CORWA: A Citation-Oriented Related Work Annotation Dataset

Xiangci Li Biswadip Mandal Jessica Ouyang
Department of Computer Science
University of Texas at Dallas
Richardson, TX 75080
`lixiangci8@gmail.com`,
`{Biswadip.Mandal, Jessica.Ouyang}@UTDallas.edu`

Abstract

Academic research is an exploratory activity to discover new solutions to problems. By this nature, academic research works perform literature reviews to distinguish their novelties from prior work. In natural language processing, this literature review is usually conducted under the "Related Work" section. The task of related work generation aims to automatically generate the related work section given the rest of the research paper and a list of papers to cite. Prior work on this task has focused on the sentence as the basic unit of generation, neglecting the fact that related work sections consist of variable length text fragments derived from different information sources. As a first step toward a linguistically-motivated related work generation framework, we present a Citation Oriented Related Work Annotation (CORWA) dataset that labels different types of citation text fragments from different information sources. We train a strong baseline model that automatically tags the CORWA labels on massive unlabeled related work section texts. We further suggest a novel framework for human-in-the-loop, iterative, abstractive related work generation.

1 Introduction

Academic research is an exploratory activity to solve problems that have never been solved before. By this nature, each academic research work must sit at the frontier of its field and present novel contributions that have not been addressed by prior work; in order to convince readers of the novelty of the current work, the authors must compare against the prior work. While the format may vary among different fields, in natural language processing (NLP), this literature review is usually conducted under the "Related Work" section. Since each paper must review the relevant prior work in its field, which is shared among papers on the same topic or task, many related work sections in a given field can be similar in both content and format. Therefore,

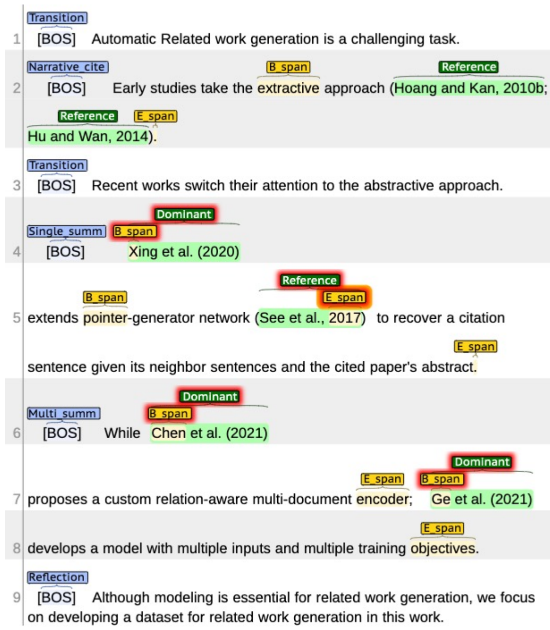

Figure 1: An example of CORWA labels displayed using the BRAT interface (Stenetorp et al., 2012).

it is a natural motivation to develop a system for generating related work sections automatically.

The task of automatic related work generation is that of generating the related work section of a target paper given the rest of the target paper and a set of papers to cite. Prior works (Hoang and Kan, 2010; Hu and Wan, 2014; Chen and Zhuge, 2019; Wang et al., 2019; Xing et al., 2020; Ge et al., 2021; Luu et al., 2021; Chen et al., 2021) mostly simplify related work generation as a general summarization task, generating related work sections using sentence-level models. This approach ignores the nature of the related work section, which consists of variable-length text fragments derived from different information sources. These text fragments refer to different cited papers, and they range in length from a few words to multiple sentences. There are also non-citation, supporting sentences that serve various discursive roles, such as introducing new topics, transitioning between topics, or reflecting on the current work. We argue it is necessary to distinguish these heterogeneous text fragments, rather than treating related work sections as concatenations of homogeneous sentences.

In addition to the heterogeneous information sources for related work section sentences, the writing styles of these sentences also vary. Khoo et al. (2011) classify literature reviews to be integrative or descriptive, depending on whether they focus on high-level ideas or provide more detailed information on specific studies. However, this document-level classification scheme was intended as a descriptive, information science study of related work sections, and it has not been previously used in automatic related work generation.

Inspired by these observations, as a first step towards linguistically-motivated related work generation, we present a Citation Oriented Related Work Annotation (CORWA) dataset of related work sections from NLP papers. We distinguish text fragments from different information sources by tagging each sentence with discourse labels and identifying the spans of tokens belonging to each citation. We further distinguish citations that give detailed explanations of cited papers and those that illustrate high-level concepts.

Our main contributions are as follows: (1) We collect a CORWA dataset that decomposes the related work section with three inter-related annotation tasks — discourse tagging, citation span detection, and citation type recognition — and demonstrate the significance of CORWA with analyses from multiple perspectives (§3). (2) We propose a strong baseline model that automatically tags the CORWA annotation scheme on massive unlabeled related work section texts (§4). (3) We show that citation spans are a better target than citation sentences with two example tasks (§5). (4) We discuss a novel framework for human-in-the-loop, iterative, abstractive related work generation (§6).

2 Related Work

Extractive Related Work Generation. Early related work generation systems employed the extractive summarization approach. Hoang and Kan (2010) pioneered the task, developing rules to select sentences following a topic hierarchy tree that was assumed to be given as input. Hu and Wan (2014) grouped sentences into topic-biased clusters with PLSA, modeled sentence importance with SVR, and applied a global optimization framework to select sentences. Chen and Zhuge (2019) selected sentences from papers that co-cited the same cited papers as the target paper in order to cover a minimum Steiner tree constructed from the paper's keywords. Wang et al. (2019) extracted Cited Text Spans (CTS), the matched text spans in the cited paper that are most related to a given citation. However, these extractive approaches aim to maximally cover the citation texts with the extracted sentences, thus mostly ignoring the *reference* type citations that are concise and abstractive (§3.1.3).

