Recommending Missed Citations Identified by Reviewers: A New Task, Dataset and Baselines

Kehan Long[1], **Shasha Li**[1,2,†] , **Pancheng Wang**[1], **Chenlong Bao**[1], **Jintao Tang**[1,†], **Ting Wang**[1,†]
[1]School of Computer, National University of Defense Technology
[2]Key Laboratory of Software Engineering for Complex Systems, National University of Defense Technology
lishasha198211@163.com
{longkehan15, wangpancheng13, baochenglong, tangjintao, tingwang}@nudt.edu.cn

Abstract

Citing comprehensively and appropriately has become a challenging task with the explosive growth of scientific publications. Current citation recommendation systems aim to recommend a list of scientific papers for a given text context or a draft paper. However, none of the existing work focuses on already included citations of full papers, which are imperfect and still have much room for improvement. In the scenario of peer reviewing, it is a common phenomenon that submissions are identified as missing vital citations by reviewers. This may lead to a negative impact on the credibility and validity of the research presented. To help improve citations of full papers, we first define a novel task of Recommending Missed Citations Identified by Reviewers (RMC) and construct a corresponding expert-labeled dataset called CitationR. We conduct an extensive evaluation of several state-of-the-art methods on CitationR. Furthermore, we propose a new framework RMCNet with an *Attentive Reference Encoder* module mining the relevance between papers, already-made citations, and missed citations. Empirical results prove that RMC is challenging, with the proposed architecture outperforming previous methods in all metrics. We release our dataset and benchmark models to motivate future research on this challenging new task.

Keywords: Citation Recommendation, Dataset, Missed Citations, Peer Review, Scientific Document Encoders

1. Introduction

Citations are essential in many writing scenarios, especially in academic writing (Färber and Jatowt, 2020). Academic researchers constantly refer to credible literature in their research fields to support their arguments and provide readers with a plausible explanation of the content of their manuscripts. Those reliable sources should be correctly cited, and this is where the concept and significance of citations enter the picture (Pears and Shields, 2013). A broad and critical literature survey is an essential component of any scientific research, as it provides a foundation for developing research questions, designing experiments, and interpreting results (Booth et al., 2012). However, the exponential growth of scientific publications has made it increasingly challenging for researchers to conduct thorough literature reviews and make comprehensive and appropriate citations.

The task of citation recommendation (CR) has been introduced by (He et al., 2010), aiming to automatically recommend appropriate citations for a given text context or a draft paper. Here we denote a scientific paper with only content as a *draft paper* or *manuscript* and a paper containing both content and complete citations as a *submission paper* or *full paper*. Specifically, researchers collect published scientific papers and take already cited papers in

Review:
......In the abstract the author mentioned "state-of-the-art" LSTM for 2D navigation/localization problems, I am not sure where is this coming from.For example in [1], CMP performs better than LSTM. **In the embodied AI community, there are a lot of works on cognitive mapping (see [1][2][3]), but this work didn't mention those, or compare with those ....**
References
[1] Gupta, Saurabh, et al. "Cognitive mapping ......
[2] Zhang, Jingwei, et al. "Neural slam: Learning ......
[3] Chaplot, Devendra Singh, et al. "Learning to ......

Figure 1: An example of missed citations identified by reviewers extracted from https://openreview.net/forum?id=R612wi_C-7w. Italic and colored texts represent papers mentioned in reviews, among which, enclosed by the dashed border, are those reviewers recommend citing. Underlined and bolded texts indicate reasons why those citations are missed and necessary.

the reference section as labels for training models to predict. Some works (Huang et al., 2015; Bhagavatula et al., 2018; Färber and Sampath, 2020) rely on semantic representations of the content of draft papers learnt by neural networks to generate recommendations. Some works (Ren et al., 2014; Xie et al., 2021; Wang et al., 2022) further adopt embeddings learnt from graphs to recommend citations. Although some of those studies construct a citation network based on partial citations of input papers, they equally treat these citations and are unaware of critical citations that significantly impact the research's foundation. In all, previous studies

† Corresponding authors.

LREC-COLING 2024, pages 13699–13711

mainly focus on manuscripts or give equal weights to all citations, neglecting the potential imperfections of already included citations and the negative impact of missing vital citations on the research's comprehensiveness and innovativeness.

It is not a rare case that submissions are considered to lack vital citations by reviewers, as illustrated in Figure 1. The omission of these essential papers in the submission often leads to deficiencies in terms of credibility, comprehensiveness, and innovation. That is to say, citations recommended by reviewers may have a great influence on the research foundation of submissions. Inspired by this common phenomenon in the peer review process, we formulate (§3) and study a novel task of Recommending Missed Citations Identified by Reviewers (RMC). Previous CR tasks generate recommendations by taking already cited papers provided by authors as labels for training. RMC enhances the reference sections of submissions by considering citations recommended by reviewers as golden labels. Considering the data flow, CR is similar to RMC. However, they differ in several aspects:

- **Recommend for**: Local CR recommends citations for a text context where specific citations should be made. Global CR takes a draft paper with no or partial citations as input. RMC identifies missed citations of a full paper with complete citations provided by the author.
- **Guarantee**: The golden citations of both CR tasks are obtained from the reference sections in papers, which are written and guaranteed by corresponding authors. In RMC, golden citations are identified by experienced experts from top-tier conferences.
- **Relevance**: For local CR, golden citations are highly related to input texts since that is where those citations should be made. Global CR aims to recommend a whole reference list, with some citations highly related and some less important and even replaceable. For RMC, all golden citations are considered highly related and important by reviewers.

