SymTax: Symbiotic Relationship and Taxonomy Fusion for Effective Citation Recommendation

Karan Goyal
IIIT Delhi, India
karang@iiitd.ac.in

Mayank Goel
NSUT Delhi, India
mayank.co19@nsut.ac.in

Vikram Goyal
IIIT Delhi, India
vikram@iiitd.ac.in

Mukesh Mohania
IIIT Delhi, India
mukesh@iiitd.ac.in

Abstract

Citing pertinent literature is pivotal to writing and reviewing a scientific document. Existing techniques mainly focus on the local context or the global context for recommending citations but fail to consider the actual human citation behaviour. We propose SymTax, a three-stage recommendation architecture that considers both the local and the global context, and additionally the taxonomical representations of query-candidate tuples and the *Symbiosis* prevailing amongst them. SymTax learns to embed the infused taxonomies in the hyperbolic space and uses hyperbolic separation as a latent feature to compute query-candidate similarity. We build a novel and large dataset ArSyTa containing 8.27 million citation contexts and describe the creation process in detail. We conduct extensive experiments and ablation studies to demonstrate the effectiveness and design choice of each module in our framework. Also, combinatorial analysis from our experiments shed light on the choice of language models (LMs) and fusion embedding, and the inclusion of section heading as a signal. Our proposed module that captures the symbiotic relationship solely leads to performance gains of 26.66% and 39.25% in Recall@5 w.r.t. SOTA on ACL-200 and RefSeer datasets, respectively. The complete framework yields a gain of 22.56% in Recall@5 wrt SOTA on our proposed dataset. The code and dataset are available at https://github.com/goyalkaraniit/SymTax.

1 Introduction

Citing has always been the backbone of scientific research. It enables trust and supports the claims made in the scientific document. The ever-growing increase in the amount of scientific literature makes it imperative to ease out the author's task of finding a list of suitable papers to follow and cite (Johnson et al., 2018; Bornmann et al., 2021; Nane et al., 2023). Citation recommendation is such a process that helps researchers to be aware of the relevant research in respective domains. There are two different approaches to recommend citations: *local* (Dai et al., 2019; Ebesu and Fang, 2017; Huang et al., 2012; He et al., 2010), and *global* (Xie et al., 2021; Ali et al., 2021; Bhagavatula et al., 2018; Guo et al., 2017). Local citation recommendation is the task of finding and recommending the most relevant prior work, mainly corresponding to a specific text passage (also known as citation context), making it context-aware. On the other hand, global citation recommendation recommends a list of suitable prior art for the entire document, mainly given the title and abstract or the whole document. In this paper, we solve the task of local citation recommendation, which is more fine-grained and provides a solution to the actual challenge the author faces. For example, consider the below citation excerpt:[1]

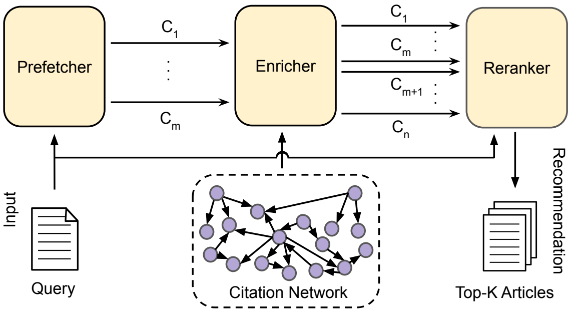

Figure 1: Proposed method consists of three essential modules. *Prefetcher* and *Reranker* takes query consisting of citation context, title, abstract and taxonomy of the citing paper as input. For each candidate paper (C_i), *Enricher* uses knowledge from citation network and *Reranker* generates the final top-K recommendations.

This can have extreme consequences in real-life scenarios such as autonomous cars CitX.

Examining the above context in isolation makes it challenging to predict the specific article cited

[1] Excerpt is borrowed from *Towards Consistency in Adversarial Classification* of (Meunier et al., 2022). The cited article is *An analysis of adversarial attacks and defenses on autonomous driving models* of (Deng et al., 2020).

Findings of the Association for Computational Linguistics: ACL 2024, pages 8997–9008
August 11-16, 2024 ©2024 Association for Computational Linguistics

at CitX. However, leveraging global information such as title, abstract, and taxonomy narrows down the search space while at the same time utilizing symbiotic relationship provides the model with an enriched pool of the most suitable candidates. Unlike ACL-200 and RefSeer datasets with curated contexts of fixed size, we curate richer contexts by incorporating complete information of adjoining sentences with respect to the citation sentence. To summarise, we make the following contributions:

- *Dataset*: We have constructed a dataset ArSyTa comprising 8.27 million comprehensive citation contexts across diverse domains, featuring richer density and relevant features, including taxonomy concepts, to facilitate the task of citation recommendation.
- *Conceptual*: We explore the concept of *Symbiosis* from Biology and draw its analogy with human citation behaviour in the scientific research ecosystem and select a better pool of candidates.
- *Methodological*: We propose a novel taxonomy fused reranker that subsequently learns projections of fused taxonomies in hyperbolic space and utilises hyperbolic separation as a latent feature.
- *Empirical*: We perform extensive experiments, ablations, and analysis on five datasets and six metrics, demonstrating SymTax consistently outperforms SOTA by huge margins.

2 Related Work

Local citation recommendation has drawn comparatively less interest than its global counterpart until recently. He et al. (2010) introduced the task of local citation recommendation by using tf-idf based vector similarity between context and cited articles. Livne et al. (2014) extracted hand-crafted features from the citation excerpt and the remaining document text, and developed a system to recommend citations while the document is being drafted. The neural probabilistic model of (Huang et al., 2015) determines the citation probability for a given context by jointly embedding context and all articles in a shared embedding space. Ebesu and Fang (2017) proposed neural citation network based on encoder-decoder architecture. The encoder obtains a robust representation of citation context and further augments it via author networks and attention mechanism, which the decoder uses to generate the title of the cited paper. Dai et al. (2019) utilised stacked denoising autoencoders for representing cited articles, bidirectional LSTMs for citation context representation and attention principle over citation context to enhance the learning ability of their framework.

