

ILCiteR: Evidence-grounded Interpretable Local Citation Recommendation

Sayar Ghosh Roy, Jiawei Han

University of Illinois Urbana-Champaign

{sayar3, hanj}@illinois.edu

Abstract

Existing Machine Learning approaches for local citation recommendation directly map or translate a query, which is typically a claim or an entity mention, to citation-worthy research papers. Within such a formulation, it is challenging to pinpoint why one should cite a specific research paper for a particular query, leading to limited recommendation interpretability. To alleviate this, we introduce the evidence-grounded local citation recommendation task, where the target latent space comprises evidence spans for recommending specific papers. Using a distantly-supervised evidence retrieval and multi-step re-ranking framework, our proposed system, ILCiteR, recommends papers to cite for a query grounded on similar evidence spans extracted from the existing research literature. Unlike past formulations that simply output recommendations, ILCiteR retrieves ranked lists of evidence span and recommended paper pairs. Secondly, previously proposed neural models for citation recommendation require expensive training on massive labeled data, ideally after every significant update to the pool of candidate papers. In contrast, ILCiteR relies *solely* on distant supervision from a dynamic evidence database and pre-trained Transformer-based Language Models without any model training. We contribute a novel dataset for the evidence-grounded local citation recommendation task and demonstrate the efficacy of our proposed conditional neural rank-ensembling approach for re-ranking evidence spans.

Keywords: Local Citation Recommendation, Evidence-Grounded, Distant Supervision

1. Introduction

A citation recommendation system retrieves a set of articles that could be cited for a given query. There are two broad types of citation recommendation tasks, namely, (a) global citation recommendation (Bethard and Jurafsky, 2010), and (b) local citation recommendation (Ebesu and Fang, 2017). A global citation recommendation model fetches possible citations for a *complete* document. In contrast, the local citation recommendation task aims to retrieve articles to cite conditioned on a much smaller text span (usually a sentence or a phrase). Fig. 1 provides a rough overview of the local citation recommendation task. Formally, the local citation recommendation task is defined as follows. Given a query q and a set of N candidate research papers $P = \{p_1, p_2, \dots, p_N\}$, produce a ranked ordering: $[p_{i_1}, p_{i_2}, \dots, p_{i_k}]$ of the top- k ($k < N$) most relevant research papers that could be cited for the query q . The query usually represents a claim or an entity mention. For example, if $q = \text{'ELMo'}$, a valid citation could be $\{\text{'title': 'Deep Contextualized Word Representations'}, \text{'year': 2018}, \dots\}$.

In this work, we introduce the evidence-grounded local citation recommendation task shifting the focus to recommendation interpretability. Typical local citation recommendation methods estimate a *direct* mapping function from the space of queries to that of research papers. Due to this formulation,

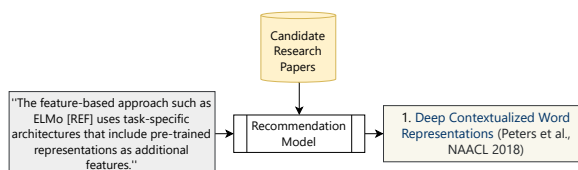


Figure 1: An overview of the local citation recommendation task for scientific research papers.

there is no explicit reasoning mechanism to identify why a particular paper should be cited for a given query.

Consider the query $q = \text{'NMT models trained on one domain tend to perform poorly when translating sentences in a significantly different domain'}$. Suppose, we find the following span e in the existing literature: $\text{'it has also been noted that NMT models trained on corpora in a particular domain perform poorly when applied in a significantly different domain [REF}_1, \text{REF}_2\text{'}$ with $[\text{REF}_1]$ referring to the paper p_1 : $\{\text{'title': 'Six challenges for neural machine translation'}, \text{'year': 2017}, \dots\}$ and $[\text{REF}_2]$ to p_2 : $\{\text{'title': 'A survey of domain adaptation for neural machine translation'}, \text{'year': 2018}, \dots\}$. We could then readily utilize e as *evidence* to recommend papers p_1 and p_2 to be cited for q . Intuitively, for a query q , a *span* of text from an existing research paper which is

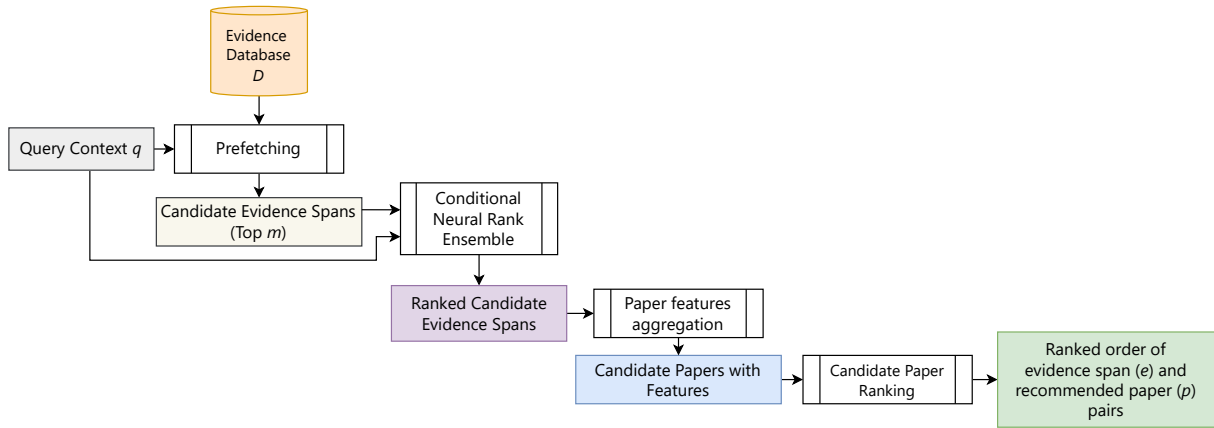


Figure 2: An overview of ILCiteR: our proposed system for evidence-grounded citation recommendation.

both similar to q and contains a citation for paper p , could readily serve as precedence to cite paper p for query q .