Abstractive Related Work Generation. Recently, Xing et al. (2020) extend the pointer-generator (See et al., 2017) to take two text inputs, allowing them to recover a masked citation sentence given its neighboring context sentences. Ge et al. (2021) encode the citation context, cited paper's abstract, and citation network and train their model with multiple objectives: sentence salience score regression of the cited paper's abstract, functional role classification of the citation sentence, and citation sentence generation. Chen et al. (2021) propose a relation-aware, multi-document encoder to generate a related work paragraph given a set of cited papers. Luu et al. (2021) fine-tune GPT2 (Radford et al., 2019) on scientific texts and explore several techniques for representing documents, such as using extracted named entities.

All of the works described above focus on the generation aspect, while neglecting dataset collection; their datasets are mostly extracted automatically. Moreover, the datasets are not reused, though they are publicly available, because these works all use slightly different problem definitions, and thus the models are not directly comparable (Li and Ouyang, 2022). In this work, we focus on collecting a dataset that is widely applicable to various related work generation settings, rather than proposing another incomparable approach.

3 CORWA Dataset

In this work, we limit our scope to publications from the NLP domain for ease of automatically extracting the related work section; existing work on related work generation has also focused on NLP in the past. We build our dataset on top of the NLP partition of the S2ORC dataset (Lo et al., 2020), a large-scale corpus of scientific papers derived from LaTeX source code and PDF files. We extract the related work section by matching the section titles. Because not all papers cited in the extracted related work sections are available in S2ORC, we prioritize annotating related work sections where the majority of their cited papers are available.

3.1 Annotation Scheme

Our CORWA dataset decomposes the related work section with three inter-related annotation tasks: discourse tagging, citation span detection, and citation type recognition.

3.1.1 Discourse Tagging

Each sentence in a related work section has a specific role and information source. Some may be general topic or transition sentences; some summarize one or multiple prior works in detail, while others describe the general relationship among prior works at a high level. Our discourse tagging task tags the role of each related work sentence with one of six labels: {*single_summ*, *multi_summ*, *narrative_cite*, *reflection*, *transition*, *other*}.

Single Document Summarization. *Single_summ* refers to sentences that summarize one single cited work in detail. Most typically, this includes sentences with explicit citation marks, as when a work is mentioned for the first time. We also include the following cases: (1) follow-up sentences without explicit citation marks that describe the same paper as a preceding *single_summ* sentence, and (2) sentences containing multiple citations that heavily focus on one of those works.

Multi-Document Summarization. *Multi_summ* refers to sentences that summarize multiple prior works of equal importance. As with *single_summ*, we include the case of follow-up sentences without explicit citation marks that continue describing the same group of prior works discussed in a preceding *multi_summ* sentence.

Narrative Citation. In contrast to *single_summ* and *multi_summ*, narrative citation (*narrative_cite*) refers to citation sentences that do not summarize specific cited works in detail, but rather convey high-level observations from the authors of the current work. *Narrative_cite* sentences may contain general statements about the field or task, or the authors' comments on or comparisons of prior works.

Reflection. In addition to describing prior works, authors discuss how they relate to the current work, highlighting the authors' novel contributions. These *reflection* sentences focus on the current work, instead of prior works.

Transition. Non-citation sentences in related work sections serve as topic introductions or transitions from one topic to another. We label these supplemental sentences that do not belong to any of the above cases as *transition* sentences.

Other. The related work sections in our dataset are extracted automatically using heuristics based on section titles, and there are occasionally some errors in section boundary detection; we label those sentences that are not actually part of the related work section as *other*.

3.1.2 Citation Span Detection

In order to understand sentences that describe prior work, it is crucial to recognize the token-level mapping between the citation text and the cited paper(s). Our citation span detection task identifies the span of text whose information is directly derived from a specific cited paper. For example, if a cited paper is explained with a summary, its citation span covers the entire summary, which may range from part of a sentence to a few consecutive sentences; if a cited paper is mentioned with an explicit citation, but is not described or discussed at all, then the citation span is just the citation mark.

In constructing the dataset, we find that a single citation rarely spans across paragraph boundaries without a new explicit citation mark, so we require our spans to be bounded by paragraph boundaries.

3.1.3 Citation Type Recognition

Our citation type recognition task indicates whether a cited work is discussed in detail or used to illustrate a high-level concept. We label these types of citations as *dominant* and *reference*, respectively.

Dominant. These citations are discussed in detail, usually via summarization of their content, and are often longer than *reference* citations.

Reference. These citations are not discussed in detail. They frequently appear in *narrative_cite* sentences, but may also appear in *single_summ* and *multi_summ* sentences when they are not the main focus of the sentence, and thus it is not sufficient to depend on the sentence-level discourse tags to distinguish them. For example, in Figure 1, line 5, the pointer-generator network (See et al., 2017) is cited for reference as part of a longer *dominant* citation span. *Reference* citations tend to be more abstractive than *dominant* citations.