Additionally, we curate a novel high-quality dataset, CitationR, by extracting recommended citations in reviews from NeurIPS and ICLR (§4). In total, we collect 76,143 official reviews and 21,598 submissions, among which around 35% of submissions are identified as lacking citations. Moreover, to better replicate the actual situation in which researchers search for papers to cite, we establish a larger and more challenging version of CitationR. This version includes additional 40,810 papers published in top venues that reviewers frequently recommend citations from.

We adapt and evaluate a wide range of existing state-of-the-art methods on our dataset and task formation, including four groups (§6.1): (1) traditional sparse retrieve models, (2) traditional citation recommendation models, (3) pre-trained scientific document encoding models, and (4) large language models. We further propose a novel framework RMCNet based on an *Attentive Reference Encoder (ARE)* module and contrastive learning objectives to solve this task. One key challenge in RMC is to assess the correlation between missed citations, already included citations and the content of submissions. Our designed *ARE* aims to effectively fuse both the content and reference sections of submissions. We use the "Citation-Informed Transformer" (Ostendorff et al., 2022) as the text encoder in our framework, which can be readily adapted to other encoders.

In experiments, we show that our method outperforms all previous methods in all metrics on CitationR. Ablation and parameter studies further prove the effectiveness of our approach. However, compared to traditional CR datasets, the performances of all methods on CitationR are much worse, highlighting the complexity of CitationR.

To conclude, our contributions are threefold:

- We introduce a new challenging task of recommending missed citations identified by reviewers, which is built from a common phenomenon in the peer review process and aims to avoid the reliability and novelty of research being undermined due to missing vital citations.
- We develop a novel high-quality dataset containing submission-citation pairs extracted from real reviews, which are actually labeled by experienced experts from top-tier conferences and are easy to extend with more reviews coming out annually.
- We evaluate several mainstream methods in other similar research tasks on our proposed CitationR and establish a new method. Our proposed method achieves the best and can serve as a solid baseline for future research. All data and code are publicly available[1].

2. Related Work

We introduce the related work from three aspects, including citation recommendation (CR), transformer-based scientific document encoders, and other review-mining tasks.

[1] https://github.com/ChainsawM/RMC

2.1. Citation Recommendation

Based on the degree the input text covers the source paper, citation recommendation (CR) can usually be divided into two types (He et al., 2010): global citation recommendation (Gupta and Varma, 2017; Jiang et al., 2018; Hu et al., 2020), which recommends citations for a draft paper, and local citation recommendation (Ebesu and Fang, 2017; Yin and Li, 2017), which recommends citations for a short text context. Although early works (Tang and Zhang, 2009; He et al., 2011; Huang et al., 2012) use probability distributions to represent and analyze the relevance of documents, the advancement and prevalence of neural networks have made embedding-based models (Huang et al., 2015; Cai et al., 2018; Han et al., 2018; Zhang and Ma, 2020) the mainstream approach. Recently, Bhagavatula et al. (2018) use shallow feed-forward networks to learn representations of content and metadata of papers and introduce a contrastive learning objective for training the model. Gu et al. (2021) adopt hierarchical transformer layers as paper encoders and use the same contrastive learning objective for training.

Besides text, citations are a vital factor in measuring the similarity of scientific papers. Node representations learned from the citation graph can be taken as representations of papers (Liu et al., 2014; Pornprasit et al., 2022). However, constructing a citation graph requires a huge number of papers with dense citations, which is beyond our current collected dataset. Thus, graph-based methods are not considered in this paper.

2.2. Transformer-based Scientific Document Encoders

Pre-trained Language Models (LMs) (Devlin et al., 2019) based on transformer architecture have shown their surprising ability on numerous natural language processing tasks. Adapting to scientific domain corpora (Beltagy et al., 2019) further improved the performance of LMs and dominates various scientific document processing tasks, such as explaining relationships between scientific papers (Luu et al., 2020), scientific fact-checking (Cai et al., 2022), citation recommendation (Gu et al., 2021; Medic and Snajder, 2022), and so on.

Recently, several works try to leverage citations between scientific documents to enhance their representation learning when pre-training LMs. SPECTER (Cohan et al., 2020) leverages citations as a signal for document-relatedness and formulates this into a triplet-loss contrastive learning objective. CiteBERT (Wright and Augenstein, 2021) trains SciBERT with the task of cite-worthiness detection and LinkBERT (Yasunaga et al., 2022) fine-tunes BERT on the extra task of document relation prediction. SciNCL (Ostendorff et al., 2022) uses citation graph embeddings for a more informative selection of negative examples with the same contrastive learning objective as SPECTER. Transformer-based scientific document encoders with citation information introduced generally achieve better performance on document-level tasks. We adapt and evaluate their performance on RMC and apply them as the text encoder in our proposed framework to help solve the task.

Besides, Large Language Models (LLMs), which undergo extensive pretraining on diverse textual sources, have showed remarkable capabilities in text generation, language understanding, and context preservation (Zhou et al., 2023; Zhu et al., 2023). LLMs have been applied across various research fields, such as natural language processing (NLP) (Brown et al., 2020; Touvron et al., 2023a), code generation (Chen et al., 2021; Zheng et al., 2023), and recommender systems (Fan et al., 2023; Liu et al., 2023; Wu et al., 2023; Hou et al., 2023). In this paper, we include LLMs as baselines and design several prompts to evaluate their performance on RMC.