Based on this principle, our proposed system, namely ILCiteR, retrieves *evidence spans* (e 's) from the existing research literature that are (a) similar to the query, and (b) cite at least one paper. It re-ranks the retrieved evidence spans in decreasing order of their similarity to the query. It further ranks the individual papers associated with these ranked evidence spans. At the end of the 2-step re-ranking process, we obtain a ranked list of pairs of citation-worthy papers and appropriate evidence spans. Thus, every paper recommendation is grounded in the evidence span, which explicitly adds interpretability to each paper recommendation.

Our code and dataset are available here¹. We make the following key contributions in this work.

1. We introduce the task of evidence-grounded local citation recommendation
2. We present a novel dataset of over 200,000 unique evidence spans with corresponding sets of cited research paper and support pairs, covering three subtopics within Computer Science
3. We propose ILCiteR, a distantly supervised approach for the evidence-grounded local citation recommendation task leveraging (I) an evidence database and (II) pre-trained Transformer language models, requiring *no* explicit model training
4. Our proposed conditional neural rank ensembling approach for re-ranking evidence spans significantly improves downstream paper recommendation performance over purely lexical and semantic similarity based retrieval as well as naive rank ensembling

2. Related Work

He et al. (2010) defined the problem of local citation recommendation, laying out the distinction between global and local citation recommendation, and performed evaluations on CiteSeerX² using their prototype system. Inspired by the IBM Model 1 (Brown et al., 1993) for Machine Translation, Huang et al. (2014) proposed a local citation recommendation system considering words within the query citation context as the source language and the list of candidate references as the target language. Recently, various supervised Deep Learning models have been proposed for local citation recommendation (Ebesu and Fang, 2017; Yang et al., 2019; Wang et al., 2023; Ali et al., 2021, 2022). Medić and Snajder (2020) enhanced the query by bringing in document level information (title and abstract of the paper from which the query originated). Jeong et al. (2019) jointly considered the query's semantics and a citation network comprising various paper nodes. Gu et al. (2022) adopted a paper pre-fetching and re-ranking approach utilizing a Transformer-based Hierarchical-Attention text encoder and a SciBERT-based (Beltagy et al., 2019) paper ranking module. Other notable approaches include semantic modeling (Saier and Faerber, 2020) and recommendation ensembling (Färber and Sampath, 2020). Saier and Faerber (2020) distinguish between two broad types of citations, namely citations for named entities and citations for claims. Färber and Sampath (2020) ensembles various existing methods such as LDA (Blei et al., 2001), Doc2Vec (Le and Mikolov, 2014), Paper2Vec (Ganguly and Pudi, 2017), and HyperDoc2vec (Han et al., 2018) using a semi-genetic recommender.

Our novel problem formulation for the *evidence-grounded* recommendation task differs from all prior works on local citation recommendation.

¹<https://github.com/sayarghoshroy/ILCiteR>

²<https://citeseerx.ist.psu.edu/>

While past studies focused on building a neural mapping or translation model from the latent space of queries to that of paper nodes, our target latent space is that of evidence spans within the existing research literature.

As consequence of this updated formulation where the primary focus is interpretability, existing datasets (Färber and Jatowt, 2020; Medić and Snajder, 2020; Jeong et al., 2019; Ebesu and Fang, 2017) for local citation recommendation are unsuitable for our evidence-grounded recommendation task. We overcome this by building a comprehensive evidence database containing evidence spans and corresponding sets of cited papers with supports, and an evaluation set having queries extracted from recent papers never processed into the evidence database.

Lastly, our proposed system, ILCiteR, only leverages pre-trained Transformer-based language models with distant supervision from the dynamic evidence database and our neural recommendation pipeline does not require any form of training. Thus, we do not need to re-train ILCiteR after every update to the pool of candidate papers.

3. Problem Definition

We formally define an evidence database and our task formulation as follows.

Evidence Database Specification An evidence database (D) would contain individual records of the form: (e, P) , where e is an evidence span and P is a set of $|P|$ pairs of the type (p^j, s^j) , ($j \in \{1, 2, \dots, |P|\}$), with p^j being the metadata for the j^{th} paper in P with s^j representing its support, i.e., the number of times the paper p^j was cited for the evidence span e . Each evidence span e appears only once in the D .

The evidence database could also be regarded as map from evidence spans (e 's) to sets of paper metadata (p) and support (s) pairs. Table 1 shows some examples of simplified records from an evidence database.

Task Formulation Given an evidence database D and a query q , construct a ranked list of n unique pairs of the form (e_k^q, p_k^q) , $k \in \{1, 2, \dots, n\}$, where p_k^q is a recommended paper for q with e_k^q being the corresponding evidence span. The pairs are ranked in decreasing order of their relevance to q .

4. Approach Overview

We briefly summarize our overall approach including evidence database creation and evidence-grounded paper recommendation.

To build the evidence database D (Sec. 5), we preprocess existing research papers to fetch individual sentences containing at least a single citation (a numbered [REF] tag). From each sentence, we

extract the relevant portion of text, i.e., the evidence span, for each individual citation (Sec. 5.2). Note that there might be multiple citations for a single evidence span. This comprehensive database D acts as an external guide to provide evidence for citing a specific paper for some novel query.

For processing a query q , ILCiteR first pre-fetches a set of m candidate evidence spans (e 's) from D based on lexical similarity (Sec. 6.1). It then utilizes a two-step re-ranking procedure to construct the ranked list of evidence span and recommended citation pairs (Sec. 6.2). In the first step, it re-ranks the fetched evidence spans using a conditional neural rank ensembling approach (Sec. 6.2.1). In the second step, it ranks the collection of all candidate papers corresponding to the m retrieved evidence spans. The candidate papers are ranked considering the best observed rank of their associated evidence spans (from step one), their composite support, and recency (Sec. 6.2.2). An overview of our evidence-grounded local citation recommendation pipeline is shown in Fig. 2.

