

Local Citation Recommendation via Heterogeneous Graph Neural Networks and Contrastive Learning

Baha Eddine Hantous^{1,2}, Bruno Yun^{1,*} and Nadia Yacoubi Ayadi¹

¹Université Claude Bernard Lyon 1, CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Lumière Lyon 2, LIRIS, UMR5205, 69622 Villeurbanne, France

²Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000, Le Bardo, Tunisia

Abstract

The exponential growth of scientific literature makes it extremely difficult for researchers to identify the most relevant prior work to cite at a given point in a manuscript. Local citation recommendation (LCR) addresses this problem by suggesting, for a given citation context, the papers that best support the underlying claim. Existing approaches to LCR are typically confined to single-stage retrieval, prefetch-then-rerank pipelines, or generative formulations that only weakly exploit the structural relationships between papers, authors, and citation contexts. In this paper, we propose a Scholarly Knowledge Graph (SKG)-based approach to LCR that jointly exploits the semantic structure of scholarly data and neural representation learning. We first construct ArgKG, a domain-specific SKG for the argumentation-theory literature, comprising 1,576 collected corpus papers, 24,438 additional referenced papers (26,014 in total), and 82,868 citation contexts, modelled as an RDF graph that reuses established scholarly ontologies and vocabularies to represent papers, authors, citation links, and citation-context semantics explicitly. Building on this representation, we introduce a two-tower architecture that learns a shared embedding space for citation contexts and candidate papers through contrastive learning. Candidate papers are represented by metapath-based features extracted from the SKG – their title and abstract, reference neighbourhood, and previously observed citation contexts – fused using a transformer-based encoder. We report experimental results across several feature-set and architecture configurations on ArgKG and on the established ACL-200 benchmark, showing that combining a paper’s own title and abstract with its citation neighbourhood yields the best-performing configuration and outperforms most of the existing LCR baselines on ACL-200. To support reproducibility, the code, trained models, and SKG construction pipeline are publicly released at <https://github.com/Ba-aH/SciteKG>.

Keywords

Citation Recommendation, Heterogeneous Graph Neural Networks, Contrastive Learning, Scholarly Knowledge Graphs, Knowledge Graph Construction

1. Introduction

Citing prior work appropriately is a foundational requirement of scientific writing: citations substantiate an author’s claims, ensure transparency and reproducibility, allow readers to trace and verify the lineage of an argument, and give credit to the researchers whose work is being built upon. As the volume of scientific literature keeps expanding – the number of records indexed in DBLP alone has grown from a few hundred thousand in the late 1990s to over eight million today [1], identifying, among this ever-growing body of work, the right papers that should be cited at a given point of a manuscript has become an increasingly demanding task for authors.

This challenge motivates the task of *local citation recommendation* (LCR): given a citation context – the span of text surrounding a citation marker – the goal is to recommend, among a large corpus of candidate papers, those that are the most suitable to be cited. LCR is often confused with two related, but distinct, tasks. *Paper recommendation* aims to suggest papers that are worth reading or studying [2, 3], generally based on a user’s reading history rather than on a specific textual claim; a recommended paper therefore only needs to be topically relevant to a broad research interest, whereas a paper recommended for a citation context must specifically support the exact statement being made in the text, which makes

6th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Oct 2026, Bari, Italy

*Corresponding author.

✉ baha-eddine.hantous@univ-lyon1.fr (B. E. Hantous); bruno.yun@univ-lyon1.fr (B. Yun); nadia.yacoubi-ayadi@univ-lyon1.fr (N. Yacoubi Ayadi)

ORCID 0000-0001-9370-3917 (B. Yun); 0000-0002-9421-8566 (N. Yacoubi Ayadi)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

LCR a more complex problem. *Citation placeholder detection* [4] is the complementary task that refers to the task of identifying *where* in a text a citation is needed to be placed; LCR, in contrast, assumes the citation context is already known and focuses solely on recommending the papers that should fill it. A first challenge specific to LCR is that a single citation context may correspond to more than one ground-truth paper at once (e.g. “... several works address this problem [A,B]”); treating each citation marker independently, or ignoring this multiplicity altogether, leads to both an inaccurate account of the task and a misleading evaluation of a system’s performance. Existing approaches to LCR range from single-stage retrieval over IR metrics such as TF-IDF or learned embeddings [5, 6], to prefetch-then-rerank pipelines [7, 8], to graph-enriched re-ranking [9], and more recently generative approaches that directly produce the citation string [10, 11]. We provide a brief presentation of the most prominent research works in Section 2. These approaches, however, either ignore the structural relationships between papers, authors, and citation contexts entirely, or exploit them only superficially, e.g. by appending a fixed number of cited neighbours to a candidate list. In this paper, we formalize the underlying corpus, references, and citation contexts as an explicit, queryable, ontology-grounded *Scholarly Knowledge Graph* (SKG), which enables both the interoperability of the resulting resource and the diversity of structural signal a downstream model can draw on, as demonstrated in [12]. Therefore, we make the following contributions:

- We describe a pipeline for constructing a SKG from a scraped corpus of scientific papers, in which references are cleaned and validated against OpenAlex, citation contexts are localized around each citation marker, and the resulting data is mapped to RDF using established Semantic Publishing and Referencing (SPAR) ontologies [13, 14].
- We propose a two-tower architecture for LCR, in which candidate papers are represented through metapath-based features (e.g., their own content (title + abstract), their references, or their citation contexts) fused through a transformer-based semantic fusion module; whereas citation contexts are encoded with a fine-tuned SciBERT model.
- We instantiate this pipeline and architecture on a domain-specific corpus of argumentation-theory papers, and report experimental results across several feature-set and architecture configurations, under both single- and multi-positive evaluation protocols.

This paper is structured as follows. Section 2 reviews existing approaches proposed in the context of LCR. Section 3 formalizes the notation needed for our approach to LCR. Section 4 introduces our approach for LCR: the modelling of the SKG and the two-tower training architecture. Section 5 reports our experimental results, and Section 6 concludes and outlines future work.

