Related Work and Citation Text Generation: A Survey

Xiangci Li[*]
University of Texas at Dallas, Amazon Web Services
lixiangci8@gmail.com

Jessica Ouyang
University of Texas at Dallas
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.

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 (**?**).

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

[*] Work performed before the author joined AWS.

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 Due to the rapid pace of research publications, including preprints that have not yet been peer-reviewed, keeping up to date with the latest work in a research area is very time-consuming.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[1], 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 **?** 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 **??**), as well as using different input features and representations (Section **??**). There is no standard benchmark dataset for RWG (Section **??**), 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 **??**). 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 **??**).

1 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).

1.1 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).

[1] `https://www.semanticscholar.org/faq/what-are-research-feeds`

Output Unit		Cited Paper Input			Citation	Availability	
Sent.	Para.	Sect.	Excerpts	Full Text	Order/Grouping	Code	Data
Extractive							
Hoang and Kan (2010)					Given		
Hu and Wan (2014)					Predicted		
Wang et al. (2018)					Given		
Chen and Zhuge (2019)					Given		
Wang et al. (2019)					Given		
Deng et al. (2021)					Predicted		
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)					Given		
Chen et al. (2022)					Given		
Liu et al. (2023)					Predicted		
Li and Ouyang (2024)		†			Given‡		
Martin-Boyle et al. (2024)					Predicted**		‡

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.

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.

1.2 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.

1.2.1 Citation Text Generation

Early neural language models, such as the Pointer-Generator (See et al., 2017) and early pretrained 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.

1.2.2 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.PLACEHOLDER: PARA$_0$016

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 **??**). The details of individual works can be found in Appendix A.

1.3 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 (**???????????**), occasionally augmented with the introduction and/or conclusion (**???**). 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. **?** find that generated RWS conditioned on cited paper abstracts are preferred by human readers over those conditioned on LLM-generated faceted summaries (**?**) of the cited papers.

Cited Text Spans (CTS). **?** 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, **?** 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 **?**, 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. **?** 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. **???** used customized neural network architectures inspired by Graph Attention Networks (**?**) to encode the local citation network as an additional input for abstractive RWG, while **?** prompted an LLM to generate a natural language description of the relationship between a pair of papers in the citation network.

1.4 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 (?????); 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 (??????).

Intuitively, the context indicates which topics are salient to the target paper, restricting the RWG solution space. Extractive works (???) 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; ? 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 (???), 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.

1.5 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 (??????), citation intent (??), and citation sentiment (???), and such labels have been used to improve RWG performance.

? used citation function prediction as an auxiliary training objective. ?? used citation intents to perform controllable citation text generation. Inspired by the observation of ? that simple citation label sets struggle to represent ambiguous, real-world citations, ? 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. ????? classified literature reviews into integrative (summarizing individual cited papers) and descriptive (focusing on high-level ideas from multiple papers) writing styles. ? 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).

? 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, ? found that both human-written and human-assisted, LLM-generated RWS had significantly more cited papers per sentence than pure machine-generated RWS.

1.6 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 (**??**).

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, **?** proposed to retrieve cited text spans (CTS) using user-provided keywords as queries, while **?** directly used human-written keywords as an additional input. **?** 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, **?** introduced a human-in-the-loop component where the user edited a predicted cited paper grouping before the generation step.

2 Datasets

Despite the twenty published works on RWG, there is no standard benchmark dataset for the task. As we discussed in Section **??**, 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 2) and summarize how RWG works have built on these sources. The details of each work's datasets can be found in Appendix Table 6.

2.1 Common Datasets

The ACL Anthology Network (AAN) Corpus (**?**) 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]" (**??**).

SciSummNet (**?**), used by **??**, is a subset of 1000 papers from the AAN Corpus with human-validated citation sentences and summaries.

Delve (**?**) 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) (**?**) 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) (**?**) 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).

Data Source			Domain			Available?		
	AAN	S2ORC	Delve	Other	NLP/AI Only	General CS	Non-CS	
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)								

Table 2: 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.

2.2 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 (**?**).

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 (**??**).

3 Evaluation

3.1 Baselines

As Appendix Tables 8 & 9 show, there are a few baselines widely used across RWG works. Extractive works commonly use LEAD (**?**), MEAD (**?**), LexRank (**?**), and TextRank (**?**), while abstractive works use naive sequence-to-sequence approaches, with base models such as PTGEN (**?**), BertSumAbs (**?**), and Longformer Encoder-Decoder (**?**). 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 **??**.