5. Evidence Database

We build an evidence database D to store evidence spans (e 's) and corresponding set of cited papers with support counts (P 's). Sec. 3 lays out the specification for D . For our experiments, we consider three popular topics within Natural Language Processing, namely, Named Entity Recognition (NER), Summarization (SUMM), and Machine Translation (MT), and build evidence databases for each.

5.1. Acquiring Research Papers

We process the available dump of the S2ORC dataset³ (Lo et al., 2020) to collect research papers tagged as 'Computer Science'. To fetch papers on a particular topic, we check whether the topic name (or its abbreviation) appears within the paper abstract. Although simple, this method promises a high recall (as a work on the topic of say, Named Entity Recognition, is highly likely to have the topic name within the abstract itself). For building our evidence database, we only consider papers for which the full-text is *publicly* available in some form and subsequently, its *parsed* full-text exists within the S2ORC dump. We processed over 100 million papers from S2ORC and collected over 20,000 CS papers on the above topics (2260 papers on NER, 8445 on SUMM, and 9522 on MT) having a publicly available full-text. From each collected paper p , we obtain a list L of individual sentences containing *at least* one citation. Thus, every sentence $s \in L$ has one or more REFS, each referring to a cited paper. We further process every s to extract one or more evidence spans.

By using paper parses from S2ORC (Lo et al., 2020), the inherent noise from the PDF parsing process is

³<https://github.com/allenai/s2orc>

Evidence span (e)	Primary cited paper metadata (p)	support (s)
'fasttext'	{'title': 'Enriching word vectors with subword information', 'year': 2016, 'authors': ...}	34
'Caffe'	{'title': 'Caffe: Convolutional Architecture for Fast Feature Embedding', 'year': 2014, 'authors': ...}	8
'Gadag's method uses R-precision metric to evaluate the result of the paraphrased sentence'	{'title': 'N-gram Based Paraphrase Generator from Large Text Document', 'year': 2016, 'authors': ...}	1
'demonstrated that AL outperformed random sampling for a simulated clinical NER task'	{'title': 'A study of active learning methods for named entity recognition in clinical text', 'year': 2015, 'authors': ...}	1
'proposed to compress LSTM-based neural machine translation models with pruning algorithms'	{'title': 'N-gram Based Paraphrase Generator from Large Text Document', 'year': 2016, 'authors': ...}	1

Table 1: Examples of evidence span and primary cited paper with their corresponding support from the evidence database D . The first evidence spans contain simple entity mentions while the last three capture various types of claims.

topic	#spans	avg #chars	avg #tokens	#cited papers
NER	23803	117.46	20.27	19041
SUMM	79345	118.89	20.37	59659
MT	108692	123.29	21.33	78743

Table 2: Key statistics on the created evidence databases for 3 popular CS subtopics

minimized since, wherever applicable, S2ORC considers the available latex source leading to a more reliable parse. Also, the [REF] to cited paper mappings for our acquired papers are of high quality since S2ORC explicitly resolves bibliographic links among in-corpus paper clusters.

5.2. Extracting Evidence Spans

From a sentence s containing at least one citation (a numbered [REF]), we extract *relevant spans* for each [REF], which serve as the evidence spans. First, for every sentence $s \in L$, we group together sets of co-occurring [REF]s into unit REFGROUPs. For example, if $s =$ "There are two broad types of text summarization approaches, namely, extractive [REF₁, REF₂, REF₃] and abstractive [REF₄]", we would group REFS 1 through 3 into REFGROUP₁ and REF₄ into REFGROUP₂. Thus, REFGROUP₁ would refer to 3 cited papers. The processed sentence s' would be: "There are two broad types of text summarization approaches, namely, extractive REFGROUP₁ and abstractive REFGROUP₂." We then extract evidence spans for each REFGROUP in s' .

We first isolate entity mentions in s' by processing its dependency parse structure. We generate a dependency parse tree for s' and begin a graph traversal from every REFGROUP-type node. Tokens visited during the traversal form the evidence span for the particular source REFGROUP. The traversal proceeds from a current node *curr* to a child node *child* only if:

1. $\text{pos}(\text{curr}) - \text{pos}(\text{child}) = 1$, where $\text{pos}(\text{token})$ is the index of *token* in the processed sentence s'
2. \exists an edge in the produced dependency parse tree from *curr* to *child* encoding one the following relationships: 'COMP' or 'AMOD'

Consider $s'_1 =$ "They used an IEX parser REFGROUP₀ to encode the ..." and $s'_2 =$ "... past few years has been focused on extractive summarization REFGROUP₂ ...". With the above rules, we capture 'IEX Parser' as the evidence span with citation REFGROUP₀ and 'extractive summarization' as that with citation REFGROUP₂.

In addition to dependency-parse based extraction, we extract evidence spans based on the sequence of tokens in s' . We split s' into contiguous spans based on positions of REFGROUPs, and map each split span to its antecedent REFGROUP. Lastly, we use two additional conditions to include the entire sentence as a possible evidence span:

1. The REFGROUP occurs at the sentence end. If $s'_3 =$ "Context embeddings were generated using Sentence Transformers REFGROUP₀", REFGROUP₀ could either be a citation for (a) 'Sentence Transformers' or (b) the complete sentence
2. Only *one* REFGROUP exists within s' . If $s'_4 =$ "They used ROUGE and METEOR metrics REFGROUP₀ for evaluating their models", it would be beneficial to consider the full sentence as an evidence span for REFGROUP₀.