2.3. Other Review-Mining Tasks

In addition to identifying missed citations in reviews, there are other review-related tasks that can be roughly categorized into three main areas: (1) Using the content of reviews to predict the citation count of submitted papers (Li et al., 2019, 2022), predict final decisions (Wang and Wan, 2018; Deng et al., 2020; Kumar et al., 2022), and predict aspect scores (Deng et al., 2020; Li et al., 2020). (2) Analyzing the content of reviews for argument mining (Hua et al., 2019), sentiment analysis (Wang and Wan, 2018; Chakraborty et al., 2020), and grading reviews (Arous et al., 2021; Bharti et al., 2022). (3) Investigating the writing patterns of reviewers and exploring ways to automate the peer review process (Wang et al., 2020; Yuan et al., 2021; Lin et al., 2023). Different from RMC, these tasks primarily involve utilizing the content of reviews for various purposes.

3. Task Formulation

Let $S = \{s_1, \dots, s_m\}$ be the set of m submissions. From their reviews, a set of missed citations recommended by reviewers $R = \{r_{1,1}, \dots, r_{1,p}, \dots, r_{m,1}, \dots, r_{m,q}\}$ can be extracted, where $r_{m,q}$ is the q-th recommended paper for submission paper s_m. Besides, to better mimic the real situation, a set of n candidate papers $C = \{c_1, \dots, c_n\}$ is collected based on the characteristics of R (§4.4).

For a recommendation model, given the set of

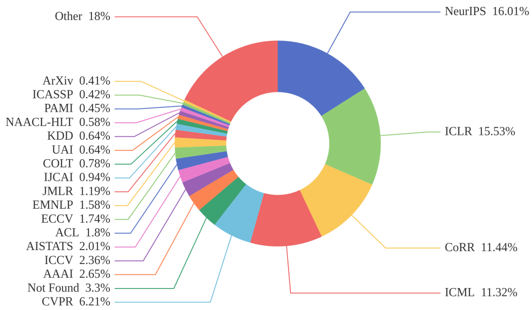

Figure 2: Distribution of venues of extracted citations recommended by reviewers.

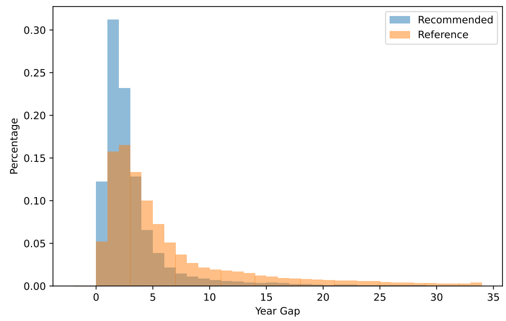

Figure 3: Distribution of year gaps between submissions and their citations.

all papers $P = S \cup R \cup C$, its input is a submission paper s, and it is supposed to calculate a rank score for each candidate paper in $P - \{s\}$ and output a paper list according to the descending rank scores.

In the task of RMC, we define $p = \{T, A, T_{r_1}, \dots, T_{r_n}\}$ represent a paper from the total paper collection P, where $w \in W$ is a word from the vocabulary set W, $T = [w_1^t, \dots, w_i^t]$ is the title consists of i words, $A = [w_1^a, \dots, w_j^a]$ is the abstract consists of j words and $T_{r_n} = [w_1^{r_n}, \dots, w_{m_n}^{r_n}]$ is the title of the n-th paper from the reference of p.

4. Dataset Construction

In this section, we present the building process and details of two versions of CitationR dataset.

4.1. Review Collection

The first step is review collection. We download reviews of scientific submissions from two conference sources: the NeurIPS[2] and the ICLR[3]. From the former source, we collect all accepted papers and corresponding reviews for NeurIPS 2013-2021, a total of 34,613 reviews for 9,526 papers. From OpenReview, we collect all submissions to ICLR 2017-2022, a total of 41,530 official anonymous reviews for 12,072 papers.

4.2. Citation Extraction

Peer review has been adopted by most journals and conferences to identify important and relevant research. Although various guidelines[4] or tutorials[5] about how to write good reviews have been proposed, there are no unified standards on the format of reviews, let alone how to cite external resources.

Empirically, we classify all mentions of papers into two categories. The first type pertains to papers that have already been cited by the authors but are mentioned by reviewers in their critiques. These mentions typically appear as brief phrases, such as a concatenation of the author name and publication year of the paper, or short phrases enclosed in brackets, such as "[ref X]", and so on. The second type of mentions includes papers that the authors do not cite but that are identified and recommended by reviewers to be included. Typically, these papers are mentioned through a formal reference section attached to the end or via URL links to external resources. Considering simplicity and practice, we make the intuitive assumption that missed citations in reviews are mentioned in the format of reference strings or URL links. Then we extract these mentions using regular expressions and manually corrected samples that are found invalid in the latter steps.

4.3. Paper Alignment

For papers mentioned in the format of URL links, we directly download according files and remove those that are not scientific papers. For reference strings, we try to align them to scientific papers via searching in bibliographic databases like DBLP[6] and Semantic Scholar[7]. In general, we adopt three rules to judge the alignment of reference strings and scientific papers: (1) The title of the retrieved paper appears in the reference string, (2) titles of retrieved papers belonging to the same reference string from two sources are matched, (3) and unmatched samples are manually checked. For down-

[2] https://proceedings.neurips.cc/
[3] https://openreview.net/group?id=ICLR.cc/
[4] https://icml.cc/Conferences/2023/ReviewerTutorial
[5] https://neurips.cc/Conferences/2022/ReviewerGuidelines
[6] https://dblp.uni-trier.de/
[7] https://www.semanticscholar.org/

Section	# Samples			# Submissions	# Reviews	Publication years
	train	val	test	(valid / collected)	(valid / collected)	
ICLR	10,646	1,556	1,492	5,127 / 12,072	5,872 / 41,530	2017-2022
NeurIPS	4,257	566	582	2,401 / 9,526	3,228 / 34,613	2013-2021
total	14,903	2,122	2,074	7,528 / 21,598	9,100 / 76,143	2013-2022
extended				0 / +40,810		2009-2022

Table 1: Statistics of CitationR dataset. Here "valid" means missed citations are found from those reviews and for those submissions.

loaded papers, the tool Doc2json[8] is used to extract their metadata and bibliographies.