Jeong et al. (2020) proposed a BERT-GCN model which uses BERT (Kenton and Toutanova, 2019) to obtain embeddings for context sentences, and Graph Convolutional Network (Kipf and Welling, 2017) to derive embeddings from citation graph nodes. The two embeddings are then concatenated and passed through a feedforward neural network to obtain relevance between them. However, due to the high cost of computing GCN, as mentioned in Gu et al. (2022), BERT-GCN model was evaluated on tiny datasets containing merely a few thousand citation contexts. It highlights the limitation of scaling such GNN models for recommending citations on large datasets.

Medić and Šnajder (2020) suggested the use of global information of articles along with citation context to recommend citations. It computes semantic matching score between citation context and cited article text, and bibliographic score from the article's popularity in the community to generate a final recommendation score. Ostendorff et al. (2022) perform neighbourhood contrastive learning over the full citation graph to yield citation embeddings and then uses k-nearest neighbourhood based indexing to retrieve the top recommendations. The most recent work in local citation recommendation by Gu et al. (2022) proposed a two-stage recommendation architecture comprising a fast prefetching module and a slow reranking module. We build upon work of Gu et al. (2022) by borrowing their prefetching module and designing a novel reranking module and another novel module named Enricher that fits between Prefetcher and Reranker. We name our model as SymTax (Symbiotic Relationship and Taxonomy Fusion).

3 Proposed Dataset

Motivation. Citation recommendation algorithms depend on the availability of the labelled data for training. However, curating such a dataset is challenging as full pdf papers must be parsed to extract citation excerpts and map the respective cited articles. Further, the constraint that cited articles should be present in the corpus eliminates

Dataset	# Contexts				# Papers	LCC	Deg	Pub Years
	Train	Val	Test	Total				
ACL-200	30,390	9,381	9,585	49,356	19,776	0.035	3.42	2009-2015
FTPR	9,363	492	6,814	16,669	4,837	0.036	2.84	2007-2017
RefSeer	3,521,582	124,911	126,593	3,773,086	624,957	0.033	4.46	-2014
arXiv	2,988,030	112,779	104,401	3,205,210	1,661,201	0.027	3.30	1991-2020
ArSyTa	8,030,837	124,188	124,189	**8,279,214**	474,341	**0.051**	**9.98**	**2007-2023**

Table 1: Statistics across various datasets indicate the largest, densest and most recent nature of our dataset, `ArSyTa`. FTPR is `FullTextPeerRead`, arXiv is `arXiv(HAtten)`, and LCC and Deg are the average local clustering coefficient and average degree of the citation context network, respectively.

a large proportion of it, thus reducing the dataset size considerably. e.g. `FullTextPeerRead` (Jeong et al., 2020) and `ACL-200` (Medić and Šnajder, 2020) datasets contain only a few thousand papers and contexts. `RefSeer` (Medić and Šnajder, 2020) contains 0.6 million papers published till 2014 and hence is not up-to-date. Gu et al. (2022) released a large and recent arXiv-based dataset (we refer to it as `arXiv(HAtten)`) by following the same strategy adopted by `ACL-200` and `FullTextPeerRead` for extracting contexts. They consider 200 characters around the citation marker as the citation context. The above mentioned datasets have limited features, which may restrict the design of new algorithms for local citation recommendation. Thus, we propose a novel dataset `ArSyTa`[2] which is latest, largest and contains rich citation contexts with additional features.

Dataset Creation. We selected $475,170$ papers belonging to Computer Science (CS) categories from over 1.7 million scholarly papers spanning STEM disciplines available on arXiv. The papers are selected from April 2007-January 2023 publication dates to ensure current relevance. arXiv contains an extensive collection of scientific papers that offer innate diversity in different formatting styles, templates and written characterisation, posing a significant challenge in parsing pdfs. We comprehensively evaluate established frameworks, namely, arXiv Vanity[3], CERMINE[4], and GROBID[5], for data extraction. arXiv Vanity converts pdfs to HTML format for data extraction but produces inconsistent results, thus turning extraction infeasible in this scenario. CERMINE uses JAVA binaries to generate BibTeX format from pdf but fails to extract many references, thereby not providing the required level of information. GROBID is a state-of-the-art tool that accurately and efficiently produces easy-to-parse results in XML format with a standard syntax. We conduct extensive manual testing to assess parsing efficacy and finally choose GROBID as it adeptly parses more than 99.99% (i.e., $474,341$) of the documents. We organise the constructed dataset into a directed graph. Nodes within the graph encapsulate a rich array of attributes, encompassing abstracts, titles, authors, submitters, publication dates, topics, categories within CS, and comments associated with each paper. Edges within graph symbolise citations, carrying citation contexts and section headings in which they appear. This provides a format that offers better visualisation and utilisation of data.

Unlike previously available datasets, which use a 200-character length window to extract citation context, we consider one sentence before and after the citation sentence as a complete citation context. We create a robust mapping function for efficient data retrieval. Since every citation does not contain a Digital Object Identifier, mapping citations to corresponding papers is challenging. The use of several citation formats and the grammatical errors adds a further challenge to the task. To expedite title-based searches that associate titles with unique paper IDs, we devise an approximate mapping function based on LCS (Longest Common Substring), but the sheer size of the number of papers makes it infeasible to run directly, as each query requires around 10 seconds. Finally, to identify potential matches, we employ an approximate hash function called MinHash LSH (Locality Sensitivity Hashing), which provides the top 100 candidates with a high probability for a citation existing in our raw database to be present in the candidate list. We then utilise LCS matching with a 0.9 similarity score threshold to give a final candidate, thus reducing the time to a few microseconds. Finally, our dataset consists of 8.27 million citation contexts whereas the largest existing dataset, RefSeer, consists of only 3.7 million contexts. The dataset is essentially comprised of contexts and the corresponding metadata only and not the research papers, as is the case with other datasets. Even after considering a relatively lesser number of papers as a raw source, we curated significantly more citation contexts (i.e., final data), thus showing the effectiveness of our data extraction technique. This is further supported

[2] ArSyTa: Arxiv Symbiotic Relationship Taxonomy Fusion

[3] https://github.com/arxiv-vanity/arxiv-vanity

[4] https://github.com/CeON/CERMINE

[5] https://github.com/kermitt2/grobid_client_python

empirically by the fact that our dataset has significantly higher values of average local clustering coefficient and average degree with respect to the other datasets (as shown in Table 1). Each citing paper and cited paper that corresponds to a citation context respectively belongs to a CS concept in the flat-level arXiv taxonomy that contains 40 classes. The distribution of category classes in arXiv taxonomy for ArSyTa is shown in Figure 3 (Appendix).