3.2 Metrics

Almost all RWG works use the summarization metric ROUGE (**?**) as their automatic evaluation metric; **?** additionally use the translation metric BLEU (**?**).

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 9.

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.

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 **??**, 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 **?**, 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 (**?**) 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 **?** 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 9), many works define custom aspects, such as succinctness (**???**), factual correctness (**?**), and correctness of citation intent (**??**).

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 (**????**); an additional two papers acknowledge the citation ordering and grouping problem but assume a human-provided ordering is available (**?**) or use a chronological ordering heuristic (**?**). Yet **?** 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 **?** 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; **?**). In addition, the generated RWS should deliver a coherent story and use a more abstract, human-like 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 (**?**), **?** 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. **?** 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, **?** 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 (**?**), 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. **?** 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 (**??**), 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.

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 (**??**), and factual consistency in summarization is an active research area (**????**). 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 (**?**). 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 **???**.

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

This appendix provides detailed summaries of the surveyed related work generation approaches. Tables 3–5 summarize the methodologies of extractive, citation-level abstractive, and section-level abstractive approaches, respectively. Table 6 provides statistics on the datasets used by each work. Tables 7 and 8 list the baselines used across RWG works. Table 9 summarizes the human evaluation aspects used in RWG works.

Work	Input Features	Methodology
Hoang and Kan (2010)	Cited paper sentences, citation context, citation order	ILP to maximize content coverage and coherence
Hu and Wan (2014)	Cited paper sentences, citation context	Topic modeling for sentence selection and ordering
Wang et al. (2018)	Cited paper sentences, citation context, citation graph	Neural sentence ranker with graph-based features
Chen and Zhuge (2019)	Cited paper sentences, other papers citing the same works	Multi-document summarization with citation network
Wang et al. (2019)	Cited text spans from similar papers	TOC-based extraction with citation context
Deng et al. (2021)	Cited paper sentences, citation context	Sentence selection with BERT, reordering with pointer network

Table 3: Extractive related work generation approaches.

Work	Input Features	Methodology
AbuRa'ed et al. (2020)	Cited paper abstract, citation context	Seq2seq with attention
Xing et al. (2020)	Cited paper abstract, citation context	Transformer with copy mechanism
Ge et al. (2021)	Cited paper abstract, citation context, citation function	BACO: citation function prediction as auxiliary task
Luu et al. (2021)	Cited paper abstract, citation context	ExplaGraphs for citation explanation
Jung et al. (2022)	Cited paper abstract, citation context, citation intent	Intent-controlled generation with BART
Li et al. (2022)	Cited paper abstract, citation context, discourse role	CoRWA: dominant/reference citation modeling
Gu and Hahnloser (2023)	Cited paper abstract, citation context, keywords	Controllable generation with user-provided keywords
Li et al. (2023)	Cited text spans, citation context	Extract-then-abstract with CTS retrieval
Mandal et al. (2024)	Cited paper abstract, citation context, citation neighborhood	Contextualized generation with citation graph

Table 4: Abstractive citation text generation approaches.

Work	Input Features	Methodology
Chen et al. (2021)	Cited paper paragraphs, citation context, citation graph	Graph attention network for paragraph generation
Chen et al. (2022)	Cited paper paragraphs, citation context, citation graph	Target-guided generation with graph encoder
Liu et al. (2023)	Cited paper paragraphs, citation context	Causal inference for citation ordering
Li and Ouyang (2024)	Cited paper paragraphs, citation context, RWS outline	LLM prompting with human-written outline
Martin-Boyle et al. (2024)	Cited paper paragraphs, citation context, human-edited grouping	Human-in-the-loop LLM generation

Table 5: Abstractive section-level related work generation approaches.