Note that we apply the above dependency parse and token sequence based extraction rules in a hierarchical fashion. As an example, for $s'_5 =$ "They used BERT REFGROUP₀, a popular Large Language Model REFGROUP₁, to generate text embeddings REFGROUP₂", we have the following evidence spans (e 's) and their corresponding sets of cited papers (captured within the REFGROUPs):

'BERT' \rightarrow REFGROUP₀
 'Large Language Model' \rightarrow REFGROUP₁
 "to generate text embeddings" \rightarrow REFGROUP₂
 Complete Sentence \rightarrow REFGROUP₂

5.3. Populating Evidence Database

We process every sentence (containing at least one citation) from each collected paper to extract evidence spans (e_i 's), $i \in \{1, 2, \dots, |D|\}$ and corresponding REFGROUPs. We then resolve REFGROUPs

Query q_1	'In the first step, we only look at one Transformer block, and describe how to learn the position representation driven by a dynamical system ...'
Retrieved Evidence e_1	'In the simplest case we use word embeddings and add position encodings to them; we use ...'
Recommendation p_1	{'title': 'Attention is all you need', 'year': 2017, 'authors': ...}
Query q_2	'FastAlign'
Retrieved Evidence e_1	'FastAlign'
Retrieved Evidence e_2	'The alignment is induced with FastAlign'
Retrieved Evidence e_3	'One such method is the FastAlign word-alignment model'
Recommendation p_1	{'title': 'A simple, fast, and effective reparameterization of ibm model 2', 'year': 2013, ...}

Table 3: Demonstration of our evidence-grounded paper recommendation system – For a particular query, ILCiteR presents a ranked ordering of retrieved evidence span and recommended citation pairs.

into their component REFS. We populate the evidence database D using pairs of e_i 's and sets of pairs of cited paper (p_i^j) and support (s_i^j), $j \in \{1, 2, \dots, |P_i|\}$. We fetch papers within P_i from the resolved REFS' metadata in S2ORC. The support count s_i^j contains the number of times we observed p_i^j being cited for the evidence span e_i . Within D , e_i 's are unique. Therefore, D could be regarded as a map from evidence spans (e 's) to sets of cited paper and support pairs (P 's). During construction, we do not perform any additional semantic conflation of e_i 's; only occurrences of matching evidence spans are continually conflated.

5.4. Dataset Statistics

We provide key statistics of our evidence databases created for the 3 CS subtopics in table 2. Also, as an illustration, we present some evidence spans and their primary cited article with corresponding support in table 1. Here, we only show the most prominent cited article from the set P with its support. The first two evidence spans in table 1 are simple entity mentions while the last three represent various types of claims. We make our evidence databases (and evaluation splits) publicly available⁴.

6. Evidence-grounded Recommendation

In this section, we describe our overall methodology for evidence-grounded paper recommendation. Our task formulation can be found in Sec. 3. Given a query q , we first pre-fetch a set of m candidate evidence spans from D (Sec. 6.1). We then adopt a conditional neural rank ensembling approach to re-rank pre-fetched evidence spans based on their similarity to q (Sec. 6.2.1). In that, we conditionally ensemble a ranking of evidence spans based on semantic similarity only for lengthier queries. Lastly, we rank the collection of candidate papers based on both the relevance of their underlying evidence span and other paper-specific features (Sec. 6.2.2). An overview of our recommendation methodology, namely ILCiteR, is presented in Figure. 2.

⁴<https://github.com/sayarghoshroy/ILCiteR>

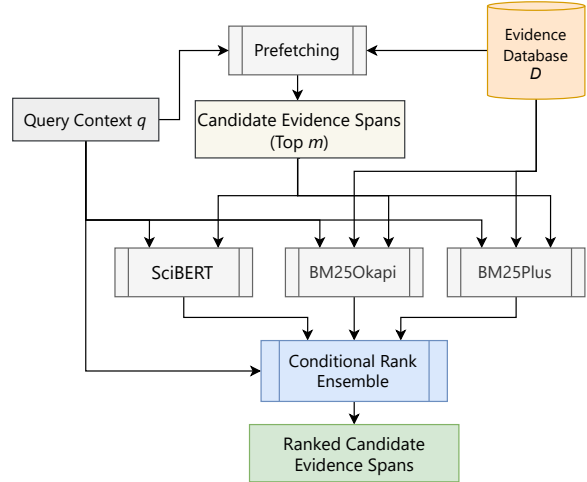


Figure 3: Conditional neural rank ensembling – re-rank candidate evidence text spans based on lexical and semantic similarity to the query q .

6.1. Pre-fetching

To reduce the search space of evidence spans, we first prefetch a set of ≤ 100 candidate evidence spans from the evidence database D that have the highest lexical similarity to the query q . These candidate evidence spans are then re-ranked in order of their similarity to the query q during the conditional rank ensembling step (Sec. 6.2.1).

Given a query text span q , we score each evidence span (e) in the evidence database D using Okapi-BM25 (Robertson and Zaragoza, 2009) and BM25Plus (Lv and Zhai, 2011). We formally write out the Okapi-BM25 score and the BM25plus score as follows.

Okapi-BM25 Given a query q with tokens q_1, q_2, \dots, q_l , the Okapi-BM25 score for a particular evidence span e can be written as: $\text{BM25Okapi_score}(e, q) = \sum_{i=1}^l \text{IDF}(q_i) \cdot \frac{f(q_i, e) \cdot (k_1 + 1)}{f(q_i, e) + k_1 \cdot (1 - b + b \cdot \frac{|e|}{a})}$. Here, $f(q_i, e)$ equals the number of times the token q_i is found within the evidence span e , $|e|$ equals the number of tokens within the evidence span e , and a is the average token length of evidence spans within the evidence database D . k_1 and b are hyperparameters.

Also, $\text{IDF}(q_i) = \ln\left(\frac{|D| - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$, where $n(q_i)$

Topic	Method	MRR	R@1	R@3	R@5	R@10
NER	BM25Okapi	0.34558	0.260	0.380	0.430	0.512
	BM25Plus	0.34735	0.262	0.386	0.436	0.510
	Conditional Ensembling	0.35155	0.266	0.390	0.438	0.514
SUMM	BM25Okapi	0.37052	0.282	0.410	0.458	0.530
	BM25Plus	0.37324	0.284	0.410	0.460	0.540
	Conditional Ensembling	0.37687	0.290	0.414	0.458	0.542
MT	BM25Okapi	0.43568	0.332	0.506	0.546	0.620
	BM25Plus	0.44071	0.338	0.500	0.558	0.636
	Conditional Ensembling	0.44456	0.342	0.516	0.556	0.628

Table 4: Performance of proposed conditional neural rank ensembling approach and lexical similarity based baselines. In general, conditional ensembling *significantly* improves recommendation performance.