Technical Merits. ArSyTa offers the following merits over the existing datasets: (i) As shown in Table 1, ArSyTa is 2.2x and 2.6x larger than RefSeer and arXiv(HAtten), respectively. Also, our citation context network is more dense than all other datasets, clearly showing that our dataset creation strategy is better. (ii) It is the most recent dataset that contains papers till January 2023. (iii) It contains longer citation contexts and additional signals such as section heading and document category. (iv) ArSyTa is suitable for additional scientific document processing tasks that can leverage section heading as a feature or a label. (v) ArSyTa is more challenging than others as it contains papers from different publication venues with varied formats and styles submitted to arXiv.

4 SymTax Model

We discuss the detailed architecture of our proposed model – SymTax, as shown in Figure 2. It comprises a fast prefetching module, an enriching module and a slow and precise reranking module. We borrow an existing prefetching module from Gu et al. (2022) whereas an enriching module and a reranking module are our novel contributions in the overall recommendation technique. The subsequent subsections elaborate on the architectures of these three modules.

4.1 Prefetcher

The task of the prefetching module is to provide an initial set of high-ranking candidates by scoring all the papers in the database with respect to the query context. It uses cosine similarity between query embedding and document embedding to estimate the relevance between query context and the candidate document. Prefetcher comprises two submodules, namely, Paragraph Encoder and Document Encoder. Paragraph Encoder computes the embedding of a given paragraph, i.e. title, abstract or citation context, using a transformer layer followed by multi-head pooling. Document Encoder takes paragraph encodings as input along with paragraph types and passes them through a multi-head pooled transformer layer to obtain the final document embedding. We adopt the prefetching module from Gu et al. (2022) and use it as a plugin in our overall recommendation technique. For brevity, we refer readers to follow the source to understand the detailed working of the prefetcher.

4.2 Enricher

In general, prefetching is often followed by reranking in recommendation systems. Let $C_L = \{c_1, c_2, ..., c_m\}$ be the candidate list generated by prefetcher, and $G = (V, E)$ be the given citation network $s.t.\ c_i \in V\ \forall\, i \in \{1, 2, ..., m\}$.

The nodes in citation network represent papers and the directed edges represent citation relationship. Here, we propose a new module Enricher that augments the candidate list generated by the prefetcher and outputs an enriched candidate list. It generates an ego network for every article in the candidate list using the citation graph and excludes the incoming neighbours and the duplicates from the expanded network list. For every node $u \in V$, we define $N_o(u) \subset V$ as the ego network of u containing only the outgoing neighbours. Let E_L denotes the enriched list, then

$$\begin{aligned} E_L =& \{c_1, c_2, ..., c_m, N_o(c_1), N_o(c_2), ..., N_o(c_m)\} \\ E_L =& \{c_1, c_2, ..., c_m, c_{m+1}, c_{m+2}, ..., c_n\} \end{aligned} \quad (1)$$

where $\{\}$ represents a set operator. We then feed this enriched list as input to the reranker. The design notion of Enricher is inspired by Symbiosis, aka Symbiotic Relationship, a concept in Biology.

Symbiosis. The idea of including cited papers of identified candidates has been pursued in the literature (Cohan et al., 2020) but from the perspective of hard negatives. To the best of our knowledge, the concept of Enrichment has never been discussed earlier for citation recommendation to model the human citation behaviour. We identify two different types of citation behaviours that prevail in the citation ecosystem and draw a corresponding analogy with *mutualism* and *parasitism* that falls under the concept of *Symbiosis*. *Symbiosis* is a long-term relationship or interaction between two dissimilar organisms in a habitat. In our work, the habitat is the citation ecosystem, and the two dissimilar organisms are the candidate article and its neighbourhood. We try to explain the citation phenomena