Disc. Label (d)	$n(d)$	$p(d)$	$p(d \mid D)$	$p(d \mid R)$	$p(D \mid d)$	$p(R \mid d)$	$p(D, d)$	$p(R, d)$
single_summ	4255	30.8%	80.8%	1.1%	98.5%	1.5%	**36.9%**	0.6%
transition	3371	24.4%	0	0.2%	12.5%	87.5%	0	0.1%
narrative_cite	2540	18.4%	0.4%	90.2%	0.4%	99.6%	0.2%	**48.9%**
reflection	2489	18.0%	0.1%	6.1%	1.5%	98.5%	0.1%	3.3%
multi_summ	671	4.8%	18.7%	2.5%	86.4%	13.6%	**8.5%**	1.3%
other	510	3.7%	0	0	0	100.0%	0	0

Table 1: Distributions of discourse labels and citation spans in CORWA dataset. d: Discourse labels. D/R: *Dominant/reference* type citation span. $n(D) = 3565, n(R) = 4228$. 2927 paragraphs in total.

3.2 Annotation Process and Agreement

Two graduate students from our university's Computer Science Department[1], manually annotated 927 related work sections. They first annotated 23 related work sections from scratch, after which we incrementally trained a transformer-based tagging model (Vaswani et al., 2017) (§4) to assist the annotation process, asking the annotators to correct the model's predictions, rather than performing manual annotation from scratch. We split the 362 annotated related work sections from papers published in 2019 and later as our test set and all 565 earlier papers as the training set.

Since each related work section is labeled by a single annotator, we calculate agreement by sampling 50 related work sections from the test set and asking the other annotator to re-annotate them from scratch[2]. We obtain strong agreement on all tasks (Cohen's κ of 0.824, 0.965 and 0.878 for discourse tagging, citation type recognition, and citation span detection, respectively; citation type recognition and citation span detection are converted to token-level labels for agreement calculation).

The automated, correction-based annotation process is much faster than annotating from scratch and allows us to collect a much larger annotated dataset. As a trade-off, the annotations may be biased by the model's predictions if the annotators fail to notice any incorrect predictions. This may explain why our model performance reported in §4.2 is higher than the inter-annotator agreement.

3.3 Analysis of CORWA

The tasks of discourse tagging, citation span detection, and citation type recognition, capture distinct but overlapping perspectives of information.

3.3.1 Relations among CORWA Subtasks

We investigate the relationships among the CORWA subtasks by calculating the co-occurrence

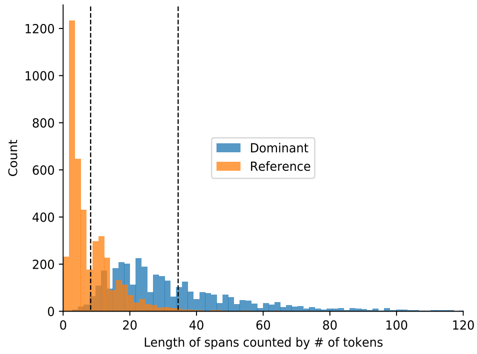

Figure 2: Histogram of the length of *dominant* and *reference*-type citation spans, excluding citation marks. The dashed vertical lines are the means of *dominant* and *reference* span lengths, 34.5 and 8.2, respectively.

distributions of discourse labels and citation span types. A citation span is considered *dominant* if it contains any *dominant* citations, and *reference* otherwise. Figure 2 shows that *dominant*-type spans (average of 34.5 tokens) are significantly longer than *reference*-type spans (average of 8.2 tokens).

Table 1 shows the count of each discourse label, the conditional probability and the joint probability of discourse labels and citation span types. *Single_summ* with *dominant* span, *multi_summ* with *dominant* span, and *narrative_cite* with *reference* span are the most frequent combinations . These observations make intuitive sense, since *dominant*-type spans describe cited papers in detail, often taking the form of a summary, while *reference*-type spans are highly abstracted, making them more likely to be mixed into *narrative*-type sentences that discuss high-level ideas, often encompassing multiple cited papers. This difference is analogous to *informative* versus *indicative* summaries, where the former serves as a surrogate for the document, and the latter characterizes what the document is about (Kan et al., 2001).

3.3.2 Related Work Writing Styles

Integrative or Descriptive? As Khoo et al. (2011) note, authors may describe the same cited paper in two different styles: descriptive, which ex-

[1]One of them later became the second author of this paper.
[2]The disagreements are adjudicated by the first author.

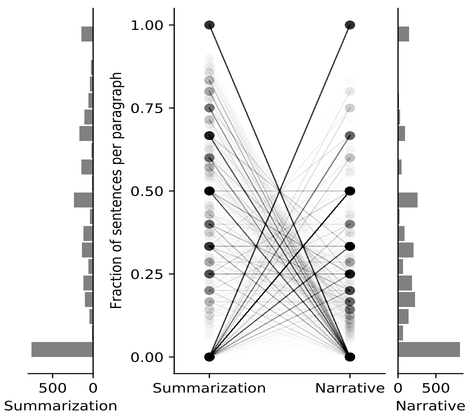

Figure 3: Parallel plot of the proportion of *summarization* and *narrative* sentences in each paragraph. Paragraphs with neither type of sentences are excluded.

plicitly summarizes the cited paper, or integrative, which describes and comments on the cited paper in a narrative form. We examine the ratio of *summarization* (both *single_summ* and *multi_summ*) and *narrative* sentences (*narrative_cite*) in related work paragraphs (Figure 3). The CORWA discourse labels capture writing style differences among papers: 34.6% of related work section paragraphs only contain *summarization* sentences, resembling Khoo et al.'s descriptive literature review, while 32.1% of paragraphs contain only *narrative* sentences, resembling an integrative literature review. Interestingly, 33.3% of paragraphs mix both styles and are neither purely descriptive nor purely integrative.