Topic	Method	MRR	R@1	R@3	R@5	R@10
NER	SciBERT only	0.26246	0.188	0.268	0.342	0.428
	Ensemble BM25Okapi + SciBERT	0.31272	0.228	0.334	0.404	0.506
	Ensemble BM25Plus + SciBERT	0.30772	0.220	0.338	0.400	0.496
	Ensemble BM25s + SciBERT	0.31502	0.222	0.344	0.418	0.510
	Conditional Ensemble	0.35155	0.266	0.390	0.438	0.514
SUMM	SciBERT only	0.31555	0.234	0.346	0.392	0.496
	Ensemble BM25Okapi + SciBERT	0.34795	0.266	0.374	0.414	0.520
	Ensemble BM25Plus + SciBERT	0.35251	0.274	0.370	0.408	0.526
	Ensemble BM25s + SciBERT	0.35822	0.270	0.392	0.438	0.530
	Conditional Ensemble	0.37687	0.290	0.414	0.458	0.542
MT	SciBERT only	0.38589	0.288	0.430	0.488	0.560
	Ensemble BM25Okapi + SciBERT	0.42961	0.342	0.458	0.514	0.614
	Ensemble BM25Plus + SciBERT	0.44116	0.352	0.472	0.542	0.630
	Ensemble BM25s + SciBERT	0.44305	0.356	0.474	0.540	0.636
	Conditional Ensemble	0.44456	0.342	0.516	0.556	0.628

Table 5: Metrics achieved by re-ranking using SciBERT-based semantic similarity only, naive ensemble of lexical and semantic similarity ranks, and our proposed conditional neural rank ensembling approach.

represents the number of evidence spans within D that contain q_i .

BM25plus BM25Plus has an additional free parameter, δ , with a default value of 1.0. The BM25Plus score can be written as: $\text{BM25Plus_score}(e, q) = \sum_{i=1}^l \text{IDF}(q_i) \cdot \left(\frac{f(q_i, e) \cdot (k_1 + 1)}{f(q_i, e) + k_1 \cdot (1 - b + b \cdot \frac{|e|}{a})} + \delta \right)$.

With standard BM25Okapi (Robertson and Zaragoza, 2009), the term frequency normalization by the length of the evidence span would not be perfectly lower bounded. Therefore, Okapi BM25 would tend to score longer evidence spans unfairly as having a relevancy score similar to that of shorter evidence spans not containing the query tokens. The BM25Plus algorithm (Lv and Zhai, 2011) resolves this drawback by introducing the parameter δ .

We consider an evidence span from D as a *candidate* evidence span for query q if it was among the top-50 highest scored evidence spans by either BM25Okapi or BM25Plus. We found that pre-fetching 50 evidence spans for each of BM25Okapi and BM25Plus provided a reasonably high recall and the effects of increasing this threshold beyond 50 were negligible. We thus create a set of m candidate evidence spans $E^q = \{e_1^q, e_2^q, \dots, e_m^q\}$ from D .

Note that the pre-fetching step in our evidence-grounded citation recommendation task differs from traditional pre-fetching in past works on local citation recommendation (Gu et al., 2022; Ebesu and Fang, 2017). Earlier works directly pre-fetched can-

didate *research articles*. In contrast, we fetch *evidence spans* that are relevant to the query q .

6.2. Two-step Ranking

We first re-rank evidence spans obtained from the pre-fetching step using our conditional neural rank ensembling approach and subsequently rank the collection of candidate papers.

6.2.1. Conditional Rank Ensembling

We formulate a conditional neural rank ensembling method to re-rank pre-fetched candidate evidence spans (evidence spans in $E^q = \{e_1^q, e_2^q, \dots, e_m^q\}$). Our design choice is based on the following observation. For smaller and simple queries, utilizing overlap of key lexical items is important since shorter queries typically contain mentions of notable named entities. However, when the query is long and complex, effectively capturing its semantics becomes crucial. To this end, for short queries, we use a rank ensemble of BM25Plus (Lv and Zhai, 2011) and BM25Okapi (Robertson and Zaragoza, 2009). When the query length exceeds a threshold, we leverage a pre-trained SciBERT (Beltagy et al., 2019) model. SciBERT is a Transformer-based (Vaswani et al., 2017) Large Language Model (similar to BERT (Devlin et al., 2019)) pre-trained on large corpora of text from scientific research papers. We define the semantic similarity score for e_i^q ($i \in \{1, 2, \dots, m\}$) as the cosine similarity between the 768-dimensional [CLS] embeddings of the query q and e_i^q .

Formally, if E_q and E_e are the SciBERT generated embeddings of the initial [CLS] tokens of the input

query q and a candidate evidence span e , the semantic similarity score between q and e could be written as: $\text{semantic_score}(q, e) = \frac{E_q \cdot E_e}{|E_q| \cdot |E_e|}$, where $E_q, E_e \in \mathbb{R}^{768}$.

We ensemble the SciBERT-based semantic similarity rank with the BM25Plus rank to produce the ranked ordering for lengthier queries. We thus obtain an evidence rank r_i^q for each e_i^q . At the end of this step, we sort e_i^q 's based on r_i^q 's in ascending order and pass the rank-sorted order: $[e_i^q, e_i^q, \dots, e_i^q]$ to the final paper recommendation step. A rough overview of our conditional rank ensembling approach is shown in Fig. 3.

6.2.2. Recommending Papers with Evidence

From the previous evidence re-ranking step, we obtain the ranked order of m candidate evidence text spans for query q : $E^q = [e_1^q, e_2^q, \dots, e_m^q]$, with P_i^q being the set of pairs of research paper and support associated to e_i^q obtained from the evidence database D ($i \in \{1, 2, \dots, m\}$). Here, for any $j < k$, e_j^q would be more similar to q than e_k^q .

