Survey on Automatic Related Work Generation: From Task Definition to Ethical Concerns

Lixiang Ci[1] Jessica Ouyang[1,2]
[1] University of Texas at Dallas
[2] Amazon Web Services
{lixiangci8, jessica.ouyang}@utdallas.edu

Abstract

To convince readers of the novelty of their research paper, authors must perform a literature review and compose a coherent story that connects and relates prior works to the current work. This challenging nature of literature review writing makes automatic related work generation (RWG) academically and computationally interesting, and also makes it an excellent test bed for examining the capability of SOTA natural language processing (NLP) models. Since the initial proposal of the RWG task, its popularity has waxed and waned, following the capabilities of mainstream NLP approaches. In this work, we survey the zoo of RWG historical works, summarizing the key approaches and task definitions and discussing the ongoing challenges of RWG.

1 Introduction

Academic research is an exploratory activity to solve problems that have never been resolved before. Each academic research paper must sit at the frontier of the field and present novelties that have not been addressed by prior work; to convince readers of the novelty of the current work, the authors must perform a literature review to compare their work with the prior work. In natural language processing (NLP), a short literature review is usually conducted under the “Related Work” section (RWS). Writing an RWS is non- trivial; it is insufficient to simply concatenate generic summaries of prior works. Instead, composing a coherent story that connects each related work and the current (citing) work, reflecting the author’s understanding of their field, is preferred (Li and Ouyang, 2024).

The challenging nature of RWS writing makes automatic related work generation (RWG) an academically and computationally interesting problem. RWG is a complex task that involves multiple NLP subtasks, such as retrieval- augmented generation, long document understanding, and query- focused multi- document summarization. Moreover, since most NLP papers have an RWS and NLP researchers are natural domain experts for evaluating these RWS, the RWG task is an excellent test bed for examining the capability of SOTA NLP models.

RWG also fills a practical need. Due to the rapid pace of research publications, including preprints that have not yet been peer- reviewed, keeping up to date with the latest work in a research area is very time- consuming. Even with daily feed tools, like the Semantic Scholar Research Feed, researchers still have to curate, read, and digest all the new papers in their feed. Thus, there is a need for concise, automatically generated literature reviews that regularly summarize the papers in a user’s feed.

Since Hoang and Kan (2010) initially proposed the task, the popularity of RWG has waxed and waned, following the capabilities of mainstream NLP approaches: from rule- based to extractive summarization, then to abstractive summarization on the sentence level, and finally to abstractive section- level RWG. Currently there is a surge of renewed interest in RWG due to the recent success of large language models (LLMs). In this work, we survey the zoo of RWG historical works.

We find that, surprisingly, most RWG works are not directly comparable because they vary drastically in task definition and simplifying assumptions (Section 2), as well as using different input features and representations (Section 3). There is no standard benchmark dataset for RWG (Section 4), as most works apply custom preprocessing to extract RWS or individual citations, reflecting differences in their task definitions. Further, many works do not release their models or generated outputs, so it is often impossible for later works to compare against earlier approaches (Section 5). Finally, we discuss ethical concerns related to RWG, such as plagiarism and non-factual statements, and the potential consequences of fully automatic RWG on the human process of scientific thinking and writing (Section 6.3).

2 Task Definition

The task definition for RWG has varied as the SOTA text summarization approach has evolved over time. Even where the overall approach is similar (e.g. extractive or abstractive approaches), different assumptions are made with respect to the availability of system inputs and the unit at which an RWS is generated (Table 1).

2.1 Extractive Related Work Generation

Hoang and Kan (2010) defined RWG as generating the RWS of a target paper given the rest of the target paper and all cited papers. This focus on extracting and concatenating salient sentences from the cited papers to form an RWS was used by most subsequent extractive RWG approaches (Hu and Wan, 2014; Wang et al., 2018; Deng et al., 2021). One key variant is that of Chen and Zhuge (2019); Wang et al. (2019), who extracted sentences from other works that also referenced the cited papers.

Otherwise, the main difference among extractive approaches is in how they order the extracted sentences: Hoang and Kan (2010); Wang et al. (2018); Chen and Zhuge (2019); Wang et al. (2019) assumed the correct ordering as input (either via a human- constructed topic tree or the ground truth ordering of the target RWS), while Hu and Wan (2014); Deng et al. (2021) used topic modeling and a sentence reordering module, respectively, to predict an ordering.

2.2 Abstractive Related Work Generation

With the advent of neural language models, two different versions of the abstractive RWG task have been proposed: generating single citation texts versus paragraphs or full RWS.

2.2.1 Citation Text Generation

Early neural language models, such as the Pointer- Generator (See et al., 2017) and early pre-trained Transformers (Vaswani et al., 2017)), were capable of fluent abstractive summarization but had severe input length restrictions. Because scientific research papers are very long documents, a new version of the RWG task arose: generating individual citation texts. The system input now needed to include only one or a few cited papers, and to further shorten the system input, researchers no longer included the full texts of the target and cited papers, but used only the target citation context and the cited paper abstract (and occasionally the introduction and conclusion sections).

The main difference among single citation text generation works is in how a citation is defined. AbuRa'ed et al. (2020); Xing et al. (2020); Ge et al. (2021); Luu et al. (2021); Gu and Hahnloser (2023) restrict citation texts to be single sentences; Jung et al. (2022) allow any number of consecutive sentences, while Li et al. (2022, 2023); Mandal et al. (2024) additionally allow citations that are shorter than a full sentence. Almost all works restrict citations to contain only one cited paper; only Li et al. (2022, 2023); Mandal et al. (2024) explicitly allow multiple cited papers.

2.2.2 Section-Level Generation

Chen et al. (2021, 2022) pioneered section- level RWG by treating the paragraph as the unit of generation; they required that a target paragraph contain at least two citations, explicitly distinguishing their work from the single citation text generation setting. While Chen et al. (2021, 2022) used the given paragraph organization of the target RWS, subsequent works focused on ordering and organizing citations into paragraphs and generating transitional sentences between citations (Liu et al., 2023; Li and Ouyang, 2024; Martin- Boyle et al., 2024).