In all, we collect 14,520 unique papers recommended by reviewers. Out of 21,598 collected submissions, 7,528 papers (around 35%) are identified as missing citations. Out of 76,143 collected reviews, 9,100 (around 12%) reviews contain citations recommended by reviewers. The average number of recommended citations per submission paper is around 2.5. Apparently, it is not a rare case that reviewers regard submissions as lacking important citations and recommend papers to cite.

4.4. Candidate Extension

To better mimic the real situation where researchers search for papers to cite, we also collect a set of candidate papers that featured the same as papers recommended by reviewers in some aspects. As illustrated in Figure 2, reviewers in NeurIPS and ICLR mostly recommend papers from top-tier conferences, and NeurIPS and ICLR are just the two venues reviewers most frequently recommend papers from. Moreover, as illustrated in Figure 3, compared to papers already cited by authors (in the reference section), papers recommended by reviewers are more up-to-date. Among papers we extracted, those published no more than three years earlier than submissions count more than 60%, and that of papers in the reference sections is no more than 40%.

Thus, considering availability and the above characteristics, we collect 40,810 papers published no more than three years earlier than recommended papers from publicly available top-tier conferences, including AAAI, ACL, AISTATS, COLT, CVPR, EMNLP, ICCV, ICML that range from 2009 to 2022. By adding collected candidate papers, we get the extended version of CitationR.

Finally, we split the dataset into training, validation, and test sets roughly in the ratio 8:1:1 based on publication years. Detailed statistics of the split dataset are listed in Table 1.

[8] https://github.com/allenai/s2orc-doc2json

5. Methodology

In this section, we introduce the RMCNet in detail, whose overall framework is shown in Figure 4.

5.1. Paper Encoder

Paper encoder aims to learn the embeddings of papers from their texts. Transformer-based encoder is adopted as the text encoder in our proposed model and can be easily replaced by other models. For a paper p, the concatenation of its title and abstract with an additional separator token inserted between them is fed into BERT, and a series of hidden states can be obtained:

$$\mathbf{h}_{[\mathrm{C}]}, \mathbf{h}_1^t, \ldots, \mathbf{h}_j^a = \mathrm{BERT}([\mathrm{C}], w_1^t, \ldots, w_i^t, [\mathrm{S}], w_1^a, \ldots, w_j^a) \tag{1}$$

where $[\mathrm{C}]$ denotes the special $[\mathrm{CLS}]$ token in BERT that is added to the front of a sequence, and $[\mathrm{S}]$ is the sentence separator token $[\mathrm{SEP}]$. Following (Devlin et al., 2019), the hidden state of $[\mathrm{C}]$ is used as the representation of the content of the input paper:

$$\mathbf{v}_{\mathrm{content}} = \mathbf{h}_{[\mathrm{C}]} \tag{2}$$

For an already cited paper in the reference section of p, its title $T_{r_x} \in P$ is input into BERT:

$$\mathbf{h}_{[\mathrm{C}]}^{r_x}, \mathbf{h}_1^{r_x}, \ldots, \mathbf{h}_{m_x}^{r_x} = \mathrm{BERT}([\mathrm{C}], w_1^{r_x}, \ldots, w_{m_x}^{r_x}, [\mathrm{S}]) \tag{3}$$

Similarly, we obtain the representation of an already cited paper in the reference section:

$$\mathbf{v}_{r_x} = \mathbf{h}_{[C]}^{r_x} \tag{4}$$

Intuitively, a paper's citing pattern can be exploited from the relations between itself and already cited papers. To model existing citations, we feed all embeddings of already cited papers into an *Attention* layer (Vaswani et al., 2017) and calculate the embedding of the reference as follows:

$$\mathbf{v}_R = \sum_{x=1}^{n} \mathbf{w}_x \mathbf{v}_{r_x} \tag{5}$$

$$\mathbf{w} = \mathrm{softmax}([\mathbf{v}_{r_1}, \ldots, \mathbf{v}_{r_n}]^\top \cdot \mathbf{v}_{\mathrm{content}}) \tag{6}$$