2. Related Work

Early approaches to local citation recommendation (LCR) formulate the task as a retrieval problem, embedding both the citation context and candidate papers into a shared representation space and ranking candidates according to their relevance to the query context. He et al. [5] introduced this context-aware formulation by representing citation contexts and candidate papers as TF-IDF vectors, establishing the retrieve-and-rank paradigm that subsequent work has extended. DualEnh [6] builds on this formulation with a Bi-LSTM encoder that captures forward and backward word dependencies in the citation context, while additionally incorporating global information from the citing article, namely its title and abstract. The final recommendation score combines this semantic matching signal with a bibliographic relevance measure based on the candidate paper’s popularity within the scientific community. However, these early work fail to exploit additional bibliographical information such as the reference list, how other papers are citing it, etc.

More recent retrieval-based approaches decouple candidate retrieval from candidate scoring, typically using a lightweight model to prefetch a small candidate set followed by a more computationally expensive re-ranking stage. HAtten [7, 8] follows this retrieve-then-rerank strategy, first prefetching candidates with a hierarchical attention-based encoder and subsequently re-ranking them using a fine-tuned SciBERT model that jointly encodes the query text and each candidate’s title and abstract. Other approaches enrich the retrieved candidate set using structural information from the citation network.

SymTax [9], for example, introduces an additional enrichment stage that traverses the citation graph to retrieve papers cited by each prefetched candidate, while excluding papers that cite it, before scoring each query-candidate pair according to a combination of textual and taxonomical similarity. Although SymTax demonstrates the value of exploiting citation-graph structure, its graph-based expansion remains a fixed, one-hop heuristic applied after candidate retrieval, rather than a learned multi-hop representation of the candidate integrated directly into the scoring process.

In parallel, LCR has increasingly been formulated as a text generation problem. CiteBART [10] introduces a citation-specific continual pre-training strategy based on BART, replacing citation markers with a mask token and training the model to generate the corresponding citation string directly from its surrounding context. Its *Global* variant further conditions generation on the citing paper’s title and abstract. Generative approaches eliminate the need for an explicit candidate-ranking stage, but this flexibility comes at the cost of reduced grounding in a verifiable and interoperable representation of the underlying scientific corpus. Moreover, because these approaches generate textual outputs, they are susceptible to hallucinations, whereby the model may produce fabricated or otherwise inaccurate references.

Across these retrieval-based and generative approaches, structural information about the citation network is exploited, at best, through fixed heuristics, such as SymTax’s one-hop candidate expansion, or through global textual and bibliographic signals, such as DualEnh’s popularity score and CiteBART-Global’s title and abstract conditioning. In contrast, our approach explicitly models the structural and semantic properties of candidate papers through learned multi-hop representations. Specifically, we ground the underlying corpus, references, and citation contexts in an explicit knowledge graph, making the resulting resource queryable and interoperable beyond the citation recommendation task itself. Then, we represent each candidate using several *metapath-based* features, which are fused through a transformer-based module inspired by SeHGNN [11], a heterogeneous graph neural network originally proposed for node classification. To the best of our knowledge, this design has not previously been applied to citation recommendation.

3. Preliminaries

3.1. Problem definition

In this paper, we focus on the task of *local citation recommendation*. Given a snippet of scientific text containing a citation marker and a database of candidate papers, the goal is to recommend a set of papers from the database that are most appropriate for filling the citation marker. In following, we formalize the notions of *citation context* and *recommendation model*.

Definition 1 (Citation context). Let P be a set of papers. A citation context is a pair (s, p) where s is a citing string containing a citation marker $\langle cite \rangle$ and $p \in P$ is an appropriate paper that can fill the citation marker in s .

Citation contexts serve as training instances for our recommendation model, as they provide contextual information about where and why a given paper can be cited. Note that s may be filled by several papers, i.e., distinct pairs $(s, p_1), (s, p_2), \dots$ sharing the same string s may occur. For a list of papers P , we write $\mathcal{C}(P)$ for the set of all citation contexts extracted from P (duplicate occurrence pairs are collapsed).

For a citing string s' , we write $\text{cit}(s') = \{p \mid (s, p) \in \mathcal{C}(P) \text{ and } s = s'\}$ for the set of *ground-truth cited papers* associated with s' in P . The set of distinct citing strings to evaluate on is $S_P = \{s \mid (s, p) \in \mathcal{C}(P)\}$. By construction, $\text{cit}(s) \neq \emptyset$ for every $s \in S_P$.

Example 1. Consider the following citing string s from a paper p_0 :

“several works have proposed using Transformer architectures for citation recommendation $\langle cite \rangle$, while others leverage citation-graph embeddings.”

If this citing string s has two ground-truth papers associated with its citation marker (e.g. p_A and p_B), we will obtain two separate citation contexts (s, p_A) and (s, p_B) .

Definition 2 (Local Citation Recommendation (LCR)). An LCR model M is a function that takes as input a citing string s and a set of papers P . It returns a finite sequence of distinct pairs $(p_i, \text{sim}_i)_{i=1}^k$ with $p_i \in P$, $\text{sim}_i \in [0, 1]$, and $p_i \neq p_j$ for $i \neq j$, where sim_i denotes the estimated relevance (or similarity) between the citing context s and the paper p_i , and k is the number of distinct papers returned by M .

For a prediction of M on s and P , let $\Pi_n^M(s)$ denote the set of the top- n papers in this ranking (i.e., $\{p_1, \dots, p_{\min(n, k)}\}$); since the p_i are distinct and the order is strict.

3.2. Knowledge Graph-based Representation

The definitions above treat P as an unstructured set of papers, which suffices to state the recommendation task and its evaluation metric. In this paper, we propose a citation recommendation approach that explicitly exploits the relational structure connecting papers, authors, and citation contexts, rather than treating P as an unstructured collection of candidates. We therefore represent P , together with citation contexts and other bibliographic entities, as nodes of a *heterogeneous graph*, and instantiate the metapath-based paper features used in Section 4.2 directly over this graph.