Frequent Discourse Label Subsequences. Scientific discourse is used by paper authors to promote their ideas (Li et al., 2021a). We analyze the patterns of CORWA discourse labels to uncover how authors promote their ideas using a mix of sentence types. We apply the rule-based PrefixSpan (Han et al., 2001) and Gap-Bide (Li and Wang, 2008) algorithms to extract frequent discourse label subsequences. We identify six typical subsequences, shown in Supplementary Tables 7 and 8. For example, the pattern of *single_summ* followed by *reflection* compares the cited paper to the current work, usually without directly criticizing the cited paper, while *single_summ* followed by *transition* is the more impersonal pattern for criticism of a cited paper, where authors tend to avoid direct comparison with the current work.

4 Joint Related Work Tagger

To help propagate our CORWA annotations to massive unlabeled related work sections, we build a joint related work tagger baseline[3] that is trained on the three annotation tasks, discourse tagging, citation span detection, and citation type recognition, via multi-task learning (Caruana, 1997).

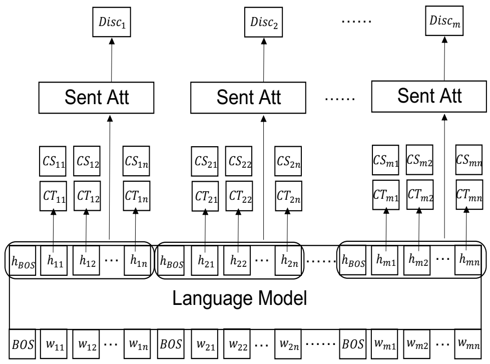

Figure 4: The architecture of our joint related work tagger, which performs discourse tagging (Disc), citation type recognition (CT), and citation span detection (CS).

4.1 Model Design

Figure 4 shows the model architecture of our joint related work tagger. We encode related work sections using a transformer-encoder (Vaswani et al., 2017) paragraph by paragraph, as we enforce the independence of paragraphs in CORWA citation span annotations. We decode citation span labels and citation type labels token by token, while our discourse tagging task uses the paragraph-level sentence tagging mechanism proposed by Li et al. (2021b). Because the three sub-tasks of CORWA are inter-related, we use multi-task learning to jointly train the tagger by sharing the encoder across tasks.

4.1.1 Paragraph Encoder

We experiment with several pre-trained transformer-encoders (Devlin et al., 2019; Beltagy et al., 2019; Liu et al., 2019; Beltagy et al., 2020), and eventually focus on SciBERT (Beltagy et al., 2019), which is a variant of the BERT model (Devlin et al., 2019) that is trained on a scientific corpus with domain-specific tokenization schemes, including NLP papers.

4.1.2 Task-specific Decoders

Citation Span Detection & Citation Type Recognition. We use the $BIO2$ tagging scheme (Sang and Veenstra, 1999) for the citation span detection and citation type recognition tasks; we use *B*, *I*, *O* for citation span detection and five labels — *B-Dominant*, *I-Dominant*, *B-Reference*, *I-Reference*,

[3] https://github.com/jacklxc/CORWA

Model	Disc	CT	CS
SciBERT	0.898	0.959	0.930
+ Distant Dataset	**0.908**	**0.963**	**0.933**

Table 2: Test set micro-F1 scores of the SciBERT-based joint related work tagger, with and without training on distantly labeled data, on the discourse tagging (Disc), citation type recognition (CT), and citation span detection (CS) tasks.

Parameter Name	Value
Encoder Learning Rate	10^{-5}
Decoder Learning Rate	5×10^{-6}
Dropout	0
Epoch	15
Batch Size	1
Steps per Update	10
γ_d	1
γ_t	3
γ_s	1.75

Table 3: Hyper-parameters of our best joint related work tagger (SciBERT + Distant Dataset).

and O — for citation type recognition. We use a two-layer feed-forward network to decode the encoded paragraph-level token embeddings to the output sequence of $BIO2$ tags.

Discourse Tagging. We apply Li et al. (2021b)'s paragraph-level sentence tagging approach for the discourse labels: a simple attention mechanism is used to aggregate token embeddings, sentence by sentence, into sentence encodings, before decoding the sentence encodings into discourse labels using a two-layer multi-layer feed-forward network.

4.1.3 Multi-task Learning

We use cross-entropy loss on all three CORWA sub-tasks. We balance the relative importance of the sub-tasks by taking a weighted sum of the sub-task losses of discourse tagging, citation span detection, and citation type recognition $\{L_d, L_s, L_t\}$:

$$L = \gamma_d L_d + \gamma_s L_s + \gamma_t L_t \quad (1)$$

where $\{\gamma_d, \gamma_s, \gamma_t\}$ are tuned hyper-parameters; their values are given in Table 3.

4.2 Experiments

We perform five-fold cross-validation to tune the model hyper-parameters. Table 2 shows the strong performance of the model[4]. We use the joint related work tagger to automatically label the unannotated 11,465 related work sections remaining in the S2ORC NLP partition and then use this distantly-supervised data to further boost the model's performance. For the citation span detection and citation type recognition tasks, we use a token-level F1 score. Our final, distantly-supervised joint related work tagger achieves more than 0.9 test F1 on all three tasks, indicating the high quality of the model's predictions. This model can be used to propagate our labels on the unannotated related work sections to create a very large training set for future work.