For each e_i^q , we obtain the associated set of $|P_i^q|$ cited papers with their supports from D :

$$P_i^q = \{(1p_i^q, 1s_i^q), (2p_i^q, 2s_i^q), \dots, (|P_i^q|p_i^q, |P_i^q|s_i^q)\}$$

With $j s_i^q$ being the support for paper $j p_i^q$ in P_i^q , $j \in \{1, 2, \dots, |P_i^q|\}$. We now rank the set of all papers among pairs in all P_i^q 's, accounting for: (1) the highest observed rank of a particular paper's associated evidence span, (2) the paper's composite support, and (3) its recency. We formally define each of these components as follows:

1. The lowest observed rank of a paper p :

$$p_r = \min_{i \in \{1, 2, \dots, m\}} R_i(p),$$

where $R_i(p) = i$ if $\exists s$ s.t. $(p, s) \in P_i^q$, else $R_i(p) = \infty$

2. The total support for a paper:

$$p_s = \sum_{i=1}^m S_i(p)$$

where $S_i(p) = s$ if $\exists s$ s.t. $(p, s) \in P_i^q$, else $S_i(p) = 0$

3. The recency of a paper (rec_p), which in our case, simply holds the year of publication

We rank papers in the following order of precedence: (1) p_r (lower is better), (2) p_s (higher is better), (3) rec_p (higher is better).

For every evidence span, we obtain a ranked order of cited papers. We aggregate this into a ranked ordering of evidence span and recommended citation pairs. Thus, each paper recommendation for a query q is grounded in the accompanying evidence span, which provides explicit interpretability. Our evidence database could be dynamically up-scaled to include evidence spans from newly published papers.

7. Experiments and Analysis

We evaluate the performance of our evidence-grounded recommendation system and present some key insights.

7.1. Evaluation Set Creation

For a particular topic, we first create an evaluation candidates set E_c containing candidate queries and their ground truth cited papers extracted⁵ from the 200 most recent papers on that topic which were not considered while creating D . This ensures that D does not contain any mappings extracted from papers that contribute to the evaluation set, reproducing a real-world evaluation setting. We then create our final evaluation set E by picking 500 datapoints (d 's) from E_c satisfying the following condition:

For some $d = (q, \{p'_1, p'_2, \dots, p'_a\}) \in E$, there is at least one pair of evidence span e and a set P of pairs of cited paper and support, with $P = \{(p^1, s^1), (p^2, s^2), \dots, (p^b, s^b)\}$ such that $(e, P) \in D$ and for some (i, j) with $i \in \{1, 2, \dots, a\}$ and $j \in \{1, 2, \dots, b\}$, $p'_i = p^j$. Here, q is the query and $\{p'_1, p'_2, \dots, p'_a\}$ is the set of ground truth cited papers for q .

This ensures that for every datapoint d in E , at least one ground truth paper exists within the set of all candidate papers in D .

7.2. Interpretable Paper Recommendation

We show two examples to demonstrate the working of our evidence-grounded local citation recommendation system (ILCiteR) in table 3. In table 4, we compare the performance of our proposed conditional neural rank ensembling approach for re-ranking evidence spans with lexical similarity based baselines. We use the Mean Reciprocal Rank (MRR) and Recall@ N measures (with $N = 1, 3, 5, 10$) as automatic evaluation metrics.

From table 3, for q_1 , we see that our system extracts a highly relevant evidence span (describing the Transformer model with a focus on position embeddings). q_2 demonstrates that handling entity mentions is relatively straightforward for our system – several pieces of evidence spans from existing papers are provided to suggest why p_2 should be cited for the entity mention 'FastAlign'.

From table 4, we see that our conditional neural rank ensembling approach improves recommendation performance over BM25Okapi and BM25Plus on Recall@1, Recall@3, and MRR, for all topics. Overall, we see performance improvements on all metrics with conditional rank ensembling with the exception of R@5 for SUMM, MT and R@10 for MT (where lexical similarity using BM25Plus leads to better recommendations). We performed t-tests

⁵Applying the procedure detailed in Sec. 5.2

Query q	'exploit monolingual data in two ways: through self-learning by forward-translating the monolingual source data to create synthetic parallel data ...'
Method	BM25 ensemble (BM25Okapi + BM25Plus)
Retrieved Evidence e_1	'By applying a self-training scheme, the pseudo parallel data were obtained by automatically translating the source-side monolingual corpora ...'
Model Reco. p_1	{'title': 'Transductive learning for statistical machine translation', 'year': 2007, ...}
Method	Conditional Neural Rank Ensembling
Retrieved Evidence e_1	'... employing a self-learning algorithm to generate pseudo data, while the second is using two NMT models to predict the translation and to reorder the source-side monolingual sentences'
Model Reco. p_1	{'title': 'Exploiting source-side monolingual data in neural machine translation', 'year': 2016, ...}

Table 6: Conditionally utilizing semantic similarity scoring (using SciBERT) for lengthier contexts helps retrieve accurate evidence text spans leading to better recommendations.

Query q_1	'The synchronous context-free grammars introduced by are transduction grammars whose productions have some restrictions'
Retrieved Evidence e_1	'The model is based on inversion transduction grammars (ITGs), a variety of synchronous context free grammars (SCFGs)'
Retrieved Evidence e_2	'Among the grammar formalisms successfully put into use in syntax based SMT are synchronous context-free grammars (SCFG) and synchronous tree substitution grammars (STSG)'
Model Reco. p_1	{'title': 'Stochastic inversion transduction grammars and bilingual parsing of parallel corpora', 'year': 1997, ...}
True Reco. p_g	{'title': 'Hierarchical phrase-based translation', 'year': 2007, ...}

Table 7: Instance of an error made by our model. Though the ground truth recommendation p_g is more appropriate, evidence spans retrieved by our model are relevant, p_1 is a sensible recommendation for q_1 .

which showed that our proposed conditional ensembling approach improves scores significantly over unsupervised BM25Okapi’s ranking with a p -value of 0.08⁶ for MRR and Recall@5 metrics.