Definition 3 (Heterogeneous graph). A heterogeneous graph is defined as $G = (V, E, \mathcal{T}^v, \mathcal{T}^e)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Each node $v \in V$ is associated with a node type through the mapping function $\phi: V \rightarrow \mathcal{T}^v$, where $\mathcal{T}^v$ denotes the set of node types. Similarly, each edge $e = (u, v) \in E$ is associated with an edge type through the mapping function $\psi: E \rightarrow \mathcal{T}^e$, where $\mathcal{T}^e$ is the set of edge types. The type of an edge is determined by both its semantic relation and the types of its endpoints, i.e., $\psi(e): \phi(u) \xrightarrow{r} \phi(v)$, where $r \in \mathcal{T}^e$.

Note that when $|\mathcal{T}^v| = |\mathcal{T}^e| = 1$, the graph reduces to a homogeneous graph. Given the SKG introduced in Section 4.1, the set of node types includes $\mathcal{T}^v = \{\text{Paper}, \text{Context}\}$, corresponding to the RDF classes `bibo:Article` and `cito:Citation`, respectively. The graph includes the edge type `cito:cites`, representing citation links between publications, as well as the edge types `cito:hasCitingEntity` and `cito:hasCitedEntity` which associate each citation context with its citing and cited papers.

Definition 4 (Metapath). A metapath is a sequence of node types connected by edge types, representing a composite semantic relation in a heterogeneous graph. Formally, a metapath is defined as $\mathcal{P} = c_1 \xrightarrow{r_1} c_2 \xrightarrow{r_2} \dots \xrightarrow{r_{l-1}} c_l$, where $c_i \in \mathcal{T}^v$ and $r_i \in \mathcal{T}^e$. Given a metapath $\mathcal{P}$, a metapath instance is a path $(v_1, e_1, v_2, \dots, e_{l-1}, v_l)$ such that $\phi(v_i) = c_i$ and $\psi(e_i) = r_i$ for every i . The terminal node v_l is called a metapath-based neighbor of v_1 under $\mathcal{P}$.

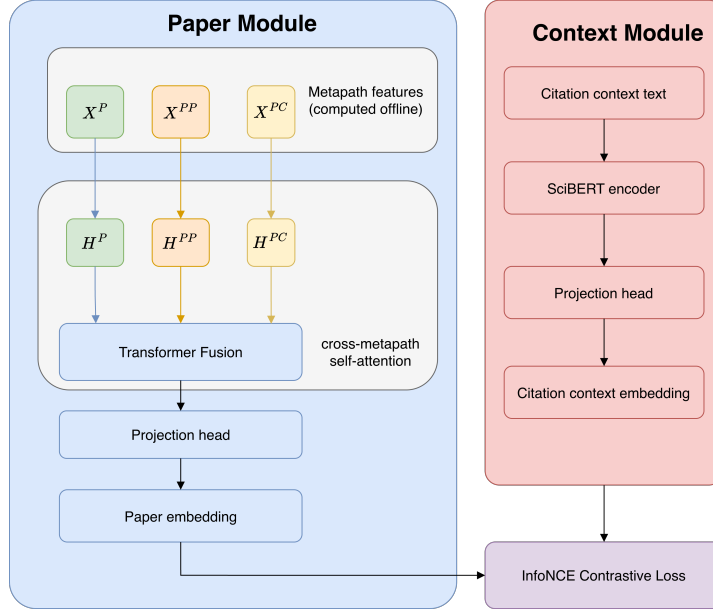
Several metapaths can be defined over a SKG. The simplest is the *0-hop metapath* `Paper`, which corresponds to the target paper itself. A *1-hop citation metapath*, `Paper` $\xrightarrow{\text{cito:cites}}$ `Paper`, connects a paper to the papers it cites. Another *1-hop metapath*, `Context` $\xrightarrow{\text{cito:hasCitedEntity}}$ `Paper`, connects a citation context to the paper being cited, while `Context` $\xrightarrow{\text{cito:hasCitingEntity}}$ `Paper` connects a citation context to the citing paper. Longer metapaths can be obtained by composing these elementary relations, thereby capturing richer semantic dependencies between scientific publications.

4. Our Approach

Our LCR approach relies on a Scholarly Knowledge Graph (SKG) built from a domain-specific corpus, structuring a set of scientific papers – together with their descriptive and bibliographic metadata (e.g. title, venue, publication year), their authors, and their citation contexts – into a shared, ontology-based representation (Section 4.1). Building on the heterogeneous-graph formalism introduced in Section 3.2, we instantiate a small set of metapaths over the obtained SKG and use them to derive structural features for each candidate paper. As illustrated in Figure 1, our approach builds upon a two-tower architecture: a Paper Module fuses metapath-based features into a single paper embedding, while a Context Module

encodes the text surrounding a citation placeholder; both towers are trained jointly with an InfoNCE contrastive objective, so that each paper’s embedding is pulled towards the citation contexts that reference it, and pushed away from unrelated papers.

Figure 1: Two-module training architecture for local citation recommendation.



4.1. SKG modelling and construction

4.1.1. Corpus pre-processing pipeline

The initial corpus was built by scraping scientific papers related to formal argumentation from OpenAlex¹. We selected argumentation as the target domain because it constitutes a well-defined research community with established terminology, strong interconnections between publications, and extensive citation practices, making it a suitable and representative domain for studying LCR. OpenAlex was chosen as the primary bibliographic source due to its open access, broad coverage, and rich metadata (authors, publication venues, DOIs, and citation links) [15]. Since the collected papers were retrieved as PDF documents, an Optical Character Recognition (OCR) step was required to convert them into plain text. We used PaddleOCR², a widely-used open-source OCR framework known for its reliable performance on complex document layouts. A dedicated script automatically locates and isolates the reference (bibliography) section of each paper, using regular expressions designed to detect common bibliography section headers and boundary patterns. This task was facilitated by PaddleOCR’s output format, which tags recognized text with section-level identifiers, making it straightforward to isolate section boundaries such as the start of the reference list. Once identified, the reference section is extracted and saved separately from the main body text of each paper, producing a raw textual file per paper containing only its reference list.