[4]Supplementary Table 6 shows the full cross-validation and test performances.

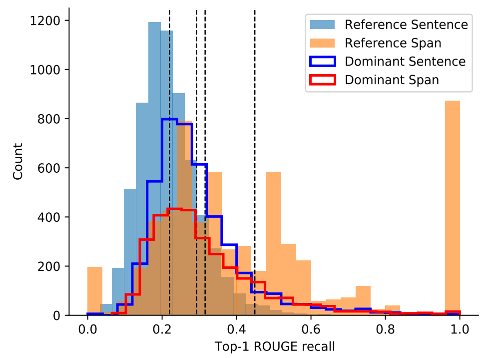

Figure 5: Histogram of top-1 ROUGE recall scores of retrieved sentences from cited papers using different queries. The dashed vertical lines are the means of reference sentence (0.220), dominant sentence (0.293), dominant span (0.316), and reference spans (0.449).

5 Spans as an Alternative to Sentences

We argue that the citation spans annotated in CORWA are a better alternative to the citation sentences that have previously been used for the tasks of ROUGE-based retrieval and citation text generation.

5.1 Queries for Relevant Sentence Retrieval

Citations focus on a small portion of the content in cited papers, and this focus is not explicitly recorded in the citation network. A popular approach for determining relevant sentences retrieves sentences from the cited papers by comparing the similarity between the gold citation sentence and candidate sentences in the cited paper (Cao et al., 2015; Yasunaga et al., 2017, 2019; Ge et al., 2021). Figure 5 compares the distribution of the top-1 average of ROUGE-1 and ROUGE-2 recall scores (Lin, 2004) of retrieved sentences from cited papers using citation spans with those using citation sentences[5]. There is no significant difference between the average ROUGE scores of *dominant* spans and sentences containing *dominant* citations, which is

[5]Only papers included in S2ORC dataset are considered.

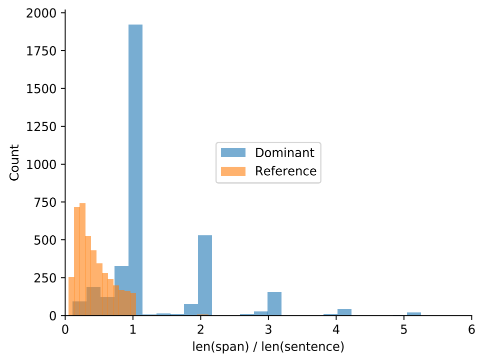

Figure 6: Histogram of the ratio of between the lengths of *dominant* and *reference* type citation spans and the corresponding citation sentences. None of the reference spans are longer than one sentence. 27.7%, 46.6%, and 25.7% of *dominant* spans are shorter than, equal to, or longer than one sentence, respectively.

reasonable because *dominant* spans are often full sentences anyway. In contrast, the average score of *reference* spans is significantly higher than that of sentences containing *reference*-type citations; *reference* spans are shorter and contain highly concentrated key information derived from their cited papers. Thus, using CORWA citation spans as queries for ROUGE-based cited sentence retrieval is superior for *reference*-type citations and comparable for *dominant*-type citations.

5.2 Span-based Related Work Generation

Existing neural network-based, abstractive related work generation systems generate citation sentences given the surrounding context sentences (Xing et al., 2020; Ge et al., 2021; Luu et al., 2021) or generate entire paragraphs containing multiple citations (Chen et al., 2021). These task settings neglect the fact that the citation text corresponding to a cited paper is not necessarily in the form of a sentence, but could be a portion of a sentence or a block of multiple sentences. Our span-based annotation scheme identifies the citation tokens that are directly derived from the cited papers.

As Figure 6 shows, *reference* spans are not full sentences, while *dominant* spans can cover multiple sentences. For *reference*-type citations, using a full sentence as the generation target includes potentially unrelated tokens outside the citation span that do not refer to the cited paper. For *dominant*-type citations, using a single sentence as the generation target can result in 1) information loss when not all sentences describing the cited paper are included in the target, and the model never learns to generate them, or 2) information leak when sentences that actually describe the cited paper are used as context sentences instead of target sentences. Thus, we propose a span-level citation text generation task and present a pilot study using a Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) baseline model.

5.2.1 Experimental Setting

The common Transformer-based language models (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020) have a limited input window size (typically 512 or 1024 tokens), which presents a major challenge for tasks like related work generation that use multiple long documents as inputs. LED (Beltagy et al., 2020) addresses this challenge by using a local self-attention mechanism, rather than global self-attention, handling in input context windows of up to 16k tokens. We present an LED-based baseline model for the citation span generation task.

We first pretrain the LED-base model on the masked language modeling (MLM) task (Devlin et al., 2019) using related work sections from S2ORC papers in the computer science domain, as well as on the cross-document language modeling (CDLM) task (Caciularu et al., 2021), which aligns masked citation sentences with their context sentences and the full text of their cited papers. We further pretrain the LED encoder with the three CORWA sub-tasks (Supplementary Table 6). All pretraining strictly excludes the texts from test set.

For the citation span generation task, we input the concatenation of {the target paper's introduction (following Luu et al. (2021)), the partial related work paragraph excluding the target citation span, and the concatenation of {explicit citation mark, title, and abstract} of each cited paper in the target span[6]}; the generation target is the ground truth citation span from CORWA. We provide the explicit citation mark (e.g. Devlin et al., 2018) because it is simple to extract but cannot be inferred from the paper text alone. Just as a human reader may remember the content of the frequently cited papers or the research topics of frequently cited authors, so the citation mark tokens may carry information about the cited paper and its authors.