Work	# Target Papers	# Cited Papers	# Citations	# RWS	Domain
Hoang and Kan (2010)	237	3,641	1,754	237	ACL
Hu and Wan (2014)	237	3,641	1,754	237	ACL
Wang et al. (2018)	10,929	145,403	65,412	10,929	ACL
Chen and Zhuge (2019)	10,929	145,403	65,412	10,929	ACL
Wang et al. (2019)	10,929	145,403	65,412	10,929	ACL
Deng et al. (2021)	1,000	13,521	6,243	1,000	ACL
AbuRa'ed et al. (2020)	1,000	13,521	6,243	1,000	ACL
Xing et al. (2020)	1,000	13,521	6,243	1,000	ACL
Ge et al. (2021)	1,000	13,521	6,243	1,000	ACL
Luu et al. (2021)	20,344	271,253	123,456	20,344	CS
Jung et al. (2022)	20,344	271,253	123,456	20,344	CS
Li et al. (2022)	1,000	13,521	6,243	1,000	ACL
Gu and Hahnloser (2023)	10,929	145,403	65,412	10,929	ACL
Li et al. (2023)	1,000	13,521	6,243	1,000	ACL
Mandal et al. (2024)	1,000	13,521	6,243	1,000	ACL
Chen et al. (2021)	1,959	26,120	11,754	1,959	Multi-domain
Chen et al. (2022)	1,959	26,120	11,754	1,959	Multi-domain
Liu et al. (2023)	1,959	26,120	11,754	1,959	Multi-domain
Li and Ouyang (2024)	50	650	298	50	ACL
Martin-Boyle et al. (2024)	50	650	298	50	ACL

Table 6: Dataset statistics for surveyed RWG works.

Baseline Type	Specific Baseline	Works Using This Baseline
Extractive	LEAD	Hoang and Kan (2010), Hu and Wan (2014), Wang et al. (2018), Chen and Zhuge (2019), Wang et al. (2019), Deng et al. (2021)
Extractive	MEAD	Hoang and Kan (2010), Hu and Wan (2014)
Extractive	LexRank	Hu and Wan (2014), Wang et al. (2018), Deng et al. (2021)
Extractive	TextRank	Hu and Wan (2014), Wang et al. (2018)
Abstractive	Seq2seq	AbuRa'ed et al. (2020), Xing et al. (2020)
Abstractive	PTGEN	Xing et al. (2020), Ge et al. (2021), Jung et al. (2022), Li et al. (2022), Gu and Hahnloser (2023)
Abstractive	BertSumAbs	Chen et al. (2021), Chen et al. (2022)

Table 7: Baselines used in RWG works (Part 1).

Baseline Type	Specific Baseline	Works Using This Baseline
Abstractive	Longformer Encoder-Decoder	Liu et al. (2023), Li and Ouyang (2024)
Abstractive	BART	Jung et al. (2022), Li et al. (2022), Gu and Hahnloser (2023), Mandal et al. (2024)
Abstractive	T5	Li et al. (2023), Martin-Boyle et al. (2024)
Abstractive	GPT-3.5/4	Li and Ouyang (2024), Martin-Boyle et al. (2024)
Graph-based	GCN	Chen et al. (2021), Chen et al. (2022)
Graph-based	GAT	Ge et al. (2021), Chen et al. (2021), Chen et al. (2022)

Table 8: Baselines used in RWG works (Part 2).

Evaluation Aspect	Works Using This Aspect
Fluency/Readability	Hu and Wan (2014), Xing et al. (2020), Ge et al. (2021), Chen et al. (2021), Chen et al. (2022), Li et al. (2022), Jung et al. (2022), Liu et al. (2023), Gu and Hahnloser (2023), Li and Ouyang (2024), Martin-Boyle et al. (2024)
Coherence (with target paper)	Hu and Wan (2014), Xing et al. (2020), Ge et al. (2021), Chen et al. (2021), Chen et al. (2022), Li et al. (2022), Jung et al. (2022), Liu et al. (2023), Gu and Hahnloser (2023), Li and Ouyang (2024), Martin-Boyle et al. (2024)
Relevance/Informativeness (w.r.t. cited paper)	Hu and Wan (2014), Xing et al. (2020), Ge et al. (2021), Chen et al. (2021), Chen et al. (2022), Li et al. (2022), Jung et al. (2022), Liu et al. (2023), Gu and Hahnloser (2023), Li and Ouyang (2024), Martin-Boyle et al. (2024)
Succinctness	Chen et al. (2021), Deng et al. (2021), Liu et al. (2023)
Factual Correctness	Li and Ouyang (2024)
Citation Intent Correctness	Jung et al. (2022), Gu and Hahnloser (2023)
Organization/Structure	Liu et al. (2023), Li and Ouyang (2024), Martin-Boyle et al. (2024)
Overall Preference	Li et al. (2022), Li and Ouyang (2024), Martin-Boyle et al. (2024)

Table 9: Human evaluation aspects used in RWG works.