As the reference sections extracted from OCR output are often noisy, containing formatting artifacts, line-break errors, and inconsistent citation styles inherent to OCR-based text extraction, we designed an automated cleaning pipeline using n8n³, where the raw reference text is passed to a Large Language Model (LLM), namely Qwen3.6-35B-A3B, which is prompted to reconstruct a cleaned, structured version of each reference entry and output the result as a JSON object. The cleaned references are then validated against OpenAlex records to confirm their existence and correctness. This validation step relies

¹<https://openalex.org/>

²<https://www.paddleocr.ai/main/en/index.html>

³<https://n8n.io/>

on fuzzy string matching using the rapidfuzz library⁴, specifically its fuzz module, which computes similarity scores based on the Levenshtein distance. A reference is considered a valid match when the similarity score exceeds a threshold of 85%, since titles are normalized beforehand (removing colons, special characters, and other formatting differences) to be able to query the OpenAlex API. As a fallback we also did author matching for validation.

For each successfully matched reference, the corresponding authors and publication venue are also retrieved from OpenAlex and stored alongside the reference metadata. In the final step to capture the textual context surrounding each citation, we developed a script that parses the full text of each paper and detect all citation markers using regular expressions. For every detected marker, the surrounding text is cleaned and a local context window is extracted, consisting of up to 600 characters before and after the citation marker. Each extracted context is then linked to its corresponding cited paper, forming the basis for the citation-context dataset used in later stages of the pipeline.

4.1.2. RDF Data modelling

Given the scientific corpus extracted and processed as described in the previous section, the next stage consists of transforming the resulting structured metadata—including publications, authors, bibliographic references, and in-text citation contexts—into a semantically enriched RDF Knowledge Graph (KG). To ensure interoperability and facilitate reuse across the scholarly knowledge ecosystem, the RDF model builds upon established Semantic Publishing and Referencing (SPAR) ontologies [13, 14]. In particular, citation semantics are represented using the Citation Typing Ontology (CiTO)⁵, while bibliographic metadata are modeled using FaBiO (FRBR-aligned Bibliographic Ontology) and BIBO (Bibliographic Ontology). Information about authors and publication relationships is represented using FOAF and Dublin Core Terms, ensuring compatibility with existing Linked Open Data resources and facilitating integration with external scholarly knowledge graphs such as OpenAlex and other bibliographic datasets [16].

While these ontologies provide comprehensive vocabularies for describing scholarly entities and citation relations, they do not explicitly capture the textual evidence associated with individual citation occurrences. To bridge this gap, we extend the model with a lightweight vocabulary dedicated to citation contexts. This extension introduces concepts for representing citation markers, citation types, and the textual fragments immediately preceding and following each citation occurrence. Treating citation contexts as first-class semantic resources enables the KG to preserve the linguistic evidence supporting citation links, thereby enriching the scholarly graph with contextual information that can be leveraged for citation intent analysis, argument mining, explainable scholarly search, and context-aware retrieval. More generally, this contextualized representation provides a richer substrate for semantic reasoning and LLM-assisted exploration of scientific literature, where both the structural citation network and the discourse surrounding citations contribute to knowledge discovery.

- **Scientific publications** are instances of `bibo:Article`. Both the publications forming the corpus and the referenced papers are described through bibliographic metadata, including title, publication venue, publication year, DOI, OpenAlex identifier, and the number of references.
- **Authors** are represented as individual resources linked to publications through the `citekg:hasAuthor` and `citekg:authorOf` properties, enabling the reconstruction of collaboration networks and authorship relations.
- **Citation relationships** between publications are modelled using the properties `cito:cites` and `cito:isCitedBy`, allowing the citation network to be traversed in both directions.
- **Citation occurrences** are represented as instances of `cito:Citation`. Rather than merely indicating that one publication cites another, each citation occurrence is modeled as a first-class resource connected to both the citing and cited publications through `cito:hasCitingEntity` and `cito:hasCitedEntity`. Each citation additionally stores contextual information, including

⁴<https://rapidfuzz.com/>

⁵<https://sparontologies.github.io/cito/current/cito.html>

the citation marker appearing in the text, the citation type, and the textual context immediately preceding and following the citation marker such illustrated in Listing 2.

Listing 1 illustrates the RDF representation of a publication within the knowledge graph. Besides its bibliographic metadata, the paper is linked to its incoming and outgoing citation relations, thereby embedding it into the global citation network.

Listing 1: Example of an RDF representation of a scientific publication

```

1 <https://citekg.org/resource/paper/4cd5a32e4d3cdce3>
2   a bibo:Article ;
3   dcterms:title "Argumentative XAI: A Survey"@en ;
4   dcterms:isPartOf "Proceedings of
5     the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-2021"@en ;
6   cito:cites
7     <https://citekg.org/resource/paper/057079ac8cdb20f0>,
8     <https://citekg.org/resource/paper/0976dee38ef32a8c>,
9     <https://citekg.org/resource/paper/0a9aae59e18aa8ff>;
10  cito:isCitedBy
11    <https://citekg.org/resource/paper/039fc7c9471b511c>,
12    <https://citekg.org/resource/paper/05d91274c975f1fb>,
13    <https://citekg.org/resource/paper/153e7c22347499bd>;
14  fabio:hasPublicationYear "2021"^^xsd:gYear ;
15  citekg:doi "https://doi.org/10.24963/ijcai.2021/600" ;
16  citekg:openalexId "https://openalex.org/W3177471157" ;
17  citekg:totalReferences 76 .