In addition to the CORWA training set, we use the distantly supervised labels predicted by our joint related work tagger (§4.2) for training. We use the default hyper-parameters of the Huggingface LED implementation (Wolf et al., 2020).

[6]We indicate whether the target span is *dominant* or *refer-*

Models	*Dominant* R-1	R-2	R-L	*Reference* R-1	R-2	R-L
LED-base w/o pretrain	0.220	0.060	0.183	0.228	0.091	0.223
LED-base Span	0.230	0.062	0.186	0.244	0.107	0.240
LED-base Sentence	0.244	0.075	0.202	0.193	0.050	0.151

Table 4: Performance of citation span/sentence generation using LED-base (Beltagy et al., 2020). Citation marks are excluded from the scores since they are trivial to generate and bring up the scores unintentionally. Note that the performance of span/sentence generations are NOT directly comparable due to different generation targets.

	Flu.	Rel.	Coh.	Overall
Dominant				
Gold Span	4.61	3.53	4.17	3.64
Span	**4.92**	**4.07**	**4.20**	3.99
Sentence	4.83	4.03	4.17	**4.02**
Reference				
Gold Span	**4.87**	4.04	4.18	**4.00**
Span	4.68	**4.24**	**4.26**	3.96
Sentence	4.86	3.64	4.09	3.70

Table 5: Average fluency, relevance, coherence and overall scores, rated by human judges.

5.2.2 Experimental Results

As Table 4 shows, the ROUGE scores of our LED-base models for citation span/sentence generation are similar to previous sentence-level citation text generation models (Xing et al., 2020; Ge et al., 2021), and our pretraining improves the citation span generation performance. Compared to sentence-level generation, span-level generation has lower scores for *dominant* citations, but higher scores for *reference* citations. However, because the span- and sentence-level tasks have different generation targets, their scores cannot be directly compared.

We perform a human evaluation following the setting of Xing et al. (2020); Ge et al. (2021). We sample 15 instances each for *dominant* and *reference* citations and compare their corresponding span- and sentence-based generation outputs, as well as the gold spans from the original related work sections. Each citation text is rated by three NLP graduate students who are fluent in English on a 1 (very poor) to 5 (excellent) point scale, with respect to four aspects: *fluency* (whether a citation span/sentence is fluent), *relevance* (whether a citation span/sentence is relevant to the cited paper(s)), *coherence* (whether a citation span/sentence is coherent within its context), and *overall quality*.

Table 5 shows human evaluation results, with moderate inter-annotator agreement (Kendall's τ of 0.298, 0.205, and 0.172 among three annotators). All citation texts are judged to be highly fluent.

Interestingly, in previous studies (Xing et al., 2020; Ge et al., 2021) the scores of gold sentences are higher than those of generated texts, but our gold spans have a significantly lower relevance scores than the generated spans. This is likely because the gold spans contain information derived from the body sections of the cited papers, which are not provided to either the models or to the human judges. As a result, some gold spans appear to be irrelevant to the human judges, echoing our earlier finding in §5.1 that citation spans contain more focused information. This observation also suggests that gold citation spans are not necessarily the best target for all task settings.

We also see that, while *dominant* sentences and spans receive similar scores, the *reference* sentences have lower relevance scores than the spans. This result makes sense because *reference* citation spans are short and focused, so the full sentences include tokens unrelated to the cited paper(s). Overall, the generated spans are rated slightly higher than the generated sentences by the human judges, confirming that span-level citation text generation is preferable to sentence-level generation.

6 Toward Full Related Work Generation

Existing extractive related work generation systems (Hoang and Kan, 2010; Hu and Wan, 2014; Chen and Zhuge, 2019; Wang et al., 2019) select sentences from the target paper and/or the cited papers, which can be concatenated to form a full related work section; neural network-based, abstractive related work generation systems generate individual citation sentences (Xing et al., 2020; Ge et al., 2021; Luu et al., 2021) or paragraphs (Chen et al., 2021). However, none of these prior works address the ordering of the extracted/generated sentences or the grouping of sentences into paragraphs, nor are they able to produce rhetorical sentences to smooth the transitions between citations. No prior work bridges the gap from generating individual citation texts to generating a full related work section.

We suggest a bottom-up, iterative approach to generate full related work sections. The process would begin with generating citation spans under

ence type, as well as the type of each citation in the span.

the settings proposed in §5.2. Then, multiple generated citation spans would be aggregated and rewritten into citation text blocks in either the *summarization* or *narrative* style. These blocks would be further aggregated and rewritten into paragraphs by generating *transition* and *reflection* sentences.

Generating and rewriting in this pipeline fashion has the following benefits: (1) It mitigates the practical issue of computational resource limitations, given that state-of-the-art models do not perform well on long text generation. (2) The auxiliary inputs, such as citation functions or discourse tags, may vary for each stage of generation. (3) As a practical system to assist researchers, it is crucial to allow user involvement in the iterative generation process. Due to the large search space, consisting of multiple valid related work section candidates with different writing styles, it is extremely challenging to precisely generate a satisfying text with a one-shot, end-to-end system. A human-in-the-loop approach allows the user to significantly prune the search space and simultaneously reduces the error-propagation issue caused by the pipeline design.

7 Other Related Tasks

7.1 Scientific Document Understanding

Besides summarization, scientific document understanding also plays an important role in related work generation.