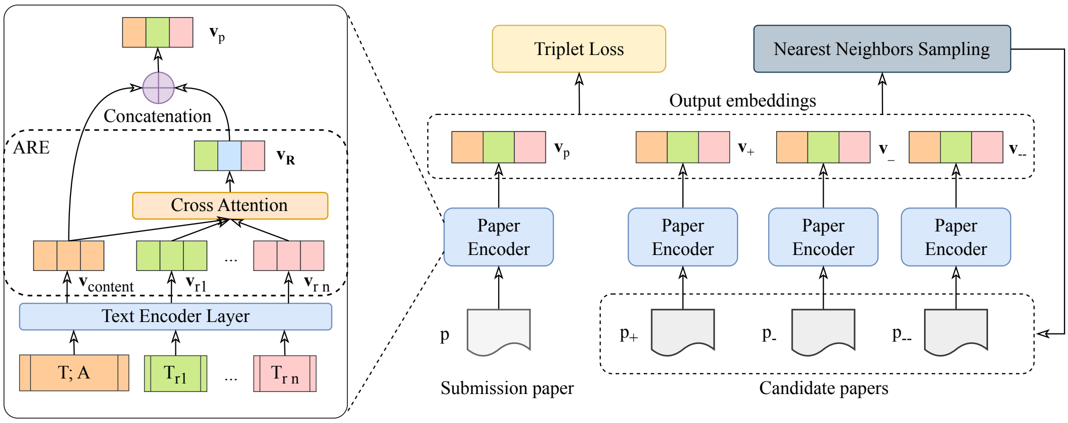

Figure 4: The overall architecture of RMCNet, which consists of three parts: (1) *Paper encoder* (left) generates representations of papers with an *Attentive Reference Encoder (ARE)* part mining the reference sections. (2) *Triplet Loss* (upper middle) computes the loss fusing positive samples and negative samples of different levels. (3) *Nearest Neighbors Sampling* (upper right) obtains negative samples of different levels based on output embeddings and their textual similarities to the submission paper.

where $\mathbf{w}$ is the weight vector and the scalar $\mathbf{w}_x$ is its x-th element.

Finally, in order to recommend missed citations relating to both the content and citation pattern, and avoid meaningless duplicates, we linearly combine the embeddings of content and reference to get the final representation of the input paper:

$$\mathbf{v}_p = (1-\alpha)\mathbf{v}_{\text{content}} + \alpha \mathbf{v}_R \quad (7)$$

where α is the parameter to balance the content and existing citations.

5.2. Triplet Loss

In particular, each training instance at least contains a triplet of papers: a submission (query) paper p, a positive paper p_+, and a negative paper p_-. The positive paper is the recommended citation extracted from reviews, and the negative paper is a paper that is not cited by p or recommended by reviewers. Via previously introduced paper encoder, respective $\mathbf{v}_p$, $\mathbf{v}_+$, $\mathbf{v}_-$ can be obtained, which represent the embeddings of papers from a training sample. We then train the model using the following triplet margin loss function:

$$\mathcal{L} = \max\{s(\mathbf{v}_p, \mathbf{v}_-) - s(\mathbf{v}_p, \mathbf{v}_+) + m, 0\} \quad (8)$$

where s is the similarity function and m is the loss margin hyper-parameter sets the span over which the loss is sensitive to the similarity of negative pairs. Following previous work (Bhagavatula et al., 2018; Gu et al., 2021), s is defined as the cosine similarity between two document embeddings.

5.3. Nearest Neighbors Sampling

The definition of positive example papers p_+ is straightforward, which are papers recommended by reviewers. However, a careful choice of negative example papers may be critical for model performance. We use two types of negative examples: (1) **Random**: Randomly selecting a paper as a negative example typically results in easy negative examples. (2) **Negative nearest neighbors**: Given a submission paper, we use the model's current checkpoint to obtain the top K_n nearest candidates excluding golden papers. The embeddings of these candidates have high similarities to the embedding of the input submission paper. Thus, we denote papers selected from those candidates as "hard negatives". It is expected that training the model to distinguish hard negative examples may improve overall performance. The checkpoint of the model is updated every N_{iter} training iterations, at which point the nearest candidates are also updated.

6. Experiments

In this section, we evaluate our method and four groups of baselines on CitationR. All reported scores are obtained by running models over at least three random seeds and are presented as percentage numbers with "%" omitted.

6.1. Approaches for Comparison

We have four groups of approaches for comparison.

The first group includes:

(1-1) **BM25** (Robertson and Walker, 1999), which is a highly effective strong baseline model representing traditional sparse retrieval models.

The second group comprises two citation recommendation models:

(2-1) **Citeomatic** (Bhagavatula et al., 2018), a global CR model uses shallow feed-forward networks to learn representations of papers.

	Dataset→	CitationR				Extended			
	Model↓ / Metric→	MAP	MRR	NDCG	R@10	MAP	MRR	NDCG	R@10
1	BM25	10.79	17.47	26.36	19.41	7.11	12.22	22.00	13.76
2	ChatGPT-turbo	11.44	19.02	26.97	19.41	7.92	14.11	22.76	13.76
3	LLaMA2-13b (2023b)	10.88	18.23	26.53	19.41	6.81	11.62	21.72	13.76
4	Citeomatic (2018)	10.11	15.91	26.34	18.59	8.82	15.07	22.88	14.96
5	H-Transformer (2021)	8.42	15.11	22.67	21.37	4.40	10.85	10.12	9.59
6	BERT (2019)	11.81	19.60	29.86	20.20	8.48	14.38	25.60	15.05
7	SciBERT (2019)	13.27	21.17	31.69	<u>23.55</u>	9.72	16.44	27.47	<u>17.94</u>
8	SPECTER (2020)	<u>13.44</u>	<u>21.87</u>	<u>31.92</u>	23.35	<u>9.83</u>	<u>16.60</u>	<u>27.48</u>	17.47
9	CiteBERT (2021)	12.70	20.61	31.23	23.15	9.19	15.44	26.91	17.03
10	LinkBERT (2022)	12.75	21.15	31.39	22.86	8.86	15.32	26.52	17.03
11	SciNCL (2022)	13.19	21.01	31.64	22.38	9.83	16.60	27.48	17.47
12	RMCNet (ours)	**14.94**	**23.13**	**33.28**	**25.59**	**10.48**	**17.14**	**28.28**	**18.76**
	$\pm\sigma$ w/ five seeds	*.185*	*.330*	*.168*	*.541*	*.241*	*.422*	*.219*	*.369*

Table 2: Performance results on CitationR and Extended CitationR. Our scores are reported as mean and standard deviation σ over five random seeds.