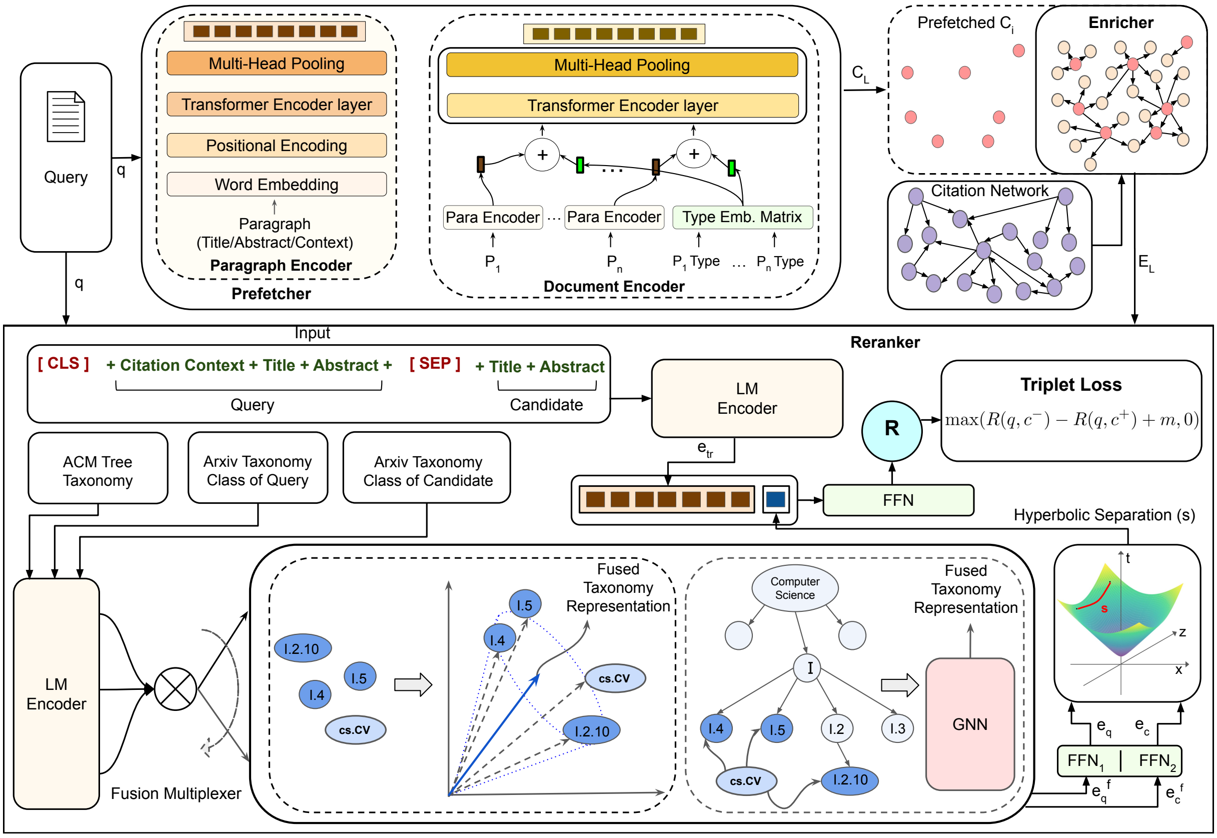

Figure 2: Architecture of SymTax. It consists of three essential modules – (a) Prefetcher, (b) Enricher, and (c) Reranker. The task of Enricher is to enrich the candidate list generated by Prefetcher and provide it as an input to Reranker. Reranker utilises taxonomy fusion and hyperbolic separation to yield final recommendation score (R). Mapping:- I.4: Image Processing and Computer Vision, I.5: Pattern Recognition, I.2.10: Vision and Scene Understanding, cs.CV: Computer Vision. Fusion Multiplexer enables switching between vector-based and graph-based taxonomy fusion. We have released the mapping config file along with the data.

through *Symbiosis* wherein the candidate and its neighbourhood either play the role of *mutualism* or *parasitism*. In *mutualism*, the query paper recommends either only the candidate paper under consideration or both the considered candidate paper and from its 1-hop outdegree neighbour network. On the other hand, in *parasitism*, the neighbour organism feeds upon the candidate to get itself cited, i.e., the query paper, rather than citing the candidate article, in turn, recommends from its outgoing edge neighbours. This whole idea, in practice, is analogous to human citation behaviour. When writing a research article, researchers often gather a few highly relevant prior art and cite highly from their references. We can interpret this tendency as a slight human bias or highly as utilising the research crowd's wisdom. Owing to this, Enricher is only required at the inference stage. Nevertheless, it is a significantly important signal, as evident from the results in Table 2 and Table 3.

4.3 Reranker

The purpose of the reranker is to rerank the candidates fetched from previous modules with higher accuracy. Therefore, the reranker is generally slower than the prefetcher. It makes a fine-grain comparison between the query and each candidate, calculates a recommendation score and returns it as the final output. Our proposed reranker considers the relevance between query and candidate at two separate levels, namely (i) text relevance and (ii) taxonomy relevance. In text relevance, we concatenate the query text and candidate text with a [SEP] token and pass it through a Language Model (LM) to obtain the joint embedding $e_{tr} \in \mathbb{R}^d$. Query text is the concatenation of citation context, title and abstract of the query article, whereas the candidate text is the concatenation of title and abstract of candidate paper. In taxonomy relevance, we perform taxonomy fusion and hyperbolic projections to calculate separation between query and candidate.

Taxonomy Fusion. The inclusion of taxonomy fusion is an important and careful design choice. Intuitively, a flat-level taxonomy (arXiv concepts) does not have a rich semantic structure in comparison to a hierarchically structured taxonomy like ACM. In a hierarchical taxonomy, we have a semantic relationship in terms of generalisation, specialisation and containment. Mapping the flat concepts into hierarchical taxonomy infuses a structure into the flat taxonomy. It also enriches the hierarchical taxonomy as we get equivalent concepts from the flat taxonomy. Each article in our proposed dataset `ArSyTa` consists of a feature category that represents the arXiv taxonomy[6] class it belongs to. Since ArSyTa contains papers from the CS domain, so we have a flat arXiv taxonomy. e.g. cs.LG and cs.CV represents Machine Learning and Computer Vision classes, respectively. We now propose the fusion of flat-level arXiv taxonomy with ACM tree taxonomy[7] to obtain rich feature representations for the category classes. We mainly utilise the subject class mapping information mentioned in the arXiv taxonomy and domain knowledge to create a class taxonomy mapping from arXiv to ACM. e.g. cs.CV is mapped to ACM classes I.2.10, I.4 and I.5 (as shown in Fig. 2). Also, we release the mapping config file in the data release phase. We employ two fusion strategies, namely vector-based and graph-based. In vector-based fusion, the classes are passed through LM and their conical vector is obtained by averaging out class vectors in feature space. In graph-based fusion, we first form a graph by injecting arXiv classes into the ACM tree and creating directed edges between them. We initialise node embeddings using LM and run Graph Neural Network (GNN) algorithm to learn fused representations. We consider GAT(Veličković et al., 2018) and APPNP(Gasteiger et al., 2019) as GNN algorithms and observe their performance as the same. The final representations of cs.{} nodes represent the fused representations learnt. Empirically, we can clearly observe that the fusion of concepts helps to attain significant performance gains (as shown in Table 3).