```

Listing 2: Example of an RDF representation of a citation context

```

1 <https://citekg.org/resource/citation/45672b25ce2a7130/51>
2   a cito:Citation ;
3   cito:hasCitingEntity
4     <https://citekg.org/resource/paper/45672b25ce2a7130> ;
5   cito:hasCitedEntity
6     <https://citekg.org/resource/paper/4cd5a32e4d3cdce3> ;
7   citekg:hasCitationMarker "[20]" ;
8   citekg:hasCitationType "single" ;
9   citekg:hasContextBefore "rst category is robustness/sensitivity, amounting
10     to the robustness of explanation methods under small perturbations of the inputs. ...." ;
11   citekg:hasContextAfter "is used to
12     measure whether the extracted AF is faithful to the underlying model. Also, transparency
13     and trust are studied as cognition-related properties of explanations [cite]...." .

```

Listing 3 retrieves every citation occurrence of a given paper from the SKG, returning the citing paper, citation marker, type, and surrounding context for each match.

Listing 3: SPARQL query retrieving all citations context of a given paper

```

1 SELECT ?citation ?context ?citingEntity ?marker ?citationType
2 WHERE {
3   ?citation a cito:Citation ;
4     cito:hasCitedEntity <https://citekg.org/resource/paper/4cd5a32e4d3cdce3> ;
5     cito:hasCitingEntity ?citingEntity ;
6     cito:hasContext ?context ;
7     citekg:hasCitationMarker ?marker ;
8     citekg:hasCitationType ?citationType .
9 }

```

4.2. Two-Tower Architecture for Citation Recommendation

We formalize LCR as a contrastive representation-learning problem over the metapaths introduced in Definition 4. As illustrated in Figure 1, a Paper Module and a Context Module map, respectively, candidate

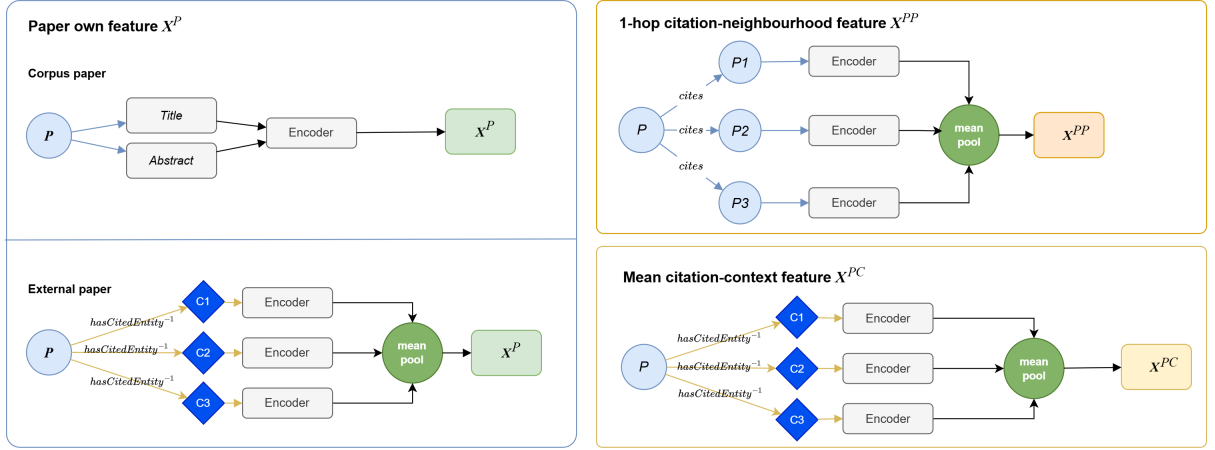
papers and citation contexts to a shared d -dimensional embedding space, such that the embedding of a cited paper lies close to the embeddings of the contexts that cite it, and far from the embeddings of unrelated papers.

4.2.1. Paper Module

The Paper Module maps each candidate paper i to a fixed-dimensional embedding $z_i^{\text{paper}} \in \mathbb{R}^d$ (in this work, we fix $d = 768$). As illustrated in Figure 2, a paper i is first represented, offline, by three metapath-based feature vectors $X_i^P, X_i^{PP}, X_i^{PC} \in \mathbb{R}^{d_0}$, instantiating Definition 4 over the SKG:

- X_i^P (own content, metapath P): for a corpus paper, the SciBERT encoding of the concatenated title and abstract; for an external paper – one for which no abstract is available – the mean-pooled SciBERT encoding of all citation contexts in which paper i is cited.
- X_i^{PP} (1-hop citation neighbourhood, metapath P-cites-P): the mean-pooled SciBERT encodings of the papers directly cited by paper i , capturing its topical context through its reference list.
- X_i^{PC} (1-hop citation-context neighbourhood, metapath P-hasCitingEntity⁻¹-C): the mean-pooled SciBERT encodings of all citation contexts in which paper i is cited.

Figure 2: Overview of the metapath feature computation



Per-metapath projection. Since each metapath encodes a structurally distinct signal, each feature vector is projected independently through its own two-layer MLP, following the per-metapath projection design of SeHGNN [11]:

$$H_i^{\mathcal{P}} = \text{MLP}_{\mathcal{P}}(X_i^{\mathcal{P}}) \in \mathbb{R}^d, \quad \mathcal{P} \in \{P, PP, PC\},$$

so that no parameters are shared across metapaths.

Transformer fusion. The three projected vectors $H_i^P, H_i^{PP}, H_i^{PC}$ are stacked into a length-3 sequence and passed through a self-attention layer that lets each metapath attend to the others, following SeHGNN’s semantic fusion module [11]. For each metapath $\mathcal{P}_k \in \{P, PP, PC\}$:

$$Q_i^{\mathcal{P}_k} = W_Q H_i^{\mathcal{P}_k}, \quad K_i^{\mathcal{P}_k} = W_K H_i^{\mathcal{P}_k}, \quad V_i^{\mathcal{P}_k} = W_V H_i^{\mathcal{P}_k},$$

$$\alpha_{k,l} = \frac{\exp(Q_i^{\mathcal{P}_k} \cdot K_i^{\mathcal{P}_l \top})}{\sum_m \exp(Q_i^{\mathcal{P}_k} \cdot K_i^{\mathcal{P}_m \top})}, \quad \tilde{H}_i^{\mathcal{P}_k} = \beta \sum_l \alpha_{k,l} V_i^{\mathcal{P}_l} + H_i^{\mathcal{P}_k},$$

where W_Q, W_K, W_V and the scalar β are learned and shared across metapaths, rather than treating all metapaths uniformly. This lets the model weight each metapath’s contribution dynamically, on a per-paper basis. The three fused vectors are concatenated into a single representation $\tilde{H}_i = [\tilde{H}_i^P \parallel \tilde{H}_i^{PP} \parallel \tilde{H}_i^{PC}]$.