(2-2) **H-Transformer** (Gu et al., 2021), a local CR model adopts hierarchical transformer layers as paper encoders in the prefetching stage and generates recommendations by pair-wise reranking via SciBERT.

The third group consists of six pre-trained scientific document encoding models:

(3-1) **BERT** (Devlin et al., 2019), the dominant pre-trained model which achieves great success on various language understanding tasks.

(3-2) **SciBERT** (Beltagy et al., 2019), a variant of BERT trained on a corpus of scientific articles with masked language modeling objectives.

(3-3) **SPECTER** (Cohan et al., 2020), a SciBERT-based scientific document encoder trained with a contrastive learning objective that minimizes the L2 distance between embeddings of citing-cited paper pairs.

(3-4) **CiteBERT** (Wright and Augenstein, 2021), a variant of SciBERT fine-tuned on cite-worthiness detection task.

(3-5) **LinkBERT** (Yasunaga et al., 2022), a variant of BERT fine-tuned on document relation prediction task.

(3-6) **SciNCL** (Ostendorff et al., 2022), a SciBERT-based Encoder that uses citation graph embeddings for a more informative selection of negative examples.

The fourth group includes two large language models:

(4-1) **ChatGPT-turbo**, a representative, powerful and widely used conversational agent.

(4-2) **LLaMA2-13b** (Touvron et al., 2023b), a popular advanced large language model.

The input of baselines (1-1)-(3-6) is the concatenation of the title, abstract, and titles in the reference section of the input paper. For LLMs, we design prompts similar to those used in news recommendation (Liu et al., 2023) and let LLMs rerank the top ten[9] results of BM25. While baselines (2-1), (2-2), (4-1), and (4-2) directly output a sorted candidates list, other baselines generate vectors of papers upon which we calculate similarities and rank candidates.

6.2. Metrics and Implementation Details

Following previous work, we use four commonly used evaluation metrics: (1) Mean Average Precision (**MAP**); (2) Mean Reciprocal Rank (**MRR**) (Voorhees, 1999); (3) Normalized Discounted Cumulative Gain (**NDCG**) (Järvelin and Kekäläinen, 2002), a widely used measure of ranking quality and is computed by

$$\text{NDCG} = \sum_{i=1}^{|M|} \frac{2^{r(i)} - 1}{log_2(i+1)} \quad (9)$$

where M is the sorted list of papers output by models and $r(i) = 1$ if the i-th paper is the one reviewers recommended to cite, otherwise $r(i) = 0$. (4) **Recall@K**. Considering the scale of CitationR, we set $K = 10$.

We adopt the weights of SciNCL (Ostendorff et al., 2022) to initialize our text encoder and the loss margin m is set as 0.05 following (Gu et al., 2021). The optimizer is AdamW (Loshchilov and Hutter, 2017) with a learning rate of $\lambda = 2^{-5}$. Models are trained on a single NVIDIA GeForce V100 (32GB) GPU for five epochs, and its checkpoint is updated every $N_{iter} = 5,000$. We set $\alpha = 0.6$, $K_n = 100$ and the ratio of positive/hard negative/easy negative papers to 1:1:1.

[9] We also tested 20 and 30, but larger numbers resulted in slightly worse performance.

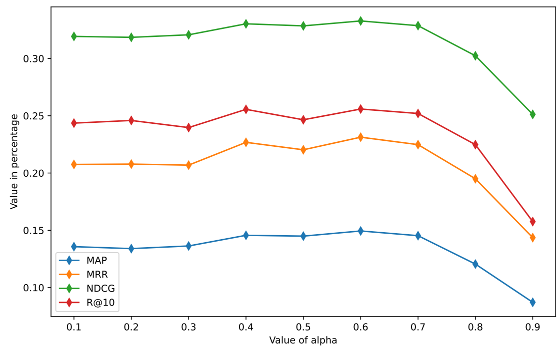

Figure 5: Results with different values of α.

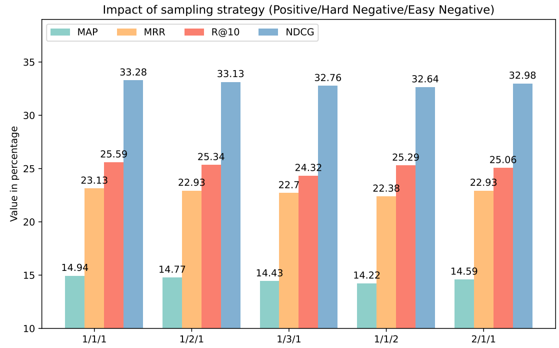

Figure 6: Results with different sampling strategies.

6.3. Main Results

The main results are summarized in Table 2. The overall best and previously best results are boldfaced and underlined, respectively. We have several observations:

Firstly, our proposed model achieves the best results in terms of all metrics on two versions of CitationR, displaying its superiority to other methods.

Secondly, pre-trained methods (Rows 6-12) perform generally better than other methods. The deeply stacked and large-scale pre-trained BERT model can better model text semantics than shallow word embeddings, which is crucial for content understanding in recommending citations. For H-Transformer, although it adopts hierarchical transformer layers as paper encoders, the lack of pre-training on large-scale corpora makes it fail in RMC. In contrast, Citeomatic obtains better results with shallow networks sufficiently trained on the dataset.

Thirdly, LLMs exhibit mediocre performance on RMC, with a notable disparity when compared to other BERT-based models. The results of reranking with LLaMA2-13b have even become worse on Extended CitationR. It seems that reranking is a complex task that goes beyond the capabilities of LLMs alone, which are good at generating texts and understanding contexts.