Further, the great success of SOTA LLMs in multiple natural language understanding and generation tasks, combined with their large context windows, have recently made it possible to generate a full RWS in a single pass (Li and Ouyang, 2024; Martin- Boyle et al., 2024). Thus, the task definition has now returned to the full RWS generation originally proposed by Hoang and Kan (2010) and previously tackled only by extractive approaches.

Table 1: Comparison of the task definitions of extractive and both single- citation and full-section abstractive approaches to related work generation. * indicates works that allow multi-sentence citations. † indicates works that extract snippets/features from the cited paper full text. ** indicates works that use human editing to improve predicted citation groupings. ‡ indicates works that provide large language model prompts.

	Generation Unit	**# Cited Papers per Citation**	**Cited Paper Input**	**Additional Notes**
Extractive	Sequence of sentences	Varies	Full text or sections	Ordering assumed or predicted
Abstractive (Citation)	Single sentence or span	Mostly single	Abstract (and intro/concl)	Context sentences used
Abstractive (Section)	Paragraph or full RWS	Multiple	Abstract	May use citation networks

3 Overview of Approaches

When Hoang and Kan (2010) proposed the RWG task, they identified three main steps: (1) Finding relevant documents, (2) Identifying the salient aspects of these documents with respect to the current work; (3) Generating a topic- biased summary. In practice, all existing works skip the document retrieval step by using the gold cited paper list in the target RWS. At a high level, the methodologies of most extractive, citation- level and section- level abstractive RWG approaches are similar within their respective categories: extractive approaches focus on the salience step and simply concatenate the extracted sentences to form the summary, while abstractive approaches focus on directly generating the summary, often without explicitly modeling salience. In this section, we do not give an exhaustive description of all methodologies, but highlight some common features and design perspectives from the overall body of RWG work (summarized in Table 2). The details of individual works can be found in Appendix A.

3.1 Representing Cited Papers

Abstracts. In abstractive RWG approaches, and some extractive approaches, the cited paper title and abstract are commonly used as a proxy for its full text (AbuRa'ed et al., 2020; Xing et al., 2020; Ge et al., 2021; Chen et al., 2021, 2022; Jung et al., 2022; Li et al., 2022; Gu and Hahnloser, 2023; Liu et al., 2023; Mandal et al., 2024; Martin- Boyle et al., 2024), occasionally augmented with the introduction and/or conclusion (Hu and Wan, 2014; Chen and Zhuge, 2019; Deng et al., 2021).

The abstract is a concise summary of the central ideas of the cited paper and can fit in a neural language model's input length limit where the full text cannot. Abstracts also play an important role in scientific communication as a preview of the paper, so they are easy to access even when their full text are blocked by paywalls. Li and Ouyang (2024) find that generated RWS conditioned on cited paper abstracts are preferred by human readers over those conditioned on LLM- generated faceted summaries (Meng et al., 2021) of the cited papers.

Cited Text Spans (CTS). Li et al. (2023) proposed to condition on automatically predicted CTS rather than cited paper abstracts. CTS refers to the specific span of the cited paper that a given citation refers to; to draw a parallel to claim verification, the citation can be thought of as the claim, and the CTS as its supporting evidence. Thus, Li et al. (2023) effectively proposed an extract- then- abstract approach to citation text generation, arguing that the cited paper abstract may not always contain sufficient information to ground the target citation. It is interesting to note that CTS had previously been used for extractive RWG by Wang et al. (2019), who extracted CTS for other citations of the cited paper in works similar to the target paper.

Citation Graphs. Since an RWS describes the relationship between the target paper and prior work, as well as among prior works, some section- level RWG approaches have modeled the local citation network of the target and cited papers. Wang et al. (2018) used a random walk on a heterogeneous bibliography graph consisting of paper, author, venue, and keyword nodes to prune the search space of salient sentences for extractive RWG. Ge et al. (2021); Chen et al. (2021, 2022) used customized neural network architectures inspired by Graph Attention Networks (Velickovic et al., 2018) to encode the local citation network as an additional input for abstractive RWG, while Li and Ouyang (2024) prompted an LLM to generate a natural language description of the relationship between a pair of papers in the citation network.

3.2 The Importance of Citation Context

Citation context refers to the text preceding or surrounding the target citation or RWS. In the case of individual citations, the context is commonly defined as several sentences before, and optionally after, the target citation (Xing et al., 2020; Ge et al., 2021; Li et al., 2022, 2023; Mandal et al., 2024); for some citation text generation works and most section- level RWG works, the context can be the full text of the target paper, or a few key sections, most commonly the title, abstract, introduction, and conclusion (Luu et al., 2021; Jung et al., 2022; Gu and Hahnloser, 2023; Chen et al., 2022; Li and Ouyang, 2024; Martin- Boyle et al., 2024).

Intuitively, the context indicates which topics are salient to the target paper, restricting the RWG solution space. Extractive works (Hu and Wan, 2014; Chen and Zhuge, 2019; Wang et al., 2019) used the context as a query to score cited paper sentences. In abstractive approaches, conditioning on the context improves the coherence of the generated text with the rest of the target paper; Mandal et al. (2024) found human readers preferred citations generated using the entire context, with the target citation embedded inside it, as the generation target.

It is interesting to note that a few works did not use any target paper context at all (Hoang and Kan, 2010; AbuRa'ed et al., 2020; Chen et al., 2021), but these were early works in their respective categories (extractive versus abstractive citation- or section- level generation), and later works all used target paper context.

Table 2: Comparison of RWG approaches. * All surveyed works used cited paper titles and abstracts, which are not list in this table. † The target citation itself is masked. $\sqrt{\ldots}$ indicates features extracted from the listed sections.