Hyperbolic Separation. Hyperbolic spaces have recently gained momentum in deep learning due to their high capacity and tree-likeliness properties, thus making them suitable for learning better representations for hierarchical data (Ganea et al., 2018). Motivated by the works of Sawhney et al. (2022) and Ganea et al. (2018), and the realisation that Euclidean space cannot fully capture the complex characteristics of hierarchical data, we use hyperbolic distance as our metric to compute taxonomy relevance. Let q and c denote query and candidate, respectively, where $c \in E_L$. Let ${e_q}^f$, ${e_c}^f \in \mathbb{R}^D$ represents the fusion embeddings for q and c respectively. We pass fusion embeddings through a feed forward network $h_\theta(.)$ and then compute hyperbolic separation between query and candidate to project the embeddings into hyperbolic space. So, we obtain the following representations

$$e_q = h_\theta({e_q}^f) \in \mathbb{R}^d; e_c = h_\theta({e_c}^f) \in \mathbb{R}^d \tag{2}$$

and then compute the hyperbolic separation s between e_q and e_c as follows

$$s(e_q, e_c) = 2\tan^{-1}\left(\|(-e_c) \oplus e_q\|\right) \tag{3}$$

where $\oplus$ represents Möbius addition and for a pair of points $a, b \in \mathcal{B}$, is defined as,

$$a \oplus b := \frac{(1 + 2\langle a, b\rangle + \|b\|^2)a + (1 - \|a\|^2)b}{1 + 2\langle a, b\rangle + \|a\|^2\|b\|^2}$$

where $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ are Euclidean inner product and norm.

Final Recommendation. We use this hyperbolic separation representing the taxonomy relevance as a latent feature and concatenate ($\odot$) it with the text relevance embedding e_{tr}. This step ensures that category classes with similar concepts in the taxonomy learn to embed themselves closely in the manifold. We then pass this relevance embedding through a feed forward network ($g_\theta(.)$) and apply the sigmoid activation function ($\sigma(.)$) to get the final relevance score, defined as,

$$R = \sigma(g_\theta(e_{tr} \odot s)) \in (0, 1) \tag{4}$$

To interpret R as the final recommendation score, we employ and minimise the Triplet loss, L:

$$L = \max(R(q, c^-) - R(q, c^+) + m, 0) \tag{5}$$

where m is margin, and c^+ and c^- are positive and negative candidates respectively. We adopt a simple technique for mining triplets for a query. We choose cited paper as the positive candidate and randomly select papers from the candidate list as negative candidates.

[6] https://arxiv.org/category_taxonomy

[7] https://www.acm.org/publications/computing-classification-system/1998/ccs98

5 Experiments and Results

This section illustrates the various baselines, evaluation metrics and datasets used to benchmark our proposed method followed by the performance comparison.

Baselines. We consider evaluating various available systems for comparison. **BM25** (Robertson et al., 2009): It is a prominent ranking algorithm, and we consider its several available implementations and choose Elastic Search implementation[8] as it gives the best performance with the highest speed. **SciNCL** (Ostendorff et al., 2022): We use its official implementation available on GitHub[9]. **HAtten** (Gu et al., 2022): We use its official implementation available on GitHub[10]. **NCN** (Ebesu and Fang, 2017) could have been a potential baseline; however, as reported by Medić and Šnajder (2020), the results mentioned could not be replicated. **DualLCR** (Medić and Šnajder, 2020): It is essentially a ranking method that requires a small and already existing list of candidates containing the ground truth, which turns it into an artificial setup that, in reality, does not exist. This unfair setup is also reported by Gu et al. (2022), which is state-of-the-art in our task. Thus for a fair comparison, we could not consider it in comparing our final results.

Evaluation Metrics. To stay consistent with the literature that uses Recall@10 and Mean Reciprocal Rank (MRR) as the evaluation metrics, we additionally use Normalised Discounted Cumulative Gain (NDCG@10) and Recall@K for different values of K to obtain more insights from the recommendation performance. Recall@K measures the percentage of cited papers appearing in top-K recommendations. MRR measures the reciprocal rank of the cited paper among the recommended candidates. NDCG takes into account the relative order of recommendations in the ranked list. The above metrics are averaged over all test queries, and higher values indicate better performance.

Model	R@5	R@10	R@20	R@50	NDCG	MRR
			ACL-200			
BM25	0.1374	0.1939	0.2531	0.3486	0.0808	0.1074
SciNCL	0.1517	0.2250	0.3176	0.4467	0.1044	0.0669
HAtten	0.4186	0.4997	0.5579	0.5962	0.3002	0.2362
SymTax (SpecG)	0.4529	0.5897	0.7038	0.8396	0.3034	0.2126
SymTax (SciV)	**0.5302**	**0.6529**	**0.7640**	**0.8803**	**0.3818**	**0.2955**
			FullTextPeerRead			
BM25	0.2688	0.3371	0.4076	0.5168	0.1750	0.2136
SciNCL	0.2173	0.3104	0.4217	0.5722	0.1452	0.0935
HAtten	0.5027	0.5788	0.6263	0.6514	0.3566	0.2847
SymTax (SpecG)	0.4611	0.6266	0.7619	0.8899	0.3087	0.2090
SymTax (SciV)	**0.6216**	**0.7505**	**0.8398**	**0.9294**	**0.4472**	**0.3500**
			RefSeer			
BM25	0.1737	0.2192	0.2677	0.3365	0.1185	0.1424
SciNCL	0.0967	0.1486	0.2114	0.3089	0.0700	0.0450
HAtten	0.2672	0.3374	0.3985	0.4637	0.1925	0.1466
SymTax (SpecG)	0.2724	0.3831	0.4960	0.6512	0.1942	0.1353
SymTax (SciV)	**0.3721**	**0.4845**	**0.5916**	**0.7264**	**0.2676**	**0.1993**
			arXiv(HAtten)			
BM25	0.1529	0.1973	0.2455	0.3160	0.1019	0.1245
SciNCL	0.1076	0.1604	0.2227	0.3192	0.0737	0.0468
HAtten	0.2426	0.3292	0.4097	0.4949	0.1651	0.1136
SymTax (SpecG)	**0.2905**	**0.4095**	0.5308	**0.6992**	**0.1983**	**0.1323**
SymTax (SciV)	0.2817	0.3997	**0.5317**	0.6987	0.1928	0.1284
			ArSyTa			
BM25	0.1777	0.2203	0.2640	0.3269	0.1155	0.1006
SciNCL	0.1612	0.2155	0.2757	0.3624	0.1088	0.0751
HAtten	0.1567	0.2046	0.2522	0.3070	0.1074	0.0766
SymTax (SpecG)	0.2061	0.2747	0.3499	0.4668	0.1421	0.1003
SymTax (SciV)	**0.2178**	**0.2976**	**0.3808**	**0.5029**	**0.1486**	**0.1018**

Table 2: Results clearly show that SymTax consistently outperforms SOTA (HAtten) across datasets on all metrics. Best results are highlighted in bold. Abbreviation: SpecG:- SPECTER_Graph; SciV:- SciBERT_Vector; R:- Recall.