Fourthly, while H-Transformer achieves **75.7%** at R@10 on local CR (Gu et al., 2021) and Citeomatic obtains **77.1%** at MRR on global CR (Bhagavatula et al., 2018), neither model achieves more than 30% at any metrics on RMC. Apparently, existing models have much worse performance on RMC compared to traditional CR datasets, highlighting the complexity of RMC, and the need for more sophisticated recommendation methods.

6.4. Ablation Study

In this section, we study the effectiveness of different components of our proposed model by removing or replacing them.

We first show the effect of the *Attentive Reference Encoder* part by removing it. The results in

Model	MAP	MRR	NDCG	R@10
Ours	**14.94**	**23.13**	**33.28**	**25.59**
w/o ARE	13.10	20.17	31.39	23.03
w/o Hard Negatives (HN)	14.36	22.45	32.70	25.13
w/o ARE & w/o HN	13.61	20.86	31.90	24.45
average pooling	12.57	19.75	30.73	21.52
concatenation	13.68	21.25	32.11	24.35

Table 3: Ablation results on CitationR

Table 3 demonstrate that *ARE* plays a crucial role, indicating that it is necessary to mine the citing patterns by modeling already cited papers in RMC. We also remove the hard negative examples when training the model. The decreased performance verifies the benefits of training the model to distinguish hard negative examples. Nevertheless, the results of the model with both components removed are between those of the model with only one component removed. It implies that *ARE* has a greater impact than hard negative examples.

To further examine the effect of the *ARE*, we explore some other ways to tackle the reference section. “average pooling” means replacing the *Attention* layer with an average layer when calculating the representation of the reference $\mathbf{v}_R$. “concatenation” means replacing the linear combination with concatenation when calculating the final representation of the input paper $\mathbf{v}_p$. However, these replaced models result in a decrease in performance, proving the superiority of balancing the content and citation pattern in our method.

6.5. Parameter Analysis

6.5.1. Effect of α

Figure 5 displays the results on CitationR of different values of α, which is the linear weight of combining $\mathbf{v}_{\mathrm{content}}$ and $\mathbf{v}_{\mathrm{R}}$. The performance of the model achieves the best when $\alpha = 0.6$ and sharply decreases when α exceeds 0.7.

6.5.2. Effect of Sampling Strategy

Figure 6 presents the results on CitationR of different sampling strategies. Here a strategy means the number of positive/hard negative/easy negative papers in a training instance. Considering the statistics of CitationR, we set the number of papers in an training instance at a low level. For all metrics, the best results are achieved by strategy “1/1/1". Increasing the number of any kind of papers in a training instance raises a slight decrease in performance. This may be due to the current scale of CitationR.

7. Challenges and Future Work

In this section, we introduce the challenges and possible further research directions of RMC from three aspects:

7.1. Relevance

One key aspect of RMC is measuring the relevance between missed citations and submissions. We have designed an *ARE* module that leverages the capability of *Attention* to exploit potential relevance. However, along with most existing methods, our method solely considers the title and abstract of a scientific paper but ignores its body text. One possible research direction is incorporating additional information from papers in modeling while ensuring efficiency. Relevance may be more comprehensively explored with more information provided, especially the body text of submissions (Sugiyama and Kan, 2013, 2015). However, accurately and effectively modeling such long texts and mining the relevance that may only be mentioned in a short context is still challenging.

7.2. Intent

Both previous methods and our method focus on mining relevance, assuming that relevance is implicit in the text. However, cases may be that the citations recommended by reviewers are completely ignored by the submissions, and there is no textual relevance. In such cases, we believe that mining the intent behind the reviewers' recommendations can be helpful. As shown in Figure 1, reviewers often provide explanations for their recommended citations. Therefore, a feasible research direction is to automatically locate these explanations from review texts and utilize them to train smarter recommenders.

7.3. Understanding

Generally, reviewers rely on their accumulated academic knowledge and understanding of submissions to identify missed citations. Meanwhile, LLMs have shown remarkable capabilities in language understanding (Yang et al., 2023) and can effectively apply their learned knowledge and reasoning abilities to tackle new tasks (Zhu et al., 2023). Thus, one promising direction is to leverage the power of LLMs to obtain a fine-grained understanding of submissions and identify potential weaknesses that need further citations to support.

8. Conclusion

In this paper, we introduce a novel challenging task of <u>R</u>ecommending <u>M</u>issed <u>C</u>itations Identified by Reviewers (RMC). RMC aims to improve the citations of submissions and avoid the reliability and novelty of research being undermined because of missing vital citations. We curate a high-quality dataset named CitationR by extracting submission-citation pairs labeled by reviewers from real reviews in top-tier conferences. We conduct an extensive evaluation of four groups of strong methods on the developed dataset. Moreover, we propose a novel framework RMCNet by integrating an *Attentive Reference Encoder* module and contrastive learning objectives. Our proposed method outperforms all baselines in all metrics and can serve as a strong baseline. We highlight challenges and several potential research directions of RMC. We make all code and data publicly available to motivate future research.

Acknowledgements

We thank all the anonymous reviewers for comments and suggestions. This work was supported by the National Key Research and Development Project of China (2021ZD0110700) and Hunan Provincial Natural Science Foundation (Grant Nos. 2022JJ30668).

Ethics Statement

The datasets used in our research are collected through open-source approaches. The whole process is conducted legally, following ethical requirements.

We see opportunities for researchers to apply our built datasets and models to help identify missed citations of academic manuscripts. However, machine learning models are liable to amplify biases in training data (Hall et al., 2022), and current collected datasets are limited in scale and diversity. Researchers must consider these implications when conducting work on our datasets and consider whether a recommended citation is feasible when using models in practice.