	Cited Paper Representation*					Target Paper Context			Citation Analys	
	Intro	RWS	Concl.	CTS	Graph	Abs.	Intro	RWS†	Concl.	MTL
Extractive										
Hoang and Kan (2010)										
Hu and Wan (2014)	√					√	√			
Wang et al. (2018)					√					
Chen and Zhuge (2019)	√		√			√	√		√	
Wang et al. (2019)				√						
Deng et al. (2021)			√							
Abstractive (citation)										
AbuRa'ed et al. (2020)										
Xing et al. (2020)										
Ge et al. (2021)					√					√
Luu et al. (2021)										
Jung et al. (2022)										
Li et al. (2022)										
Gu and Hahnloser (2023)										
Li et al. (2023)				√						
Mandal et al. (2024)										
Abstractive (section)										
Chen et al. (2021)					√					√
Chen et al. (2022)					√	√				√
Liu et al. (2023)										
Li and Ouyang (2024)						√	√	√	√	
Martin-Boyle et al. (2024)										

3.3 Applying Citation Analysis

Citation analysis is a related area of research studying the properties of citations in scientific writing. Several studies have proposed taxonomies such as citation function (Garfield et al., 1965; Teufel et al., 2006; Dong and Schafer, 2011; Jurgens et al., 2018; Tuarob et al., 2019; Zhao et al., 2019), citation intent (Cohan et al., 2019; Lauscher et al., 2021), and citation sentiment (Athar, 2011; Athar and Teufel, 2012; Ravi et al., 2018), and such labels have been used to improve RWG performance.

Ge et al. (2021) used citation function prediction as an auxiliary training objective. Jung et al. (2022); Gu and Hahnloser (2023) used citation intents to perform controllable citation text generation. Inspired by the observation of Lauscher et al. (2022) that simple citation label sets struggle to represent ambiguous, real- world citations, Li and Ouyang (2024) used LLM-generated, natural language descriptions of function of a cited paper in other, similar works that also cited it.

Other work has studied the discourse properties and organization of citations. Jaidka et al. (2010, 2011); Khoo et al. (2011); Jaidka et al. (2013b,a) classified literature reviews into integrative (summarizing individual cited papers) and descriptive (focusing on high- level ideas from multiple papers) writing styles. Li et al. (2022) proposed a more fine- grained taxonomy at the citation level, labeling citations as dominant (the main focus of their sentence) or reference (tangential to the rest of their sentence).

Li and Ouyang (2024) used this taxonomy to analyze the writing style of LLM- generated RWS and observed a strong correlation between the proportion of reference- type citations and human preference scores, concluding that human readers prefer integrative RWS supported by reference- type citations. Similarly, Martin- Boyle et al. (2024) found that both human- written and human- assisted, LLM- generated RWS had significantly more cited papers per sentence than pure machine- generated RWS.

3.4 Human-Assisted Generation

While RWG models are optimized to reconstruct the original citation texts or RWS in their training datasets, the ultimate goal of the task is to generate an RWS that satisfies a user. Human readers are sensitive to errors in cited paper organization (e.g. papers cited in the same paragraph are not sufficiently related to each other) and emphasis (e.g. less salient papers are described in greater detail than more salient ones); currently, even SOTA LLMs are not capable of organizing and emphasizing a set of cited papers without human guidance (Li and Ouyang, 2024; Martin- Boyle et al., 2024).

Thus, human input has been included in several RWG works. To determine the most salient aspects of a cited paper for single citation text generation, Li et al. (2023) proposed to retrieve cited text spans (CTS) using user- provided keywords as queries, while Gu and Hahnloser (2023) directly used human- written keywords as an additional input. Li and Ouyang (2024) extended this idea to section- level RWG by proposing to use a human- written short summary of the main ideas of the target RWS. Also for section- level RWG, Martin- Boyle et al. (2024) introduced a human- in- the- loop component where the user edited an predicted cited paper grouping before the generation step.

4 Datasets

Despite the twenty published works on RWG, there is no standard benchmark dataset for the task. As we discussed in Section 2, most RWG works define their own version of the task; they also create their own datasets, adapted to their particular task definition. In this section, we describe the most commonly used sources of scientific articles (Table 3) and summarize how

RWG works have built on these sources. The details of each work's datasets can be found in Appendix Table 6.

4.1 Common Datasets

The ACL Anthology Network (AAN) Corpus (Radev et al., 2013) consists of papers published by the Association for Computational Linguistics (ACL). For each paper, it annotates the set of sentences in any other AAN paper that cite that paper. Both in the construction of AAN, as well as in single citation text generation works that use it, individual citation texts are extracted via string search for citation marks, such as Smith et al. (2024) or "[1]" (Xing et al., 2020; Ge et al., 2021).

SciSummNet (Yasunaga et al., 2019), used by AbuRa'ed et al. (2020); Deng et al. (2021), is a subset 1000 papers from the AAN Corpus with human- validated citation sentences and summaries.

Delve (Akujuobi and Zhang, 2017) consists of papers from several computer science conferences spanning multiple fields of research. It includes automatically extracted paper abstracts and full text, as well as citation texts and links.

The Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020) contains open- access papers from multiple disciplines. The papers are annotated with automatically detected inline mentions of citations, figures, and tables, which saves researchers the need to process raw PDF files.

Citation Oriented Related Work Annotation (CORWA) (Li et al., 2022) is derived from the ACL partition of S2ORC and is annotated specifically for citation text generation. CORWA labels citations and their discourse roles (dominant or reference).

Table 3: List of common datasets used in related work generation. * indicates works that use the SciSummNet subset of AAN. ‡ indicates works that use the CORWA subset of S2ORC. † indicates works that published their datasets, but the repositories are no longer accessible.

Dataset	Description	Used By
AAN Corpus	ACL papers with citation sentence annotations	Many early works
SciSummNet*	Subset of AAN with human-validated citations	AbuRa'ed et al. (2020); Deng et al. (2021)
Delve	CS papers with abstracts, full text, citation texts	Several works
S2ORC	Multi-disciplinary open-access corpus with citation mentions	Many recent works
CORWA‡	ACL papers annotated for citation text generation	Li et al. (2022); subsequent works

4.2 Discussion

One common challenge with all existing datasets is that, for a given target paper, not all of its cited papers are necessarily in the dataset (e.g. because they are behind a paywall). In single citation text generation works, such missing cited papers are simply omitted from training and testing. For section- level RWG, missing cited papers are a bigger problem, as their absence may disrupt the flow of the generated RWS (Li and Ouyang, 2024).