Performance Comparison. As evident from Table 2, our evaluation shows the superior performance of SymTax on all metrics across all the datasets. We consider two different variants of SymTax in our main results comparison (i) SpecG: with SPECTER (Cohan et al., 2020) as LM and graph-based taxonomy fusion, and (ii) SciV: with SciBERT (Beltagy et al., 2019) as LM and vector-based taxonomy fusion. SPECTER and SciBERT are two state-of-the-art LMs trained on scientific text. SciV performs as the best model on ACL-200, FullTextPeerRead, RefSeer and ArSyTa on all metrics. SpecG performs best on arXiv(HAtten) on all metrics and results in a marginally less R@20 score than SciV. We observe the highest scores on FullTextPeerRead followed by ACL-200. It is due to the fact that these datasets lack diversity to a large extent. e.g. FullTextPeerRead is extracted from papers belonging to Artificial Intelligence field, and ACL-200 contains papers published at ACL venues. In contrast, we observe the lowest scores on ArSyTa followed by arXiv(HAtten). The common reason driving these performance trends is that both of these arXiv-based datasets contain articles from different publication venues with various formats, styles and domain areas, making the learning difficult and recommendation challenging. Our reasoning is further supported by the fact that ArSyTa is the latest dataset, and thus

[8] https://github.com/kwang2049/easy-elasticsearch

[9] https://github.com/malteos/scincl

[10] https://tinyurl.com/yckhe7d6

SymTax Variant	R@5	R@10	R@20	R@50	NDCG	MRR
SciBERT_vector	0.2178	0.2976	0.3808	0.5029	0.1486	0.1018
– Symbiosis	0.1794	0.2234	0.2651	0.3105	0.1194	0.0860
– Taxonomy	0.1614	0.2377	0.3215	0.4611	0.1162	0.0787
– Hyperbolic	0.1905	0.2678	0.3507	0.4719	0.1316	0.0891
SPECTER_graph	0.2061	0.2747	0.3499	0.4668	0.1421	0.1003
– Symbiosis	0.1749	0.2178	0.2598	0.3079	0.1181	0.0862
– Taxonomy	0.1795	0.2507	0.3384	0.4733	0.1263	0.0874
– Hyperbolic	0.2028	0.2669	0.3444	0.4641	0.1386	0.0981

Table 3: Ablation shows importance of *Symbiosis*, taxonomy fusion and hyperbolic space on ArSyTa. Excluding Symbiosis reduces the metrics more as compared to the exclusion of taxonomy and hyperbolic space.

SymTax Variant	R@5	R@10	R@20	R@50	NDCG	MRR
SciBERT_graph	0.1589	0.2421	0.3310	0.4607	0.1116	0.0715
SPECTER_vector	0.1552	0.2324	0.3252	0.4607	0.1092	0.0712
SPECTER_graph	*0.2025*	*0.2667*	*0.3456*	*0.4654*	*0.1385*	**0.0981**
SciBERT_vector	**0.2097**	**0.2876**	**0.3754**	**0.4939**	**0.1432**	*0.0978*

Table 4: Analysis on choice of LM and taxonomy fusion on 10k random samples from ArSyTa. Best results are highlighted in bold and second best are italicised.

contains the maximum amount of diverse samples and is shown to be the toughest dataset for recommending citations. To summarise, we obtain performance gains in Recall@5 of 26.66%, 23.65%, 39.25%, 19.74%, 22.56% with respect to SOTA on ACL-200, FullTextPeerRead, RefSeer, arXiv(HAtten) and ArSyTa respectively. The results show that NDCG is a tough metric compared to the commonly used Recall, as it accounts for the relative order of recommendations. Since the taxonomy class attribute is only available for our proposed dataset, we intentionally designed SymTax to be highly modular for better generalisation, as evident in Table 2.

6 Analysis

We conduct extensive analysis to assess further the modularity of SymTax, the importance of different modules, combinatorial choice of LM and taxonomy fusion, and the usage of hyperbolic space over Euclidean space. Furthermore, we analysed the effect of using section heading as an additional signal (shown in Appendix A).

6.1 Ablation Study

We perform an ablation study to highlight the importance of *Symbiosis*, taxonomy fusion and hyperbolic space. We consider two variants of SymTax, namely SciBERT_vector and SPECTER_graph. For each of these two variants, we further conduct three experiments by (i) removing the Enricher module that works on the principle of *Symbiosis*, (ii) not considering the taxonomy attribute associated with the citation context and (iii) using Euclidean space to calculate the separation score.

As evident from Table 3, *Symbiosis* exclusion results in a drop of 21.40% and 24.45% in Recall@5 and NDCG respectively for SciBERT_vector whereas for SPECTER_graph, it leads to a drop of 17.84% and 20.32% in Recall@5 and NDCG respectively. Similarly, taxonomy exclusion results in a drop of 34.94% and 27.88% in Recall@5 and NDCG respectively for SciBERT_vector whereas for SPECTER_graph, it leads to a drop of 14.81% and 12.51% in Recall@5 and NDCG respectively. It is clear from Table 3 that the use of Euclidean space instead of hyperbolic space leads to performance drop across all metrics in both variants. Exclusion of *Symbiosis* impacts higher recall metrics more in comparison to excluding taxonomy fusion and hyperbolic space.