References

Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2021. Peer grading the peer reviews: A dual-role approach for lightening the scholarly paper review process. *Proceedings of the Web Conference 2021*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *North American Chapter of the Association for Computational Linguistics*.

Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agarwal, and Asif Ekbal. 2022. A method for automatically estimating the informativeness of peer reviews. In *ICON*.

Andrew Booth, Diana Papaioannou, and Anthea Sutton. 2012. *Systematic Approaches to a Successful Literature Review*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiaoyan Cai, Junwei Han, and Libin Yang. 2018. Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation. In *AAAI*.

Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. 2022. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999 – 103999.

Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, S. Arun Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Annual Meeting of the Association for Computational Linguistics*.

Zhongfen Deng, Hao Peng, Congying Xia, Jianxin Li, Lifang He, and Philip S. Yu. 2020. Hierarchical bi-directional self-attention networks for paper review rating recommendation. *ArXiv*, abs/2011.00802.

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. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *ArXiv*, abs/2307.02046.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, pages 1 – 31.

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.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2021. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *ECIR*.

Shashank Gupta and Vasudeva Varma. 2017. Scientific article recommendation by using distributed representations of text and graph. *Proceedings of the 26th International Conference on World Wide Web Companion*.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron B. Adcock. 2022. A systematic study of bias amplification. *ArXiv*, abs/2201.11706.

Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. 2018. hyperdoc2vec: Distributed representations of hypertext documents. In *ACL*.

Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *WSDM '11*.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. 2010. Context-aware citation recommendation. In *WWW '10*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *ArXiv*, abs/2305.08845.

Yanli Hu, Chunhui He, Zhen Tan, Chong Zhang, and Bin Ge. 2020. Fusion of domain knowledge and text features for query expansion in citation recommendation. In *KSEM*.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. *ArXiv*, abs/1903.10104.

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

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A neural probabilistic model for context based citation recommendation. In *AAAI*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018. Cross-language citation recommendation via hierarchical representation learning on heterogeneous graph. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Sandeep Kumar, Hardik Arora, Tirthankar Ghosal, and Asif Ekbal. 2022. Deepaspeer: Towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews. *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–11.

Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. Multi-task peer-review score prediction. In *SDP*.

Siqing Li, Yaliang Li, Wayne Xin Zhao, Bolin Ding, and Ji rong Wen. 2022. Interpretable aspect-aware capsule network for peer review based citation count prediction. *ACM Transactions on Information Systems*.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji rong Wen. 2019. A neural citation count prediction model based on peer review text. In *Conference on Empirical Methods in Natural Language Processing*.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Inf. Fusion*, 98:101830.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023. Once: Boosting content-based recommendation with both open- and closed-source large language models.

Xiaozhong Liu, Yingying Yu, Chun Guo, and Yizhou Sun. 2014. Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 121–130. ACM.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2020. Explaining relationships between scientific documents. In *Annual Meeting of the Association for Computational Linguistics*.

Zoran Medic and Jan Snajder. 2022. Large-scale evaluation of transformer-based article encoders on the task of citation recommendation. In *SDP*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. *ArXiv*, abs/2202.06671.

Richard Pears and Graham Shields. 2013. *Cite Them Right: The Essential Referencing Guide*.

Chanathip Pornprasit, Xin Liu, Pattararat Kiattipadungkul, Natthawut Kertkeidkachorn, Kyoung-Sook Kim, Thanapon Noraset, Saeed-Ul Hassan, and Suppawong Tuarob. 2022. Enhancing citation recommendation using citation network embedding. *Scientometrics*, 127:233 – 264.

Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 2014. Cluscite: effective citation recommendation by information network-based clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Stephen E. Robertson and Steve Walker. 1999. Okapi/keenbow at trec-8. In *Text Retrieval Conference*.

Kazunari Sugiyama and Min-Yen Kan. 2013. Exploiting potential citation papers in scholarly paper recommendation. In *13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, Indianapolis, IN, USA, July 22 - 26, 2013*, pages 153–162. ACM.

Kazunari Sugiyama and Min-Yen Kan. 2015. A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *Int. J. Digit. Libr.*, 16(2):91–109.

Jie Tang and Jing Zhang. 2009. A discriminative approach to topic-based citation recommendation. In *PAKDD*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. *ArXiv*, abs/2010.06119.

Wei Wang, Tao Tang, Feng Xia, Zhiguo Gong, Zhikui Chen, and Huan Liu. 2022. Collaborative filtering with network representation learning for citation recommendation. *IEEE Transactions on Big Data*, 8:1233–1246.

Dustin Wright and Isabelle Augenstein. 2021. Citeworth: Cite-worthiness detection for improved scientific document understanding. *ArXiv*, abs/2105.10912.

Likang Wu, Zhilan Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2023. A survey on large language models for recommendation. *ArXiv*, abs/2305.19860.

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

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ArXiv*, abs/2304.13712.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *ArXiv*, abs/2203.15827.

Jun Yin and Xiaoming Li. 2017. Personalized citation recommendation via convolutional neural networks. In *APWeb/WAIM*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *J. Artif. Intell. Res.*, 75:171–212.

Yang Zhang and Qiang Ma. 2020. Dual attention model for citation recommendation. In *COLING*.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zi-Yuan Wang, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. 2023. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guan Wang, Kaichao Zhang, Cheng Ji, Qi Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, Lichao Sun Michigan State University, Beihang University, Lehigh University, Macquarie University, Nanyang Technological University, University of California at San Diego, Duke University, University of Chicago, and Salesforce AI Research. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *ArXiv*, abs/2302.09419.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv*, abs/2308.07107.