Citation Analysis. Citations are the core of related work sections. There has been a line of research on citation analysis, including citation function (Teufel et al., 2006; Dong and Schäfer, 2011; Jurgens et al., 2018; Tuarob et al., 2019), citation intent (Cohan et al., 2019; Lauscher et al., 2021), citation sentiment (Athar, 2011; Athar and Teufel, 2012; Ravi et al., 2018; Vyas et al., 2020), etc. These studies annotate citations with different labeling schemes to study the various usages and purposes of citations.

Discourse Analysis. Scientific discourse analysis studies the rhetorical components of clauses, sentences, or text spans that are not limited to citations, uncovering how authors persuade expert readers with their claims. There is a significant amount of prior work proposing discourse schemes and developing models for discourse tagging for scientific articles (Teufel and Moens, 1999, 2002; Hirohata et al., 2008; Liakata, 2010; Liakata et al., 2012; Guo et al., 2010; De Waard and Maat, 2012; Burns et al., 2016; Dernoncourt and Lee, 2017; Huang et al., 2020; Li et al., 2021a).

Our CORWA discourse tagging task focuses on distinguishing the source of the information in each related work sentence, which is complementary to the discourse tagging work listed above.

7.2 Cited Text Span

AbuRa'ed et al. (2020) extend Hoang and Kan (2010)'s RWSData dataset by annotating the Cited Text Span (CTS) (Wang et al., 2019). They annotate the specific sentences in cited papers that each citation in the target paper is based on. For each cited paper, they further collect a set of papers that co-cite this cited paper. Jaidka et al. (2018, 2019) propose the CL-Scisumm shared task, which includes identifying the CTS in reference papers for each citation instance. This shared task provides a valuable dataset for the precise generation of citation texts from a CTS, in contrast to most recent work, which uses the cited paper's abstract or introduction.

7.3 Studies of Literature Reviews

From an information studies perspective, Khoo et al. (2011) largely classify literature reviews into two styles: integrative and descriptive. Descriptive literature reviews summarize individual studies and provide detailed information on each, such as methods, results, and interpretation; integrative literature reviews provide fewer details of individual studies, instead focusing on synthesizing ideas and results extracted from these papers. Jaidka et al. (2010, 2011, 2013) analyze the properties of these two types of literature reviews.

8 Conclusion

We present the CORWA dataset of three interrelated annotation tasks: discourse tagging, citation span detection, and citation type recognition. We demonstrate the significance of CORWA with analyses from multiple perspectives, such as writing style and discourse patterns. We propose a strong baseline model that can automatically propagate the CORWA annotation scheme to massive unlabeled related work sections. Furthermore, we show that citation spans are a better alternative to citation sentences for both the relevant sentence retrieval and citation generation tasks. Finally, we discuss a novel framework for human-in-the-loop iterative abstractive related work generation.

References

Ahmed AbuRa'ed, Horacio Saggion, and Luis Chiruzzo. 2020. A multi-level annotated corpus of scientific papers for scientific document summarization and cross-document relation discovery. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6672–6679, Marseille, France. European Language Resources Association.

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

Awais Athar and Simone Teufel. 2012. Context-enhanced 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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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

Gully APC Burns, Pradeep Dasigi, Anita de Waard, and Eduard H Hovy. 2016. Automated detection of discourse segment and experimental types from the text of cancer pathway results sections. *Database*, 2016.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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

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. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Anita De Waard and Henk Pander Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. *CoRR*, abs/1710.06071.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107. Association for Computational Linguistics.

Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *proceedings of the 17th international conference on data engineering*, pages 215–224. Citeseer.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

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

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.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2018. Insights from cl-scisumm 2016: the faceted scientific document summarization shared task. *International Journal on Digital Libraries*, 19(2):163–171.

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.

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

Kokil Jaidka, Michihiro Yasunaga, Muthu Kumar Chandrasekaran, Dragomir Radev, and Min-Yen Kan. 2019. The cl-scisumm shared task 2018: Results and key insights. *arXiv preprint arXiv:1909.00764*.

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.

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.

Min-Yen Kan, Kathleen R. McKeown, and Judith L. Klavans. 2001. Applying natural language generation to indicative summarization. In *Proceedings of the ACL 2001 Eighth European Workshop on Natural Language Generation (EWNLG)*, Toulouse, France. Association for Computational Linguistics.

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

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*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chun Li and Jianyong Wang. 2008. Efficiently mining closed subsequences with gap constraints. In *proceedings of the 2008 SIAM International Conference on Data Mining*, pages 313–322. SIAM.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021a. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562.

Xiangci Li, Gully A Burns, and Nanyun Peng. 2021b. A paragraph-level multi-task learning model for scientific fact-verification. In *SDU@ AAAI*.

Xiangci Li and Jessica Ouyang. 2022. Automatic related work generation: A meta study. *arXiv preprint arXiv:2201.01880*.

Maria Liakata. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 1–4. Association for Computational Linguistics.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Simone Teufel and Marc Moens. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. *Towards Standards and Tools for Discourse Tagging*.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vishal Vyas, Kumar Ravi, Vadlamani Ravi, V Uma, Srirangaraj Setlur, and Venu Govindaraju. 2020. Article citation study: Context enhanced citation sentiment detection. *arXiv preprint arXiv:2005.04534*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

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.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462.

A Appendix

A.1 Training Configurations

For the joint related work tagger training, we use GeForce GTX 1080 11 GB GPUs. The training process lasts 2.5 hours on a single GPU using Huggingface's (Wolf et al., 2020) SciBERT, BERT-base or Roberta-base as the paragraph encoders, and it lasts 6.5 hours using LED-base encoder. We train the models for 15 epochs. It takes approximately one week to run the hyper-parameter search using five-fold cross-validation for all language models, using 8 GPUs in total.