It is also interesting to note that the majority of RWG works have used NLP datasets, and almost no works use papers from outside the domain of computer science. It is likely that RWG researchers prefer to use NLP papers because they include a separate RWS that is easy to extract, which is not the case in all fields of research; they are within the researchers' own domain of expertise, making system development easier; and they are in the domain of the researchers' colleagues, making it easier to recruit human judges for evaluation.

Finally, with the advent of LLM- based approaches, RWG researchers must contend with the possibility that a target paper was part of the training data of their model. As a result, LLM-based works have explicitly targeted recent papers (Li and Ouyang, 2024; Martin- Boyle et al., 2024).

5 Evaluation

5.1 Baselines

As Appendix Tables 8 & 9 show, there are a few baselines widely used across RWG works. Extractive works commonly use LEAD (Wasson, 1998), MEAD (Radev et al., 2004), LexRank (Erkan and Radev, 2004), and TexRank (Mihalcea and Tarau, 2004), while abstractive works use naive sequence- to- sequence approaches, with base models such as PTGEN (See et al., 2017), BertSumAbs (Liu and Lapata, 2019), and Longformer Encoder- Decoder (Beltagy et al., 2020). These common baselines are relatively easy to replicate because they are well- documented, general- purpose summarization approaches. In contrast, most specialized RWG approaches are not easy to replicate and are thus rarely used as baselines for later works; we discuss this issue further in Section 6.1.

5.2 Metrics

Almost all RWG works use the summarization metric ROUGE (Lin, 2004) as their automatic evaluation metric; Luu et al. (2021) additionally use the translation metric BLEU (Papineni et al., 2002).

Most works additionally conduct human evaluations, as is common in natural language generation tasks. While there is no fixed standard for how to conduct an RWG human evaluation, most works evaluate at least 15 samples, with three human judges per sample. Judges are generally asked to rate the fluency or readability, the coherence with respect to the target paper, and the relevance or informativeness with respect to the cited paper on a five- point Likert scale.

The relatively small number of human- evaluated samples in RWG works is likely due to the difficulty of recruiting human judges with the expertise to understand the generated citation texts or RWS, as well as the high time commitment and difficulty of the task, which requires judges to read multiple, highly specialized documents. A more detailed summary of metrics used in RWG works can be found in Appendix Table 10.

6 Conclusions and Discussion

Having surveyed the field of RWG from the perspectives of task definition, approach, datasets, and evaluation methods, we conclude by identifying three main challenges in modern RWG and make recommendations for future work in this area.

6.1 Lack of Comparability

Work in RWG is fragmented in terms of task definitions, datasets used for training and evaluation, and how evaluations are conducted. Unlike most NLP tasks, there are no standard benchmarks for RWG. Table 1 shows that around half of existing works do not release their models or generated citation texts/RWS, making it impossible to reproduce or directly compare approaches.

As we discuss in Section 2, RWG works do not agree on the definition of citation (one or more cited papers discussed in one or more sentences, or just part of a sentence) or related work section (a concatenation of individual citations or paragraphs versus one continuous and coherent piece of text). Thus, the target outputs of most RWG systems are not directly comparable to those of other systems.

A deeper problem with the varying definitions of citation is noted by Li et al. (2022), who argue that human annotators can easily find examples of human- written citations that are longer or shorter than a single sentence, or that contain more than one cited paper, so ignoring citations that are longer than a single sentence or discuss more than a single cited paper is unrealistic. They further argue that restricting citations to be single sentences is problematic when the approach uses citation context; in the case of a multi- sentence citation, an RWG system that assumes each citation can only be one sentence and uses the surrounding sentences as context will actually use the rest of the sentences from the target citation as context, creating an information leakage problem.

Variation in datasets comes partly from differences in the task definition and partly from the fact that, of the commonly used source corpora, only the CORWA partition of S2ORC (Li et al., 2022) is explicitly designed for RWG; the others (AAN, S2ORC, and Delve) are general-purpose scholarly document and citation analysis datasets. As a result, these other source corpora either automatically extract citations by searching for sentences containing citation marks or do not label citations at all; in the latter case, RWG researchers extract citations themselves by searching for sentences containing citation marks and imposing assumptions about the number of cited papers a citation can contain. Besides CORWA, only the annotations of Xing et al. (2020) provide human- labeled citations.

Finally, variation in evaluation stems from the existing problem in general summarization research where automated metrics, such as the commonly used ROUGE scores, do not correlate well with human judgments, so many RWG works perform human evaluation. While fluency, coherence, and relevance are commonly used aspects of human evaluation (Appendix Table 10), many works define custom aspects, such as succinctness (Chen et al., 2021; Deng et al., 2021; Liu et al., 2023), factual correctness (Li and Ouyang, 2024), and correctness of citation intent (Jung et al., 2022; Gu and Hahnloser, 2023).

6.2 Common Limitations and Suggestions for Future Work

We find several limitations common to existing work on RWG for future work to consider.

Citation ordering and organization. Out of twenty surveyed RWG works, only four attempt to predict the correct ordering and/or grouping of citations into paragraphs (Hu and Wan, 2014; Deng et al., 2021; Liu et al., 2023; Martin- Boyle et al., 2024); an additional two papers acknowledge the citation ordering and grouping problem but assume a human- provided ordering is available (Hoang and Kan, 2010) or use a chronological ordering heuristic (Li and Ouyang, 2024). Yet Li and Ouyang (2024) observed that human readers noticed and disliked errors in citation grouping, such as when chronologically adjacent cited papers about different topics were placed in the same paragraph, and Martin- Boyle et al. (2024) found significant differences in the organization of generated RWS with and without human- assisted citation grouping.

We suggest fully automatic citation ordering and grouping as an important area for further investigation. For example, cited papers might be clustered based on their faceted summaries (e.g. their task objectives or methodologies; Meng et al., 2021). In addition, the generated RWS should deliver a coherent story and use a more abstract, humanlike writing style, perhaps by using LLMs with multi- stage prompting to simulate human authors' thinking processes. Existing human- in- the- loop approaches can be extended to develop RWS that are truly helpful to users.