Projection head. $\tilde{H}_i$ is passed through a linear projection head and ℓ_2 -normalised onto the unit sphere, yielding the final paper embedding

$$z_i^{\text{paper}} = \frac{W_o \tilde{H}_i + b_o}{\|W_o \tilde{H}_i + b_o\|_2},$$

so that similarity between paper and context embeddings can be computed as a dot product, as required by the InfoNCE objective below. We leave the evaluation of this design choice, e.g. assessing whether ℓ_2 -normalisation onto the unit sphere is necessary in the absence of explicit dimensionality reduction, to future work.

4.2.2. Context Module

The Context Module maps a citation context s to an embedding $z_s^{\text{ctx}} \in \mathbb{R}^d$ in the same space. It encodes s with SciBERT [17], takes the [CLS] token representation, and projects it through a linear head to 768 dimensions, ℓ_2 -normalised in the same way as z_i^{paper} . During training, SciBERT’s weights are fine-tuned at a low learning rate (2×10^{-6}) to preserve its pretrained scientific-language representations, while the projection head is trained at a higher rate (1×10^{-4}) to align the resulting embeddings with the Paper Module’s output space.

4.2.3. Training

Both modules are trained jointly with the InfoNCE contrastive objective. At each step, a batch of $B = 128$ citation records $\{(s_k, p_k)\}_{k=1}^B$ is drawn, where s_k is a citation context and $p_k \in P$ is its ground-truth cited paper (cf., Definition 1). The Context Module encodes each s_k into $z_{s_k}^{\text{ctx}}$ and the Paper Module encodes each p_k into $z_{p_k}^{\text{paper}}$; the loss

$$\mathcal{L} = -\frac{1}{B} \sum_{k=1}^B \log \frac{\exp(z_{s_k}^{\text{ctx}} \cdot z_{p_k}^{\text{paper}} / \tau)}{\sum_{j=1}^B \exp(z_{s_k}^{\text{ctx}} \cdot z_{p_j}^{\text{paper}} / \tau)}$$

pulls each context embedding toward its paired paper embedding while pushing it away from the $B - 1$ other papers in the batch, where τ is a fixed temperature hyperparameter. Gradients flow through both modules jointly.

Group-aware batch sampling. The loss above implicitly treats every other paper in a batch as a negative for s_k . This assumption breaks down for citation markers referencing several papers at once (e.g. “[1,2,3]”): per Definition 1, such a marker yields several *sibling* records $(s, p_1), (s, p_2), \dots$ that share the same context string s but differ in their cited paper. If two sibling records were drawn into the same batch, the Context Module would necessarily produce identical embeddings $z_{s_k}^{\text{ctx}} = z_{s_l}^{\text{ctx}}$ for both, yet the loss above would still treat p_l as a negative for s_k and vice versa – penalising the model for correctly recognising a paper that the context does, in fact, cite. To prevent this, we use a custom batch sampler that guarantees no two sibling records are ever drawn into the same batch, i.e. for every batch $\{(s_k, p_k)\}_{k=1}^B$, $s_k = s_l \Rightarrow k = l$. This constraint is only necessary for our domain-specific dataset, which retains multi-positive citation markers; the baseline datasets used for comparison contain single-positive records only and are trained with standard random batch sampling.

Data split. The data split on the corpus dataset is done by splitting the papers using a 80/10/10 ratio for train, validation, and test. If a paper is assigned to the training split, all contexts it contains are also assigned to this split. This is to prevent citation contexts with the same citing string from being split across different partitions. Additionally, during metapath feature calculation for X_i^P for external papers and X_i^{PC} for all papers, we use only citation contexts obtained from papers in the train split.

Empty metapath. Papers with no valid feature under a given metapath set e.g., papers with no outgoing edges (no PP feature) or papers never cited in any existing context in the dataset (no PC feature) are represented by a zero feature vector. At evaluation time, since we use cosine similarity to determine which papers are closest to a given context, papers with an invalid metapath (i.e., zero feature vectors) are excluded from the candidate pool before ranking is computed. This is necessary because cosine similarity is computed by dividing the dot product of two vectors by the product of their magnitudes, since a zero vector has magnitude of zero, this division is undefined, which in practice yields a similarity score of zero. Left unfiltered, such papers could be counted as a hit at Recall@10.

5. Experimental Study

5.1. Evaluation Metrics

We evaluate all configurations using the $R@n$ and MRR metrics introduced in Definition 5 and 6, computed under two protocols. The first evaluates over the full set S_P , including contexts with $|\text{cit}(s)| > 1$ (contexts citing multiple papers at once). The second restricts S_P to single-positive contexts with $|\text{cit}(s)| = 1$ (contexts citing exactly one paper). Comparing both protocols shows how sensitive each configuration is to multi-reference contexts.

For a candidate paper p and a ground-truth citation set $C \subseteq P$, we define the relevance indicator $X(p, C) = 1$ if $p \in C$, and 0 otherwise. Moreover, for a prediction of M on s and P , and a paper $p \in \text{cit}(s)$, let $\text{rank}_M(s, p)$ denote the position i such that $p = p_i$ in M 's ranking.