7.3. Ablation Studies

We perform several ablations to showcase the efficacy of our proposed conditional neural rank ensembling approach. In table 5, we present metrics for re-ranking using SciBERT-based semantic similarity only, naive ensemble of lexical and semantic similarity ranks, and our proposed conditional rank ensembling. Firstly, we notice that re-ranking with plain SciBERT massively drops recommendation performance. This suggests that lexical similarity is a useful signal for evidence span re-ranking. Secondly, we observe that our conditional ensembling approach consistently produces performance improvements on all metrics across all topics (R@1 and R@5 for MT are the only exceptions where naive rank ensembling using SciBERT and lexical similarity achieves better numbers). We performed t-tests which showed that our proposed conditional rank ensembling approach significantly improved model performance over re-ranking evidence spans purely using SciBERT, and a naive ensemble of SciBERT and BM25, with a p -value of 5%⁷. In general, unconditionally ensembling with semantic similarity for queries of all lengths worsens performance compared to using semantic similarity ranks only when necessary, i.e., when the query is lengthy enough. We observed that a length threshold ≈ 2.5 times the average token length for evidence spans (≈ 50

tokens) was optimal for ensembling with semantic similarity-based ranks.

In table 6, we present a concrete example of how conditionally leveraging semantic similarity scoring (using SciBERT) for lengthy contexts helps evidence span retrieval. For q , we are able to extract a moderately relevant evidence span using ensembled lexical similarity. But although the suggested evidence span shares multiple common entities with the query, the downstream recommendation is incorrect (recommends a paper on statistical MT). Without the semantic similarity based signal, the ground truth paper does not appear among the top-3 recommendations. Here, conditional rank ensembling captures the semantics of q and accurately retrieves an evidence span pointing to the ground truth paper on neural machine translation.

7.4. Error Analysis

Consider the query q_1 from table 7. The top-2 evidence spans retrieved by our model were e_1 and e_2 with the recommended citation being p_1 , which was incorrect. We observe that the retrieved evidences are relevant to q_1 , and p_1 could possibly be cited for such a context. However, the ground truth citation p_g is more appropriate for q_1 , especially since the underlying topic is MT. Deciding whether to cite a paper on the method itself (in this case, Stochastic ITGs) or a paper on how the method is applied to a specific task (here, Machine Translation) would require additional input fields like the exact topic or area of the research paper from which the query originates. Handling such cases is tricky for our system since we have query as the *only* input field. Note that our model suggests p_g as recommendation #9 with the following evidence span: 'use some grammar at their core, for instance: (a) synchronous context-free grammars (SCFG)'.

⁶An exception being Recall@5 for topic: NER where conditional ensembling led to slightly worse results.

⁷An exception being topic: MT where the naive ensemble’s MRR was similar to our proposed approach.

Query q_2	'bidirectional long short-term memory with conditional random field model (BiLSTM-CRF) exhibited promising results'
Retrieved Evidence e_1	'Among others, the model of bidirectional Long Short Term Memory with a conditional random field layer (BiLSTM-CRF), exhibits promising results'
Retrieved Evidence e_2	'which is based on a bidirectional long short-term memory (LSTM) network with a conditional random field (CRF) over the output layer'
Retrieved Evidence e_3	'Our framework is based on a bidirectional long short-term memory network with a conditional random fields (BiLSTM-CRFs) for sequence tagging'
Model Reco. p_1	{'title': 'Neural architectures for named entity recognition', 'year': 2016, 'venue': 'NAACL', ...}
Model Reco. p_2	{'title': 'Bidirectional LSTM-CRF models for sequence tagging', 'year': 2015, ...}
True Reco. p_g	{'title': 'End-to-end sequence labeling via bi-directional lstm-cnns-crf', 'year': 2016, 'venue': 'ACL', 'authors': ...}

Table 8: Instance of an error made by our model – though the ground truth recommendation p_g does not match the primary model recommendation p_1 , p_1 is an appropriate recommendation for q_1 as it was cited for claims that are very similar to q_2 .

Query q_3	'The use of tokens to condition the output of NMT started with the multilingual models'
Retrieved Evidence e	'... the idea to multilingual NMT by concatenating parallel data of various language pairs and marking the source ...'
Model Reco. p	{'title': 'Google's multilingual neural machine translation system: Enabling zero-shot translation', 'year': 2017, 'venue': 'TACL', ...}
True Reco. p_g	{'title': 'Google's multilingual neural machine translation system: Enabling zero-shot translation', 'year': 2016, ...}

Table 9: Recommendation error due to multiple versions of the same paper.

We have a similar case in table 8. Here, our model retrieves highly relevant evidence spans (e_1 , e_2 , and e_3) and recommends papers p_1 and p_2 . We observe that the evidence span e_1 is very similar to q_2 and p_1 is indeed a valid citation for q_2 . Here, the ground truth citation points to a paper on the more general use of LSTM-CNNs-CRFs (in end-to-end sequence-labeling) while p_1 is a paper on the topic of interest, i.e., neural models for Named Entity Recognition, from the same year. p_2 , which is ranked second by the model, is very similar to p_g . And one could even argue that since p_2 is about LSTM-CRFs for sequence tagging, it is more appropriate for q_2 compared to p_g which is a work on using LSTM-CNNs-CRFs for general sequence labeling. Thus, we observe that even when the model recommendations for a query do not match the ground truth, the retrieved evidence spans are highly relevant and the model recommended papers are quite appropriate.

We also find another cause of error due to multiple versions of the same paper (Table 9). The model recommendation p was deemed as a failure (as we do not conflate paper nodes from different years) since the ground truth citation was a pre-print version of the same paper from an earlier year.