6.2 Quantitative Analysis

We consider two available LMs, i.e. SciBERT and SPECTER, and the two types of taxonomy fusion, i.e. graph-based and vector-based. This results in four variants, as shown in Table 4. As evident from the results, SciBERT_vector and SPECTER_graph are the best-performing variants. So, the combinatorial choice of LM and taxonomy fusion plays a vital role in model performance. The above observations can be attributed to SciBERT being a LM trained on plain scientific text. In contrast, SPECTER is a LM trained with Triplet loss using 1-hop neighbours of the positive sample from the citation graph as hard negative samples. So, SPECTER embodies graph information inside itself, whereas SciBERT does not.

6.3 Qualitative Analysis

We assess the quality of recommendations given by different algorithms by randomly choosing an example. Though random, we choose the example that has multiple citations in a given context so that we can present the qualitative analysis well by investigating the top-10 ranked predictions. As shown in Table 5, we consider an excerpt from Liu et al. (2020) that contains five citations. As we can see that Symtax correctly recommend three citations in the top-10, whereas HAtten only rec-

Citation Context:- "Self-training methods such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLNet (Yang et al., 2019) have brought significant performance gains, but it can be challenging to determine which aspects of the methods contribute the most. Training is computationally expensive, limiting the amount of tuning that can be done, and is often done with private training data of varying sizes, limiting our ability to measure the effects of the modeling advances."
Query Title:- RoBERTa: A Robustly Optimized BERT Pretraining Approach

#	BM25 recommendation	HAtten recommendation	SymTax recommendation
1	Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks	**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**	**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
2	Language-agnostic BERT Sentence Embedding	Deep Residual Learning for Image Recognition	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
3	FlauBERT: Unsupervised Language Model Pre-training for French	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension	Deep Residual Learning for Image Recognition
4	Passage Re-ranking with BERT	Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
5	Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings	Neural Machine Translation by Jointly Learning to Align and Translate	SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems
6	BERTweet: A pre-trained language model for English Tweets	Adam: A Method for Stochastic Optimization	StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding
7	Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling	Unified Language Model Pre-training for Natural Language Understanding and Generation	MPNet: Masked and Permuted Pre-training for Language Understanding
8	Pre-trained language models as knowledge bases for Automotive Complaint Analysis	SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems	**Cross-lingual Language Model Pretraining**
9	Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing	BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning	**XLNet: Generalized Autoregressive Pretraining for Language Understanding**
10	**Cross-lingual Language Model Pretraining**	StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding	Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing

Table 5: The table shows the top-10 citation recommendations given by various algorithms for a randomly chosen example from ArSyTa. Valid predictions are highlighted in bold. It clearly shows that SymTax (SciBERT_vector) is able to recommend three valid articles in the top-10. In contrast, each of the HAtten and BM25 could recommend only one valid article for the given citation context. # denotes the rank of the recommended citations.

ommend one citation correctly at rank 1 and BM25 only suggest one correct citation at rank 10. The use of title is crucial to performance, as we can see that many recommendations consist of the words "BERT" and "Pretraining", which are the keywords present in the title. One more observation is that the taxonomy plays a vital role in recommendations. The taxonomy category of the query is 'Computation and Language', and most of the recommended articles are from the same category. SymTax gives only one recommendation (Deep Residual Learning for Image Recognition) from a different category, i.e."Computer Vision", whereas HAtten recommends three citations from different categories, i.e. (Deep Residual Learning for Image Recognition) from "Computer Vision" and (Batch Normalization, and Adam) from "Machine Learning".

7 Conclusion

In this paper, we present a model for local citation recommendation that leverages the notion of *Symbiosis* from Biology, and we draw its analogy with human citation behaviour. We propose the notion of taxonomy fusion for learning rich concept representations and project them into hyperbolic space to derive a latent feature. We introduce a novel dataset that is comparatively large, dense, recent and more challenging than other existing datasets. Through several experiments and analyses, we prove our model as highly modular, which can run on datasets with comparatively few signals and accommodate additional signals as well. Our model consistently outperforms SOTA by huge margins for all evaluation metrics across all datasets.

8 Acknowledgement

We would like to thank Vertika and Garima for their inputs in the data curation.

9 Limitations

The current work marks the initial step towards incorporating human behaviour in designing a recommendation system for citation. We show empirically that such an inclusion leads to significant gains in performance. However, additional signals that resemble the actual citation behaviour can be incorporated to yield better performance. In the current setting, our system is limited to work in offline mode. We intend to transform our system to operate in the online setting, providing real-time recommendations.

10 Ethics Statement

Our work focuses on advancing citation recommendation and assisting the researchers in their academic writing process, where we are committed to maintain ethical standards. We will release our curated dataset and it can serve as a large and suitable benchmark for future research. Upholding transparency, our methodologies adhere to ethical guidelines, ensuring the responsible considerations. We assert that our work contributes positively to the citation ecosystem without raising ethical or moral concerns. We remain vigilant in addressing any unforeseen ethical challenges, driven by a commitment to principled research conduct. Our goal is to foster collaboration, uphold privacy, and enhance scholarly discourse.

References

Zafar Ali, Guilin Qi, Khan Muhammad, Pavlos Kefalas, and Shah Khusro. 2021. Global citation recommendation employing generative adversarial network. *Expert Syst. Appl.*, 180(C).

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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Tao Dai, Li Zhu, Yaxiong Wang, and Kathleen M Carley. 2019. Attentive stacked denoising autoencoder with bi-lstm for personalized context-aware citation recommendation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:553–568.

Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. 2020. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE.

Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1093–1096.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. *Advances in neural information processing systems*, 31.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. 2019. Combining neural networks with personalized pagerank for classification on graphs. In *International Conference on Learning Representations*.

Nianlong Gu, Yingqiang Gao, and Richard HR Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *European Conference on Information Retrieval*, pages 274–288. Springer.