Definition 5 (Recall at n). The recall at n ($R@n$) of model M over S_P is defined as

$$R@n = \frac{\sum_{s \in S_P} \sum_{p \in \Pi_n^M(s)} X(p, \text{cit}(s))}{\sum_{s \in S_P} \min(|\text{cit}(s)|, n)}.$$

Definition 6 (Mean Reciprocal Rank). The mean reciprocal rank (MRR) of model M over S_P is defined as

$$\text{MRR} = \frac{1}{|S_P|} \sum_{s \in S_P} \left(\frac{1}{|\text{cit}(s)|} \sum_{p \in \text{cit}(s)} \frac{1}{\text{rank}_M(s, p)} \right),$$

Because $\Pi_n^M(s)$ consists of distinct papers and $\text{cit}(s) \neq \emptyset$ for all $s \in S_P$, the denominator is always strictly positive and each context contributes at most $\min(|\text{cit}(s)|, n)$ to the numerator, so $R@n \in [0, 1]$. Note that $R@n$ is a *micro-averaged* (pooled) recall: hits and denominators are summed across all contexts before dividing, so contexts with more ground-truth citations are implicitly weighted more heavily in the score. When $|\text{cit}(s)| \leq n$, a perfect model can recover all ground-truth papers for context s ; when $|\text{cit}(s)| > n$, that context's maximum achievable contribution is capped below its full citation count, reflecting the intrinsic difficulty of the context rather than a limitation of the model.

Unlike the standard MRR, which only credits the single best-ranked ground-truth paper per context, this formulation averages reciprocal ranks over *all* papers in $\text{cit}(s)$ before averaging across contexts, so that multi-reference contexts are neither penalized nor over-weighted relative to single-reference ones.

Figure 3: Dataset statistics and experiments on the impact of the MLP hidden-layer size.

(a) Statistics on the number of papers per dataset				
Dataset	Type	#Papers		
ArgKG	Corpus	1,576		
	External	24,438		
	<i>Total</i>	<i>26,014</i>		
ACL-200	Corpus	13,211		

(b) Split and local citation context statistics per experimental setting.				
Dataset	Train	Val	Test	#Contexts
ArgKG (Corpus+external)	62,354	7,154	9,857	79,365
ArgKG (Corpus only)	12,354	1,290	1,435	15,098
ACL-200	50,860	6,376	6,080	63,316

5.2. Dataset Description

We constructed our own dataset by scraping papers related to argumentation and argumentation theory from OpenAlex, extracting their text via OCR, and preprocessing the results into a knowledge graph of papers and citation contexts (cf. section 4.1). Our dataset distinguishes two types of papers: corpus papers, which we scraped directly and for which full text and metadata are available, and external papers, which appear only in the reference lists of corpus papers. Starting from 1,576 corpus papers, reference extraction expanded the knowledge graph to 26,014 distinct papers overall, the majority being external papers reached only through citations (Table 3a). For each citing-cited paper pair, we extracted a localized citation context from the surrounding text, yielding 79,365 citation context records after cleaning. These were split into train, validation, and test splits, fixed across all experiments for consistency across model variants (Table 3b).

To benchmark our results with existing work, we evaluated our approach on the ACL-200 dataset, a widely used benchmark for local citation recommendation. ACL-200 is a subcategory of ACL-ARC [18], derived from the ACL Anthology Reference Corpus and was constructed by [6] for local citation recommendation and has been adopted in prior work such as [7, 8, 9]. Unlike our dataset, ACL-200 does not distinguish between corpus and external papers, as abstracts are available for all papers and the two categories are therefore equivalent in representation. Starting from the original ACL-200 dataset, we reuse the methodology introduced by [10] and filtered out papers without abstracts, resulting in a subset of 13,211 papers and 63,316 citation context records (Table 3b).

5.3. Results

Table 4a compares the different metapath feature configurations on both variations of ArgKG and on ACL-200. Considering the paper’s own feature P, we observe strong baseline, confirming that the content of a candidate paper is well captured through its title and abstract. Similarly, we note that the PC feature on its own also holds rich semantic signal, indicating that a model can learn a useful representation of a paper purely from the contexts in which it is cited. This is particularly relevant for external papers, which lack an abstract by definition and whose only available semantic representation is PC (or, equivalently, P computed from pooled citation contexts, as described in 4.2.1). This largely explains why P and PC yield nearly identical performance on Corpus+external with R@10 of 0.453 and MRR of 0.377 for P, against R@10 of 0.454 and MRR of 0.374 for PC, since external papers make up 93.94% of the dataset, and for these papers both features

Figure 4: Evaluation of our approach on ArgKG and ACL-200 and comparison with other approaches. Results are grouped by feature set. Bold results refer to the best performing configuration.

(a) Evaluation on ArgKG and ACL-200 dataset.

Feature Set	Corpus + external		Corpus only		ACL-200	
	R@10	MRR	R@10	MRR	R@10	MRR
P	0.453	0.377	0.637	0.520	0.644	0.434
PP	0.002	0.002	0.004	0.003	0.001	0.001
PC	0.454	0.374	0.576	0.463	0.644	0.416
P+PP	0.453	0.379	0.639	0.514	0.628	0.419
P+PC	0.456	0.375	0.618	0.525	0.670	0.449
PC+PP	0.455	0.368	0.567	0.468	0.648	0.419
P+PP+PC	0.460	0.383	0.610	0.521	0.629	0.414

(b) Comparison with state-of-the-art on LCR benchmarks (ACL-200).