7.5. Implementation Details

We use publicly available implementations of BM25⁸, with default values for all hyperparameters. For generating dependency parses, we utilized spaCy's⁹ default dependency parser based on Honnibal and Johnson (2015) and Nivre and Nilsson (2005). We used HuggingFace's¹⁰ implementation of SciBERT-cased. Our codebase for ex-

⁸https://github.com/dorianbrown/rank_bm25

⁹<https://spacy.io/>

¹⁰<https://huggingface.co/>

perimentation and analysis has been shared¹¹.

8. Conclusion

In this work, we introduced the task of evidence-grounded local citation recommendation, with an explicit focus on recommendation interpretability. We contribute a novel dataset with over 200,000 evidence spans and sets of cited paper and support pairs, covering three subtopics within Computer Science. Our proposed recommendation system, ILCiteR, solely uses distant supervision from a dynamic evidence database and pre-trained Transformer-based language models, requiring *no* explicit training. Our findings demonstrate how a conditional neural rank ensembling approach, which utilizes semantic similarity-based ranks for lengthier queries, significantly improves downstream paper recommendation performance over purely lexical and semantic similarity based retrieval and naive rank ensembling techniques. In future, we plan to study the effects of jointly considering distant supervision from an evidence database and a secondary database of claims, findings, contributions, and named entities extracted from scientific papers for the evidence-grounded local citation recommendation task.

9. Acknowledgements

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and IN-CAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329.

¹¹<https://github.com/sayarghoshroy/ILCiteR>

10. Bibliographical References

- Z. Ali, Guilin Qi, Pavlos Kefalas, Shah Khusro, Inayat Khan, and Khan Muhammad. 2022. [Sprsmn: scientific paper recommendation employing specter with memory network](#). *Scientometrics*, 127:6763–6785.
- Zafar Ali, Guilin Qi, Khan Muhammad, S. Bhattacharyya, Irfan Ullah, and Waheed Ahmad Abro. 2021. [Citation recommendation employing heterogeneous bibliographic network embedding](#). *Neural Computing and Applications*, 34:10229 – 10242.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. [Neural machine translation by jointly learning to align and translate](#). *CoRR*, abs/1409.0473.
- 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.
- Steven Bethard and Dan Jurafsky. 2010. [Who should i cite: learning literature search models from citation behavior](#). *Proceedings of the 19th ACM international conference on Information and knowledge management*.
- David M. Blei, A. Ng, and Michael I. Jordan. 2001. [Latent dirichlet allocation](#). *J. Mach. Learn. Res.*, 3:993–1022.
- Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. [The mathematics of statistical machine translation: Parameter estimation](#). *Comput. Linguistics*, 19:263–311.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *ArXiv*, abs/1810.04805.
- 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, SIGIR '17*, page 1093–1096, New York, NY, USA. Association for Computing Machinery.
- Michael Färber and Adam Jatowt. 2020. [Citation recommendation: approaches and datasets](#). *International Journal on Digital Libraries*, 21:375 – 405.
- Michael Färber and Ashwath Sampath. 2020. [Hybridcite: A hybrid model for context-aware citation recommendation](#). *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*.
- Soumyajit Ganguly and Vikram Pudi. 2017. [Paper2vec: Combining graph and text information for scientific paper representation](#). In *European Conference on Information Retrieval*.
- Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. [Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking](#). In *Advances in Information Retrieval*, pages 274–288, Cham. Springer International Publishing.
- Ben Hachey. 2009. [Multi-document summarisation using generic relation extraction](#). In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 420–429, Singapore. Association for Computational Linguistics.
- Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. 2018. [hyperdoc2vec: Distributed representations of hypertext documents](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2384–2394, Melbourne, Australia. Association for Computational Linguistics.
- Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. 2010. [Context-aware citation recommendation](#). In *The Web Conference*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Computation*, 9:1735–1780.
- Matthew Honnibal and Mark Johnson. 2015. [An improved non-monotonic transition system for dependency parsing](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. 2014. [Refseer: A citation recommendation system](#). *IEEE/ACM Joint Conference on Digital Libraries*, pages 371–374.
- Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Lucy Park, and Sungchul Choi. 2019. [A context-aware citation recommendation model with bert and graph convolutional networks](#). *Scientometrics*, 124:1907 – 1922.
- Thomas Kipf and Max Welling. 2016a. [Semi-supervised classification with graph convolutional networks](#). *ArXiv*, abs/1609.02907.

- Thomas Kipf and Max Welling. 2016b. [Variational graph auto-encoders](#). *ArXiv*, abs/1611.07308.
- Quoc V. Le and Tomas Mikolov. 2014. [Distributed representations of sentences and documents](#). In *International Conference on Machine Learning*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. [S2ORC: The semantic scholar open research corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.
- Yuanhua Lv and ChengXiang Zhai. 2011. [Lower-bounding term frequency normalization](#). In *International Conference on Information and Knowledge Management*.
- Zoran Medić and Jan Snajder. 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, Online. Association for Computational Linguistics.
- Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6826–6833.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. [Efficient estimation of word representations in vector space](#). In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.
- Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005. [Evaluating automatic summaries of meeting recordings](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 33–40, Ann Arbor, Michigan. Association for Computational Linguistics.
- Ani Nenkova and Rebecca Passonneau. 2004. [Evaluating content selection in summarization: The pyramid method](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Joakim Nivre and Jens Nilsson. 2005. [Pseudo-projective dependency parsing](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: bm25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3:333–389.
- Tarek Saier and Michael Faerber. 2020. [Semantic Modelling of Citation Contexts for Context-Aware Citation Recommendation](#), pages 220–233.
- Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. [TaxoClass: Hierarchical multi-label text classification using only class names](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online. Association for Computational Linguistics.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NIPS*.
- Shanshan Wang, Xiaohong Li, Jin Yao, and Ben You. 2023. [Multi-information fusion based on dual attention and text embedding network for local citation recommendation](#). *Advances in Computational Intelligence*, 3.
- Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. [Universal decompositional semantics on Universal Dependencies](#). In

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Libin Yang, Zeqing Zhang, Xiaoyan Cai, and Tao Dai. 2019. [Attention-based personalized encoder-decoder model for local citation recommendation](#). *Computational Intelligence and Neuroscience*, 2019.