Lantian Guo, Xiaoyan Cai, Fei Hao, Dejun Mu, Changjian Fang, and Libin Yang. 2017. Exploiting fine-grained co-authorship for personalized citation recommendation. *IEEE Access*, 5:12714–12725.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430.

Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C Lee Giles, and Lior Rokach. 2012. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1910–1914.

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C Giles. 2015. A neural probabilistic model for context based citation recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124:1907–1922.

Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. The stm report. *An overview of scientific and scholarly publishing. 5th edition October*, page 94.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Avishay Livne, Vivek Gokuladas, Jaime Teevan, Susan T Dumais, and Eytan Adar. 2014. Citesight: supporting contextual citation recommendation using differential search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 807–816.

Zoran Medić and Jan Šnajder. 2020. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the first workshop on scholarly document processing*, pages 97–103.

Laurent Meunier, Raphael Ettedgui, Rafael Pinot, Yann Chevaleyre, and Jamal Atif. 2022. Towards consistency in adversarial classification. In *Advances in Neural Information Processing Systems*, volume 35, pages 8538–8549. Curran Associates, Inc.

Gabriela F Nane, Nicolas Robinson-Garcia, François van Schalkwyk, and Daniel Torres-Salinas. 2023. Covid-19 and the scientific publishing system: growth, open access and scientific fields. *Scientometrics*, 128(1):345–362.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ramit Sawhney, Ritesh Soun, Shrey Pandit, Megh Thakkar, Sarvagya Malaviya, and Yuval Pinter. 2022. Ciaug: Equipping interpolative augmentation with curriculum learning. In *NAACL*, pages 1758–1764.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Yifan Wang, Yiping Song, Shuai Li, Chaoran Cheng, Wei Ju, Ming Zhang, and Sheng Wang. 2022. Disencite: Graph-based disentangled representation learning for context-specific citation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11449–11458.

Qianqian Xie, Yutao Zhu, Jimin Huang, Pan Du, and Jian-Yun Nie. 2021. Graph neural collaborative topic model for citation recommendation. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–30.

A Appendix

We conduct another quantitative analysis using the section heading as an additional signal in our reranking module.

A.1 Additional Experiment

We concatenate the section heading with query context in reranker and run our two SymTax variants. From Table 6, we can observe that using section heading leads to a significant performance drop in SciBERT_vector for all the metrics. However, for SPECTER_graph, the overall performance remains nearly the same. Both of these patterns clearly indicate that using section heading as a feature acts as a noise, and thus the citation contexts are already rich. Since our proposed dataset contains this additional feature, it is suitable for two additional tasks: context-specific citation generation (Wang et al., 2022), and section heading prediction for a given citation context.

A.2 Implementation Details

We run all experiments on an NVIDIA DGX A100 GPU cluster, and our model is highly efficient in that it only requires about 5GB of GPU memory for training SymTax. We use Adam optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$. We set learning rate to $1e^{-4}$ and weight decay to $1e^{-5}$ in prefetcher, whereas for reranker these were set to $1e^{-5}$ and to $1e^{-2}$ respectively for fine-tuning LM. We choose the top 100 candidates returned by prefetcher as input to Enricher. We choose random seed = 12 for sampling 10k citation contexts from ArSyTa for conducting Quantitative Analysis as discussed in Table 4 and Table 5. We set margin $m = 0.1$ in the triplet loss function. The maximum sequence length for LM is 512. The values of D and d in the reranker are 768 and 512 respectively. Since ArSyTa is a highly dense network, we sort the enriched candidate list by frequency count and take the top 300 candidates with the highest frequency count to run the reranker further. All the results are reported as an average of 3 runs.

A.3 Datasets

ACL-200. This dataset contains papers published at ACL venues. It is a processed version of the ACL-ARC dataset created using ParsCit[11], a string parsing package based on conditional random field. It contains citation contexts by considering ±200 characters around the citation placeholder.

SymTax Variant	R@5	R@10	R@20	R@50	NDCG	MRR
SciBERT_vector	0.2097	0.2876	0.3754	0.4939	0.1432	0.0978
+ Section	0.1556	0.2289	0.3193	0.4707	0.1123	0.0763
SPECTER_graph	0.2025	0.2667	0.3456	0.4654	0.1386	0.0981
+ Section	0.2005	0.2772	0.3632	0.5001	0.1385	0.0950

Table 6: Analysis on the inclusion of section heading as a feature on 10k random samples from ArSyTa data. The results indicate that using section heading as a feature acts as a noise as the citation contexts are already rich.

FullTextPeerRead. It is an expansion of PeerRead dataset that contains the peer reviews of papers submitted to top venues in the Artificial Intelligence domain. So, FullTextPeerRead contains the citation contexts from the papers present in the PeerRead dataset.

RefSeer. This dataset is curated by extracting scientific articles belonging to various engineering domains. A citation excerpt is taken as the text of ±200 characters around the citation marker. It is a large dataset that contains 3.7 million citation contexts.

arXiv (HAtten). It is created using arXiv papers from a large and diverse corpus of scientific articles contained in S2ORC[12]. For every paper having its full text available, a citation excerpt is considered if the cited paper is also present in the arXiv database. Following the similar trend setup by ACL-200 and RefSeer, this dataset is also curated by considering the words in the ±200 character window around the citation marker.

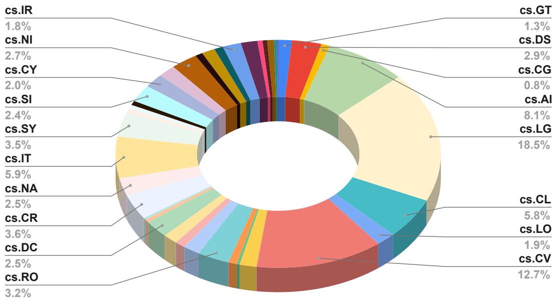

Figure 3: Statistics show the distribution of major category classes of flat-level arXiv taxonomy corresponding to ArSyTa. The highest number of research papers belong to Machine Learning (cs.LG), Computer Vision (cs.CV), and Artificial Intelligence (cs.AI).

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

[12] https://github.com/allenai/s2orc