Transition sentences and writing style. Based on the terms from general summarization (Klavans et al., 2001), Hoang and Kan (2010) distinguished informative sentences, which "give detail on a specific aspect of the problem... definitions, purpose or application of the topic", and indicative sentences, which “make the topic transition explicit and rhetorically sound”. However, modern abstractive approaches have focused on informative sentences: single citation generation approaches completely ignore indicative transition sentences, and section- level approaches include them only in that they are part of the target paragraphs. Li and Ouyang (2024) found that human readers asked for more transition sentences, complaining about RWS that simply concatenated one cited paper summary after another. Further, in their analysis of RWS writing style and citation clusters, Martin- Boyle et al. (2024) have shown that generated RWS do not draw enough connections among cited papers.

Thus, the generation of transition sentences and multi- paper citations remains an open problem. Where existing works have often explicitly excluded multi- paper citations, future works should explicitly target them. Similarly, the distinction between the reference- style citations (Li et al., 2022), which are more like extreme summarization, and the dominant- style citations that current models tend to produce, should be accounted for; future works can use different models for these two very different citation styles.

Retrieval- augmented related work generation. Existing RWG works assume the list of cited papers is available as input, but this assumption is unrealistic, as evidenced by the existence of “missing citations” questions on many conference and journal peer review forms. Li and Ouyang (2024) reported that several human judges expressed the desire for a system that would not only help them draft a RWS, but also alert them to any other relevant papers they should consider citing.

Given the recent success and popularity of retrieval- augmented generation (RAG) approaches (Lewis et al., 2020; Shuster et al., 2021), applying RAG to RWG is a promising direction for future RWG research. Future works may start with a partial list of works that should definitely be cited, alongside a set of candidate works that might be related. They could then use RAG to iteratively select a candidate paper and generate its transition/citation sentences. This functionality is crucial because RWG systems are much less practically useful without the ability to search for additional related works.

6.3 Ethical Concerns

Finally, we discuss three ethical issues related to the RWG task. First, abstractive RWG works must be concerned with the problems of plagiarism and factual errors. In extractive approaches, the generated RWS is by its very nature plagiarized, since its sentences are copied directly from the cited papers; it was presumably well- understood by extractive RWG researchers that their systems could never be used to directly write the RWS for a new paper. However, extractive approaches cannot hallucinate, so their outputs are less likely to contain factual errors about the cited papers.

With modern abstractive RWG, the situation is muddier. It is well- known in general summarization research that abstractive models can still copy significant chunks of text directly from their inputs (Grusky et al., 2018; Narayan et al., 2018), and factual consistency in summarization is an active research area (Cao et al., 2018; Goodrich et al., 2019; Falke et al., 2019; Kryscinski et al., 2019). Thus, it is possible for an abstractive RWG system to output plagiarized or hallucinated text, which should be of concern to any user who wishes to use such a system to write an RWS.

Second, the use of RWG to write an RWS for a paper one intends to submit for publication raises questions of academic dishonesty. Is it ethical for a researcher to put an automatically generated RWS in a submitted manuscript? Does this mean the researcher is claiming to have written that RWS, as they presumably wrote the rest of the paper? Do the answers to these questions change if the researcher has edited the automatically generated RWS? As with many concerns relating to the use of powerful modern LLMs, these questions are very new, and there is as yet no consensus among the scientific community on how to answer them. While automatically generated RWS as currently easy to recognize, we nonetheless urge caution on the part of RWG researchers and users.

Third, RWG is a challenging task even for humans; in many doctoral programs, writing a formal literature review is part of their candidacy qualifying exams (Knopf, 2006). Thus, the process of writing an RWS may be considered an important process for researchers where they must read broadly and think deeply about how their contributions fit into the bigger picture of their field. Some RWG works have argued that writing an RWS is arduous and time- consuming, and so RWG should save researchers from having to do it, but we argue this position ignores the value of RWS writing as a learning and thinking experience. We urge RWG researchers to consider human- in- the- loop frameworks, following Gu and Hahnloser (2023); Li and Ouyang (2024); Martin- Boyle et al. (2024).

Limitations of this Survey

There is currently a surge of interest in RWG, so new papers are being published that may not be included in this survey.

Due to the length limit, we are not able to give a detailed discussion of each work's methodology and implementation. We include cheat sheets in Appendix A to summarize the surveyed works from various perspectives. We also do not compare the specific performance scores of the surveyed works because they are generally not directly comparable.

As with any survey paper, the opinions and interpretations are ours and may not reflect what the authors of the surveyed papers believe about their own work.

A Appendix A: Summary Tables

A.1 Table 4: Problem Formulations for Extractive RWG

Table 4: A summary of the problem formulations of the prior works on extractive related work generation. All of their generation targets are a sequence of extracted sentences.

Prior Work	Inputs
Hoang and Kan (2010)	Topic hierarchy tree of the target related work, full cited papers
Hu and Wan (2014)	Target paper (abstract, introduction), cited papers (abstract, introduction, related work, conclusion)
Wang et al. (2018)	Full-texts of cited papers
Chen and Zhuge (2019)	Title, abstract, introduction, and conclusion for both target paper and cited papers; papers that co-cite the cited papers
Wang et al. (2019)	Full papers of target paper and cited papers, citation sentences that co-citing the cited papers
Deng et al. (2021)	Abstract or conclusion sections of the cited papers

A.2 Table 5: Problem Formulations for Abstractive RWG

A.3 Table 6: Dataset Summary

A.4 Table 7: Approach Summary

A.5 Table 8: Evaluation for Extractive RWG

A.6 Table 9: Evaluation for Abstractive RWG

A.7 Table 10: Human Evaluation Perspectives

References

References

[1] Cong Duy Vu Hoang and Min- Yen Kan. 2010. Towards automated related work summarization. In *Coling 2010: Posters*, pages 427–435, Beijing, China.