For training the citation span generation model, we use Tesla V100s-PCIE-32GB GPUs. The training process lasts for 2 days on a single GPU. We run the training for a maximum of 3 epochs with early stopping based on the validation loss.

A.2 Ethical Considerations

We present a new dataset that is derived from the S2ORC dataset (Lo et al., 2020), which is released under CC BY-NC 2.0 license. The Huggingface models (Wolf et al., 2020) we develop upon are released under Apache License 2.0.

Our annotators were compensated for their work at a rate of double the minimum wage in our local area.

	Five-fold cross-validation scores			Test-set scores		
Models	**Disc**	**CT**	**CS**	**Disc**	**CT**	**CS**
SciBERT (Beltagy et al., 2019)	**0.900** (0.0099)	**0.961** (0.0038)	**0.926** (0.0059)	**0.898**	**0.959**	**0.930**
Roberta-base (Liu et al., 2019)	0.886 (0.0050)	0.956 (0.0036)	0.922 (0.0048)	0.885	0.956	0.929
BERT-base (Devlin et al., 2019)	0.879 (0.0070)	0.954 (0.0055)	0.910 (0.0064)	0.875	0.952	0.915
LED-base (Pretrained)	0.872 (0.0253)	0.948 (0.0117)	0.905 (0.0088)	0.869	0.910	0.907
LED-base (Beltagy et al., 2020)	0.865 (0.0090)	0.922 (0.0128)	0.907 (0.0074)	0.842	0.874	0.909

Table 6: Micro-F1 scores for the joint related work tagger using different language models as the encoder. The tasks are discourse tagging (Disc), citation type recognition (CT), and citation span detection (CS). Five-fold cross-validation scores are reported as the mean (standard deviation) across all folds. The pretraining of LED is explained in §5.2.1.

Discourse Subsequence
transition, *narrative_cite*, *single_summ*
Functionalities
Introducing an approach and providing background knowledge.
Examples
1. Joint POS tagging with parsing is not a new idea.
2. In PCFG-based parsing (Collins, 1999; Charniak, 2000; Petrov et al., 2006), POS tagging is considered as a natural step of parsing by employing lexical rules.
3. For transition-based parsing, Hatori et al. (2011) proposed to integrate POS tagging with dependency parsing.

Discourse Subsequence
single_summ, *reflection*
Functionalities
Comparing the prior work to the current work.
Examples
1. Haghighi et al. (2009) confirm and extend these results, showing BLEU improvement for a hierarchical phrase-based MT system on a small Chinese corpus.
2. As opposed to ITG, we use a linguistically motivated phrase-structure tree to drive our search and inform our model.

Discourse Subsequence
reflection, *single_summ*
Functionalities
Supporting the current work with a previous work.
Examples
1. Our baseline semi-supervised model can be viewed as an extension of these approaches to a reading comprehension setting.
2. Dai et al. (2015) also explore initialization from a language model, but find that the recurrent autoencoder is superior, which is why we do not consider language models in this work.

Discourse Subsequence
transition, *narrative_cite*, *transition*
Functionalities
Topic sentence, narration of prior work followed by critique.
Examples
1. Traditional work on relation classification can be categorized into feature-based methods and kernel-based methods.
2. The former relies on a large number of human-designed features (Zhou et al., 2005; Jiang and Zhai, 2007; Li and Ji, 2014) while the latter leverages various kernels to implicitly explore a much larger feature space (Bunescu and Mooney, 2005; Nguyen et al., 2009).
3. However, both methods suffer from error propagation problems and poor generalization abilities on unseen words.

Table 7: Frequent discourse label subsequences detected by applying PrefixSpan (Han et al., 2001) and Gap-Bide algorithm (Li and Wang, 2008).

Discourse Subsequence *single_summ*, *single_summ*, *transition* **Functionalities** Commenting previous works summarized. **Examples** 1. Walker et al. (2012) extract rules representing characters from their annotated movie subtitle corpora. 2. Miyazaki et al. (2015) propose a method of converting utterances using rewriting rules automatically derived from a Twitter corpus. 3. These approaches have a fundamental problem to need some manual annotations, which is a main issue to be solved in this work.
Discourse Subsequence *narrative_cite*, *transition*, *single_summ* **Functionalities** Criticizing the previously cited work and citing an improved work. **Examples** 1. There have also been several classical studies based on nonneural approaches to headline generation (Woodsend et al., 2010; Alfonseca et al., 2013; Colmenares et al., 2015) , but they basically addressed sentence compression after extracting important linguistic units such as phrases. 2. In other words, their methods can still yield erroneous output, although they would be more controllable than neural models. 3. One exception is the work of Alotaiby (2011) , where fixed-sized substrings were considered for headline generation.
Discourse Subsequence *narrative_cite*, *transition*, *single_summ* **Functionalities** Describing an idea following by a comment and then citations implementing the idea. **Examples** 1. One of the classes of errors in the Helping Our Own (HOO) 2011 shared task (Dale and Kilgarriff, 2011) was punctuation. 2. Comma errors are the most frequent kind of punctuation error made by learners. 3. Israel et al. (2012) present a model for detecting these kinds of errors in learner texts.

Table 8: Frequent discourse label subsequences detected by applying PrefixSpan (Han et al., 2001) and Gap-Bide algorithm (Li and Wang, 2008), continued.