–748.

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 α and γ are the adversarial temperature and the margin value (RotatE-only), respectively.

Table 4
Hyperparameters of KGE algorithms.

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
α	0	0.25	1
γ	-	-	6

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.

A.3. Multi-Hop Link Prediction

We run a grid search over the following ranges: *number of layers* $\in \{0, 1, 2, 3\}$, *dimension* $\in \{10, 50, 100, 200\}$, *learning rate* $\in \{0.03, 0.01, 0.003, 0.001\}$, The optimal hyperparameters are as follows: *number of layers* $= 1$, *dimension* $= 100$, *learning rate* $= 0.01$. We use Adam as the optimizer through the tuning process.

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 α and γ are the adversarial temperature and the margin value (RotatE-only), respectively.