[2] Xiangci Li and Jessica Ouyang. 2024. Explaining relationships among research papers. *arXiv preprint arXiv:2402.13426.*

[3] Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: an optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633.

[4] Yongzhen Wang, Xiaozhong Liu, and Zheng Gao. 2018. Neural related work summarization with a joint context- driven attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786, Brussels, Belgium.

[5] Zekun Deng, Zixin Zeng, Weiye Gu, Jiawen Ji, and Bolin Hua. 2021. Automatic related work section generation by sentence extraction and reordering.

[6] Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, 31(3):e4261.

[7] Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang. 2019. Toc- rwg: Explore the combination of topic model and citation information for automatic related work generation. *IEEE Access*, 8:13043–13055.

[8] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer- generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada.

[9] 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*, pages 5998–6008.

[10] Ahmed AbuRa'ed, Horacio Saggion, Alexander Shvets, and Alex Bravo. 2020. Automatic related work section generation: experiments in scientific document abstracting. *Scientometrics*, 125(3):3159–3185.

[11] Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190.

[12] Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A background knowledge- and content- based framework for citing sentence generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1466–1478, Online.

[13] Kelvin Luu, Xinyi Wu, Rik Koncel- Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2130–2144, Online.

[14] Nianlong Gu and Richard H. R. Hahnloser. 2023. Controllable citation sentence generation with language models.

[15] Shing- Yun Jung, Ting- Han Lin, Chia- Hung Liao, Shyan- Ming Yuan, and Chuen- Tsai Sun. 2022. Intentcontrollable citation text generation. *Mathematics*, 10(10):1763.

[16] Xiangci Li, Biswadip Mandal, and Jessica Ouyang. 2022. CORWA: A citation- oriented related work annotation dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5426–5440, Seattle, United States.

[17] Xiangci Li, Yi- Hui Lee, and Jessica Ouyang. 2023. Cited text spans for citation text generation. *arXiv preprint arXiv:2309.06365*.

[18] Biswadip Mandal, Xiangci Li, and Jessica Ouyang. 2024. Contextualizing generated citation texts. *arXiv preprint arXiv:2402.18054*.

[19] Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6068–6077, Online.

[20] Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Target- aware abstractive related work generation with contrastive learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 373–383.

[21] Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, Liang Hu, and Ivor Tsang. 2023. Causal intervention for abstractive related work generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2148–2159, Singapore.

[22] Anna Martin- Boyle, Aahan Tyagi, Marti A Hearst, and Dongyeop Kang. 2024. Shallow synthesis of knowledge in gpt- generated texts: A case study in automatic related work composition. *arXiv preprint arXiv:2402.12255*.

[23] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *ArXiv*, abs/1710.10903.

[24] Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089, Online.

[25] Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu- Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

[26] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content- impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.

[27] Uchenna Akujuobi and Xiangliang Zhang. 2017. Delve: a dataset- driven scholarly search and analysis system. *ACM SIGKDD Explorations Newsletter*, 19(2):36–46.

[28] 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.

[29] Mark Wasson. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1364–1368.

[30] Dragomir R Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

[31] Gunes Erkan and Dragomir R Radev. 2004. Lexrank: Graph- based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

[32] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

[33] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China.

[34] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long- document transformer. *arXiv:2004.05150.*

[35] Chin- Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[37] Eugene Garfield et al. 1965. Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings*, volume 269, pages 189–192. Washington.

[38] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110.

[39] Cailing Dong and Ulrich Schafer. 2011. Ensemble- style self- training on citation classification. In *Proceedings of 5th international joint conference on natural language processing*, pages 623–631.

[40] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

[41] Suppawong Tuarob, Sung Woo Kang, Poom Wettayakorn, Chanatip Pornprasit, Tanakitti Sachati, Saeed- Ul Hassan, and Peter Haddawy. 2019. Automatic classification of algorithm citation functions in scientific literature. *IEEE Transactions on Knowledge and Data Engineering*, 32(10):1881–1896.

[42] He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A context-based framework for modeling the role and function of on- line resource citations in scientific literature. 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 5206–5215.

[43] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

[44] Anne Lauscher, Brandon Ko, Bailey Kuhl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. Multicite: Modeling realistic citations requires moving beyond the single-sentence single- label setting. *arXiv preprint arXiv:2107.00414*.

[45] Awais Athar. 2011. Sentiment analysis of citations using sentence structure- based features. In *Proceedings of the ACL 2011 student session*, pages 81–87.

[46] Awais Athar and Simone Teufel. 2012. Contextenhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601, Montreal, Canada.

[47] Kumar Ravi, Srirangaraj Setlur, Vadlamani Ravi, and Venu Govindaraju. 2018. Article citation sentiment analysis using deep learning. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 78–85. IEEE.

[48] Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single- sentence single- label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States.

[49] Kokil Jaidka, Christopher Khoo, and Jin- Cheon Na. 2010. Imitating human literature review writing: An approach to multi- document summarization. In *International Conference on Asian Digital Libraries*, pages 116–119. Springer.

[50] Kokil Jaidka Jaidka, Christopher Khoo Khoo, and Jin- Cheon Na Na. 2011. Literature review writing: a study of information selection from cited papers/kokil jaidka, christopher khoo and jin- cheon na.

[51] Christopher SG Khoo, Jin- Cheon Na, and Kokil Jaidka. 2011. Analysis of the macro- level discourse structure of literature reviews. *Online Information Review*.

[52] Kokil Jaidka, Christopher Khoo, and Jin- Cheon Na. 2013a. Deconstructing human literature reviews- a framework for multi- document summarization. In *proceedings of the 14th European workshop on natural language generation*, pages 125–135.

[53] Kokil Jaidka, Christopher SG Khoo, and Jin- Cheon Na. 2013b. Literature review writing: how information is selected and transformed. In *Aslib Proceedings*. Emerald Group Publishing Limited.

[54] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana.

[55] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic- aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

[56] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[57] Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA.

[58] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy.

[59] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China.

[60] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen- tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge- intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

[61] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic.