Model	ACL-200	
	R@10	MRR
BM25	0.194	0.107
HAtten	0.499	0.242
SymTax	0.653	0.296
Our approach	0.670	0.449

are effectively derived from the same source of information. This is confirmed by comparing Corpus-only and ACL-200, where all candidate papers are corpus papers and their title and abstract are available. Regarding Corpus-only dataset, results show that P clearly outperforms PC, with R@10 of 0.637 and MRR of 0.520 for P against R@10 of 0.576 and MRR of 0.463 for PC, though this gap might partly be explained by the small size of the dataset. On the other hand, with the ACL-200 baseline dataset we notice that P and PC achieve identical performance, with an R@10 of 0.644 for both, showing that the model indeed learns rich semantic signals about a paper equally well from its own title and abstract as from the contexts in which it is cited. However, PP alone holds no value on its own in either dataset, and even when combined with other metapaths, it either introduces noise and significantly reduces performance or yields only a negligible improvement. On ACL-200, for instance, R@10 drops from 0.644 with P alone to 0.628 with P+PP, while on ArgKG the R@10 stayed the same for Corpus+external and improves by only 0.002 for Corpus only. In contrast, combining P with PC yields the best performance of all combinations on the ACL-200 dataset, with an R@10 of 0.670 and an MRR of 0.449, showing that the contexts in which a paper is cited, combined with its title and abstract, yield the most informative representation for the model to learn from. However, with ArgKG, we cannot draw the same conclusion as confidently, for Corpus+external, as noted above, most papers are external and thus lack a genuine P and PC distinction, while for Corpus only, the dataset is very limited in size.

6. Conclusion and Future Work

We presented an LCR approach based on a two-tower architecture that embeds citation contexts and candidate papers into a shared space, using metapath-based features fused through a transformer-based semantic fusion module. These metapath-based features are computed over ArgKG, an RDF Scholarly Knowledge Graph (SKG), dedicated to the domain of argumentation, selected because it constitutes a well-defined research community with established terminology, strong interconnections between publications, and extensive citation practices, making it a suitable and representative domain for studying LCR. Our experiments show that a paper’s own title and abstract (P) and its citation-context feature (PC) are each, on their own, strong and largely complementary sources of semantic signal, and that combining them (P+PC) yields the best overall performance. In contrast, the 1-hop reference feature (PP) holds little value on its

own and, when combined with other metapaths, either introduces noise or yields only negligible gains.

Several directions remain open for future work. First, the citation-context metapath should be investigated further, for instance by distinguishing a paper's own citation contexts (PC) from contexts in which it was cited alongside specific neighbouring papers (PCP), which our current feature set does not yet exploit. Second, we plan to compare our architecture against the prefetch-then-rerank, graph-enriched, and generative baselines discussed in Section 2 (HAtten, SymTax). Third, we intend to extend the SKG beyond the argumentation-theory domain, and to formulate SPARQL competency questions over it to validate its use as a general-purpose, queryable resource for citation-based scholarly knowledge discovery.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) solely to check grammar, spelling, and language clarity. Following the use of this tool, the author(s) reviewed and edited the generated suggestions as necessary and take full responsibility for the final content of the publication.

References

- [1] M. Färber, A. Jatowt, Citation recommendation: approaches and datasets, *International Journal on Digital Libraries* 21 (2020) 375–405.
- [2] J. Beel, B. Gipp, S. Langer, C. Breiting, Paper recommender systems: a literature survey, *International Journal on Digital Libraries* 17 (2016) 305–338.
- [3] L. Steinert, *Beyond Similarity and Accuracy: A New Take on Automating Scientific Paper Recommendations*, Verlag Dr. Hut, 2017.
- [4] D. Buscaldi, D. Dessì, E. Motta, M. Murgia, F. Osborne, D. Reforgiato Recupero, Citation prediction by leveraging transformers and natural language processing heuristics, *Inf. Process. Manage.* 61 (2024). doi:10.1016/j.ipm.2023.103583.
- [5] Q. He, J. Pei, D. Kifer, P. Mitra, L. Giles, Context-aware citation recommendation, in: *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 421–430.
- [6] Z. Medić, J. Šnajder, Improved local citation recommendation based on context enhanced with global information, in: *Proceedings of the First Workshop on Scholarly Document Processing*, Association for Computational Linguistics, Online, 2020, pp. 97–103.
- [7] N. Gu, Y. Gao, R. H. R. Hahnloser, Local citation recommendation with hierarchical-attention text encoder and SciBERT-based reranking, in: *Advances in Information Retrieval (ECIR 2022)*, volume 13185 of *Lecture Notes in Computer Science*, Springer, 2022.
- [8] T. D. Khoa, T. N. Dinh, P. Pham, B. Vo, N. N. Son, Enhancing the HAtten model for local citation recommendation using BiLSTM and attention pooling, 2026. Vol. 1, no. 1.
- [9] K. Goyal, M. Goel, V. Goyal, M. Mohania, Symtax: symbiotic relationship and taxonomy fusion for effective citation recommendation, in: *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 8997–9008.
- [10] E. Y. Çelik, S. Tekir, Citebart: learning to generate citations for local citation recommendation, in: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Suzhou, China, 2025, pp. 1703–1719.
- [11] X. Yang, M. Yan, S. Pan, X. Ye, D. Fan, Simple and efficient heterogeneous graph neural network, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [12] N. Yacoubi Ayadi, S. Bernard, R. Bossy, M. Courtin, B. G. Happi Happi, P. Larmande, F. Michel, C. Nédellec, C. Roussey, C. Faron, A unified approach to publish semantic annotations of agricultural documents as knowledge graphs, *Smart Agricultural Technology* 8 (2024) 100484. doi:<https://doi.org/10.1016/j.atech.2024.100484>.
- [13] S. Peroni, D. Shotton, The SPAR ontologies, in: *The Semantic Web – ISWC 2018*, Springer International Publishing, 2018, pp. 119–136.

- [14] S. Peroni, D. Shotton, FaBiO and CiTO: ontologies for describing bibliographic resources and citations, *Journal of Web Semantics* 17 (2012) 33–43. doi:10.1016/j.websem.2012.08.001.
- [15] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. [arXiv:2205.01833](https://arxiv.org/abs/2205.01833).
- [16] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph of research entities and claims in computer science, in: *International Semantic Web Conference*, Springer, 2022, pp. 678–696.
- [17] I. Beltagy, K. Lo, A. Cohan, 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)*, 2019, pp. 3615–3620.
- [18] F. Tan, The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics, in: *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC 08)*, 2008.