

# Citance-Contextualized Summarization of Scientific Papers

Shahbaz Syed, Ahmad Dawar Hakimi, Khalid Al-Khatib, Martin Potthast



UNIVERSITÄT  
LEIPZIG



university of  
 groningen

# Motivation

- In scientific papers, a claim is typically supported by citing related work.

## 2.2 Citation-based Summarization

In citation-based summarization, citances from the source paper are used as queries to extract relevant content from the target paper, and to generate a summary. [Qazvinian and Radev \(2008\)](#) analyzed the citation network of target papers and collected citances from different sources. These citances were clustered, and the central sentences identi-

# Motivation

- ❑ In scientific papers, a claim is typically supported by citing related work.
- ❑ Readers may not know *how* a related work supports the claim.

## 2.2 Citation-based Summarization

In citation-based summarization, citances from the source paper are used as queries to extract relevant content from the target paper, and to generate a summary. Qazvinian and Radev (2008) analyzed the citation network of target papers and collected citances from different sources. These citances were clustered, and the central sentences identi-

# Motivation

- ❑ In scientific papers, a claim is typically supported by citing related work.
- ❑ Readers may not know *how* a related work supports the claim.
- ❑ Looking up a citation is a context switch which negatively affects the reading flow.

## 2.2 Citation-based Summarization

In citation-based summarization, citances from the source paper are used as queries to extract relevant content from the target paper, and to generate a summary. Qazvinian and Radev (2008) analyzed the citation network of target papers and collected citances from different sources. These citances were clustered, and the central sentences identi-

# Motivation

- ❑ In scientific papers, a claim is typically supported by citing related work.
- ❑ Readers may not know *how* a related work supports the claim.
- ❑ Looking up a citation is a context switch which negatively affects the reading flow.
- ❑ Generating an informative summary of cited related work would help.

## 2.2 Citation-based Summarization

In citation-based summarization, citances from the source paper are used as queries to extract relevant content from the target paper, and to generate a summary. Qazvinian and Radev (2008) analyzed the citation citances from the source paper and used them to generate a summary of the target paper.

Qazvinian and Radev (2008)

Scientific paper summarization using citation summary networks

V Qazvinian, DR Radev - arXiv preprint arXiv:0807.1560, 2008

Quickly moving to a new area of research is painful for researchers due to the vast amount of scientific literature in each field of study. One possible way to overcome this problem is to summarize a scientific topic. In this paper, we propose a model of summarizing a single article, which can be further used to summarize an entire topic. Our model is based on analyzing others' viewpoint of the target article's contributions and the study of its citation summary network using a clustering approach.

☆ Save 📄 Cite Cited by 355 Related articles All 12 versions

[PDF] arxiv.org

See in References

# Related Work: Types of Scientific Paper Summaries

## Citance-Contextualized Summarization of Scientific Papers

### **TL:DR**

SciTLDR (Cachola et al., 2020)

This work proposes a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence containing the citation of a reference (a so-called "citance") that outlines the content of the cited paper relevant to the citation location.

# Related Work: Types of Scientific Paper Summaries

## Citance-Contextualized Summarization of Scientific Papers

### **TL:DR**

SciTLDR (Cachola et al., 2020)

This work proposes a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence containing the citation of a reference (a so-called "citance") that outlines the content of the cited paper relevant to the citation location.

### **Abstract**

(Author-provided)

Current approaches to automatic summarization of scientific papers generate informative summaries in the form of abstracts. However, abstracts are not intended to show the relationship between a paper and the references cited in it. We propose a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence ....

# Related Work: Types of Scientific Paper Summaries

## Citance-Contextualized Summarization of Scientific Papers

### TL:DR

SciTLDR (Cachola et al., 2020)

This work proposes a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence containing the citation of a reference (a so-called "citance") that outlines the content of the cited paper relevant to the citation location.

### Abstract

(Author-provided)

Current approaches to automatic summarization of scientific papers generate informative summaries in the form of abstracts. However, abstracts are not intended to show the relationship between a paper and the references cited in it. We propose a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence ....

### Rhetorical Role Highlights

SCIM (Fok et al., 2022)

Result

Method

(2019) uses a loss based on predicting whether the second segment in a pair has been swapped with a segment from another document. We compare to this loss in our experiments and find that sentence ordering is a more challenging pretraining task and more useful for certain downstream tasks. Concurrently to our work, Wang et al. (2019) also try to predict the order of two consecutive segments of text, but they combine it with the original next sentence prediction in a three-way classification task rather than empirically comparing the two.