[62] Judith L Klavans, Min- yen Kan, and Kathleen McKeown. 2001. Domain- specific informative and indicative summarization for information retrieval. *Proceedings of the Document Understanding Workshop.*

[63] Jeffrey W Knopf. 2006. Doing a literature review. *PS: Political Science & Politics*, 39(1):127–132.

Table 5: A summary of the problem formulations of the prior works on neural network-based related work generation. “Context” refers to those sentences or paragraphs around the target citation sentences.

Prior Work	Inputs	Target
AbuRa’ed et al. (2020)	Cited title, abstract	Citation sentence w/ single reference
Xing et al. (2020)	Context sentences, single cited abstract	Citation sentence w/ single reference
Ge et al. (2021)	Citation network, single cited abstract, context sentences	Citation sentence, citation function, salient sentence in cited abstracts
Luu et al. (2021)	Intro of the citing paper, named entities of the cited papers	Citation sentence w/ single reference
Li et al. (2022)	Context sentences w/o the target span, 1+ cited ab-	Citation span w/ 1+ citations

Table 6: A summary of the datasets of the prior works on related work generation.

Prior Work	Source Domain	Size	Dataset
Hoang and Kan (2010)	Papers from NLP and IR, manually curated topic tree	20 papers	RWSData a
Hu and Wan (2014)	ACL Anthology	1050 papers	N/A
Wang et al. (2018)	ACM digital library	8080 papers	Available b
Chen and Zhuge (2019)	ACL Anthology & IJCAI	25 papers	RWS-Cit c
Wang et al. (2019)	NLP conferences	50 papers	NudtRwG d
Deng et al. (2021)	ScisummNet (ACL)	11954 examples	N/A
AbuRa'ed et al. (2020)	ScisummNet (ACL)	940 + 15574 pairs	N/A
Xing et al. (2020)	ACL Anthology Network	1k + 85k examples	Available e
Ge et al. (2021)	ACL Anthology Network	1.2k + 84k examples	N/A
Luu et al. (2021)	S2ORC (CS)	622k citations from 154k papers	Extraction from S2ORC f
Li et al. (2022)	S2ORC (NLP)	Annotated 3565 dominant spans & 4228 reference spans from 2927 paragraphs; 565+362+11465 train/test/distant RWS	CORWA g
Jung et al. (2022)	SciCite (CS)	8243/916/1861 train/validation/test samples	Available h
Li et al. (2023)	CORWA, S2ORC (NLP)	1654/1206/19784 train/test/distant dominant citation spans	Available i
Gu and Hahnloser (2023)	arXiv computer science papers	233.6k/1.3k/1.1k train/validation/test samples	Available j
Mandal et al. (2024)	CORWA	565/362/11465 train/test/distant RWS	N/A
Chen et al. (2021)	S2ORC (Multi-domain), Delve (CS)	150k, 80k examples	Available k
Chen et al. (2022)	S2ORC (Multi-domain), Delve (CS)	107.7k/5k/5k, 208.3k/5k/5k train/dev/test examples	N/A
Liu et al. (2023)	S2ORC (Multi-domain), Delve (CS)	126k/5k/5k, 72k/3k/3k train/dev/test pairs	N/A
Li and Ouyang (2024)	PDFs from NLP, ML, Speech, CV, etc.	38 papers	N/A
Martin-Boyle et al. (2024)	2023 ACL best papers	10 papers	N/A

Table 7: A summary of the approaches of the prior works.

Prior Work	Approaches
Hoang and Kan (2010)	Heuristic approach to generate general and specific content separately given a topic tree
Hu and Wan (2014)	PLSA for topic modeling, SVR for sentence importance score, and global optimization for sentence selection
Wang et al. (2018)	Custom neural seq2seq model (CNN, LSTM, attention), random walk for encoding heterogeneous bibliography graph
Chen and Zhuge (2019)	Considering papers co-cite the cited papers; Representing graph for relationship modeling of papers, then finding sentence nodes that cover the minimum Steiner tree of the graph
Wang et al. (2019)	Leveraging both topic model and cited text spans
Deng et al. (2021)	BERT-based sentence extraction & reordering
AbuRa'ed et al. (2020)	Applying PTGen and Transformer
Xing et al. (2020)	Manual annotation + automatic annotation of citation sentences; PTGEN-Cross based on cross-attention mechanism
Ge et al. (2021)	Citation network as auxiliary input; citation function & salient sentences in cited papers as auxiliary output; multi-task learning
Luu et al. (2021)	SciGPT2; IE-Extracted Term Lists; ranking based on entity matching
Li et al. (2022)	LED-based citation span generation
Jung et al. (2022)	BART/T5-based citation sentence generation with citation intents
Li et al. (2023)	RAG & LED; ROUGE-based CTS retrieval
Gu and Hahnloser (2023)	Fine-tuned GPT-Neo & Galactica with Proximal Policy Optimization
Mandal et al. (2024)	Using citation context along with citation spans as generation target
Chen et al. (2021)	Transformer-based hierarchical encoder; relationship modeling module
Chen et al. (2022)	Improved over Chen et al. (2021) by encoding target paper's abstract
Liu et al. (2023)	Proposed a custom Causal Intervention Module (CaM) inserted between Transformer blocks
Li and Ouyang (2024)	GPT-3.5 for feature generation, e.g. faceted summary, relationship & usage of citations; GPT-4 based RWG
Martin-Boyle et al. (2024)	GPT-4 with human-in-the-loop

Table 8: A summary of the evaluation methods of the extractive related work generation works.

Prior Work	Baselines	Automatic	Human Evaluation
Hoang and Kan (2010)	LEAD, MEAD	ROUGE recall (1, 2, S4, SU4)	Correctness, novelty, fluency, usefulness
Hu and Wan (2014)	MEAD, LexRank	ROUGE F1 (1, 2, SU4)	Correctness, readability, usefulness
Wang et al. (2018)	Luhn, MMR, LexRank, SumBasic, NltkSum, Pointer Network	ROUGE F1 (1, 2, L)	Compliance to target paper, intuitiveness, usefulness
Chen and Zhuge (2019)	MEAD, LexRank, RoWoS	ROUGE F1 (1, 2)	N/A
Wang et al. (2019)	LexRank, SumBasic, JS-Gen, TopicSum	ROUGE recall & F1 (1, 2, SU4)	N/A
Deng et al. (2021)	MEAD	ROUGE precision, recall, F1 (1, 2, L)	informativeness, fluency, succinctness

Table 9: A summary of the evaluation methods of the abstractive related work generation works.

Prior Work	Baselines	Automatic	Human Evaluation
AbuRa'ed et al. (2020)	MEAD, TextRank, SUMMA, SEQ3	ROUGE precision, recall, F1 (1, 2, L, SU4)	N/A
Xing et al. (2020)	RandomSen, MaxSimSen, EXT-Oracle, COPY-CIT, PTGEN	ROUGE F1 (1, 2, L)	Readability, Content, Coherence, Overall
Ge et al. (2021)	LexRank, TextRank, EXT-Oracle, PTGEN, PTGEN-Cross	ROUGE F1 (1, 2, L)	Fluency, relevance, coherence, overall
Luu et al. (2021)	N/A	BLEU, ROUGE-L	Correct, Specific
Li et al. (2022)	Citation sentence generation	ROUGE F1 (1, 2, L)	Fluency, coherence, relevance, overall
Jung et al. (2022)	EXT-Oracle, ablations	ROUGE F1 (1, 2, L), SciBERTScore, citation intent accuracy	Correct, specific, plausible, intent
Li et al. (2023)	Citation span generation based on cited abstracts, & human-annotated CTS	BLEU, ROUGE-F1-L, METEOR, QuestEval, ANLI	Fluency, coherence, relevance, overall
Gu and Hahnloser (2023)	BART-base & -large, GPT-Neo 125M & 1.3B, Galactica 125M & 1.3B &6.7B, LLaMA-7B ablations, GPT-3.5-turbo	ROUGE F1 (1, 2, L), Intent alignment score, keyword recall, fluency score	Intent alignment, keyword recall, fluency & similarity to the ground truth
Mandal et al. (2024)	Ablations	N/A	Fluency, coherence, relevance, overall
Chen et al. (2021)	LEAD, TextRank, Bert-SumEXT, MGSum-ext, PTGen+Cov, TransformerABS, BertSumAbs, MGSum-abs, GS	ROUGE F1 (1, 2, L)	QA, informativeness, coherence, succinctness
Chen et al. (2022)	LEAD, LexRank, NES, BertSumEXT, MGSum, EMS, RRG, BertSumAbs	ROUGE F1 (1, 2, L, SU)	QA, informativeness, coherence, succinctness
Liu et al. (2023)	TexRank, BertSumEXT, MGSum-ext & -abs, TransformerABS, RRG, BertSumAbs, GS, T5-base, BART-base, Longformer, NG-abs, TAG	ROUGE F1 (1, 2, L)	QA, informativeness, coherence, succinctness
Li and Ouyang (2024)	Ablations	ROUGE F1 (1, 2, L)	Fluency, coherence, relevance (cited, target), factuality, usefulness, writing, overall, # of errors
Martin-Boyle et al. (2024)	Human & Human-in-the-loop	# of edges, average node degree, density, cluster coefficient	Qualitative analysis

Table 10: A summary of the perspectives for human evaluation.

Perspective	Definition	Used By
Fluency, Read-ability	Does the summary's exposition flow well, in terms of syntax as well as discourse?	Hoang and Kan (2010); Hu and Wan (2014); Deng et al. (2021); Xing et al. (2020); Ge et al. (2021); Chen et al. (2021, 2022); Li et al. (2022, 2023); Gu and Hahnloser (2023); Mandal et al. (2024); Li and Ouyang (2024)
Correctness	Is the summary content relevant to (express the factual relationship with) the hierarchical topics/cited papers given?	Hoang and Kan (2010); Hu and Wan (2014); Luu et al. (2021); Jung et al. (2022)
Novelty	Does the summary introduce novel information that is significant in comparison with the human created summary?	Hoang and Kan (2010)
Usefulness	Is the summary useful in supporting the researchers to quickly grasp the related works given hierarchical topics?	Hoang and Kan (2010); Hu and Wan (2014); Wang et al. (2018); Li and Ouyang (2024)
Content, Rele-vance	Whether the citation text is relevant to the cited paper's abstract	Wang et al. (2018); Xing et al. (2020); Ge et al. (2021); Li et al. (2022, 2023); Gu and Hahnloser (2023); Mandal et al. (2024); Li and Ouyang (2024)
Coherence	Whether the citation text is coherent with the citing paper's context	Xing et al. (2020); Ge et al. (2021); Chen et al. (2021, 2022); Li et al. (2022, 2023); Liu et al. (2023); Mandal et al. (2024); Li and Ouyang (2024)
Informativeness	Does the related work convey important facts about the topic question?	Deng et al. (2021); Chen et al. (2021, 2022); Liu et al. (2023)
Succinctness	Does the related work avoid repetition?	Deng et al. (2021); Chen et al. (2021); Liu et al. (2023)
Overall	Overall quality	Xing et al. (2020); Ge et al. (2021); Li et al. (2022, 2023); Mandal et al. (2024); Li and Ouyang (2024)
Intuitiveness	How intuitive is the related work section for readers to grasp the key content?	Wang et al. (2018)
QA	Retain the key information?	Chen et al. (2021, 2022); Liu et al. (2023)
Specific	Whether the explanation describes a specific relationship between the two works	Luu et al. (2021); Jung et al. (2022)
Factuality, # of errors	Does the output contain factual errors?	Li and Ouyang (2024)
Plausible, writ-ing	Writing style of citation text / RWS	Jung et al. (2022); Li and Ouyang (2024)
Qualitative analysis	Descriptive case study	Li and Ouyang (2024); Martin-Boyle et al. (2024)
Intent align-ment	Whether the output aligns with the input intent.	Jung et al. (2022); Gu and Hahnloser (2023)
Keyword recall	Whether the output contains the input key words.	Gu and Hahnloser (2023)