# Related Work: Types of Scientific Paper Summaries

## Citance-Contextualized Summarization of Scientific Papers

### TL:DR

SciTLDR (Cachola et al., 2020)

This work proposes a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence containing the citation of a reference (a so-called "citance") that outlines the content of the cited paper relevant to the citation location.

### Abstract

(Author-provided)

Current approaches to automatic summarization of scientific papers generate informative summaries in the form of abstracts. However, abstracts are not intended to show the relationship between a paper and the references cited in it. We propose a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence ....

### Rhetorical Role Highlights

SCIM (Fok et al., 2022)

Result

Method

2019) uses a loss based on predicting whether the second segment in a pair has been swapped with a segment from another document. We compare to this loss in our experiments and find that sentence ordering is a more challenging pretraining task and more useful for certain downstream tasks. Concurrently to our work, Wang et al. (2019) also try to predict the order of two consecutive segments of text, but they combine it with the original next sentence prediction in a three-way classification task rather than empirically comparing the two.

## Unsupervised Cross-Domain Image Generation

**Paper Abstract:** We study the problem of **transferring a sample in one domain to an analog sample in another domain**. Given two related domains,  $S$  and  $T$ , we would like to learn a generative function  $G$  that maps an input sample from  $S$  to the domain  $T$ , such that the output of a given function  $f$ , which accepts inputs in either domains, would remain unchanged. Other than the function  $f$ , the training data is unsupervised and consist of a set of samples from each domain. **The Domain Trans-**

**Citation Text:** Taigman et al. [8] proposed **the Domain Transfer Network (DTN) to map a sample from one domain to an analog sample in another domain and**

### Citation Span

CiteSum (Mao et al., 2022)

# Related Work: Types of Scientific Paper Summaries

## Citance-Contextualized Summarization of Scientific Papers

### TL:DR

SciTLDR (Cachola et al., 2020)

This work proposes a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence containing the citation of a reference (a so-called "citance") that outlines the content of the cited paper relevant to the citation location.

### Abstract

(Author-provided)

Current approaches to automatic summarization of scientific papers generate informative summaries in the form of abstracts. However, abstracts are not intended to show the relationship between a paper and the references cited in it. We propose a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence ....

### Rhetorical Role Highlights

SCIM (Fok et al., 2022)

Result

2019) uses a loss based on predicting whether the second segment in a pair has been swapped with a segment from another document. We compare to this loss in our experiments and find that sentence ordering is a more challenging pretraining task and more useful for certain downstream tasks. Concurrently to our work, Wang et al. (2019) also try to predict the order of two consecutive segments of text, but they combine it with the original next sentence prediction in a three-way classification task rather than empirically comparing the two.

Method

## Unsupervised Cross-Domain Image Generation

**Paper Abstract:** We study the problem of **transferring a sample in one domain to an analog sample in another domain**. Given two related domains, S and T, we would like to learn a generative function G that maps an input sample from S to the domain T, such that the output of a given function f, which accepts inputs in either domains, would remain unchanged. Other than the function f, the training data is unsupervised and consist of a set of samples from each domain. **The Domain Trans-**

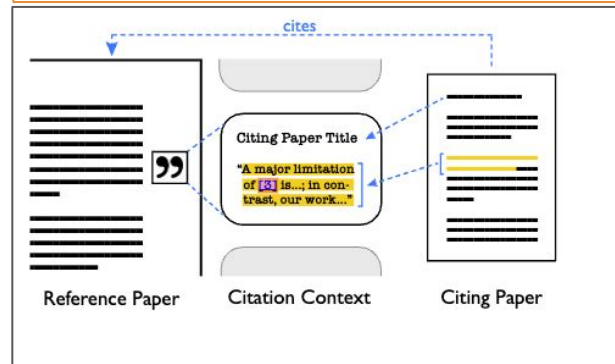
**Citation Text:** Taigman et al. [8] proposed **the Domain Transfer Network (DTN) to map a sample from one domain to an analog sample in another domain and**

### Citation Span

CiteSum (Mao et al., 2022)

## Citation Contexts from Subsequent Papers

CiteRead (Rachatsumrit et al. 2022)



# Related Work: Limitations of Existing Summary Types

- ❑ TL;DR summaries mention only the key contribution.
- ❑ Abstracts are incomplete and/or biased towards specific contents.
- ❑ Highlighting by rhetorical role does not eliminate referencing the cited paper.
- ❑ Citation spans may be misinterpreted by the citing authors.
- ❑ Aggregating citation sentences from subsequent papers results in an extractive summary which may be incoherent.

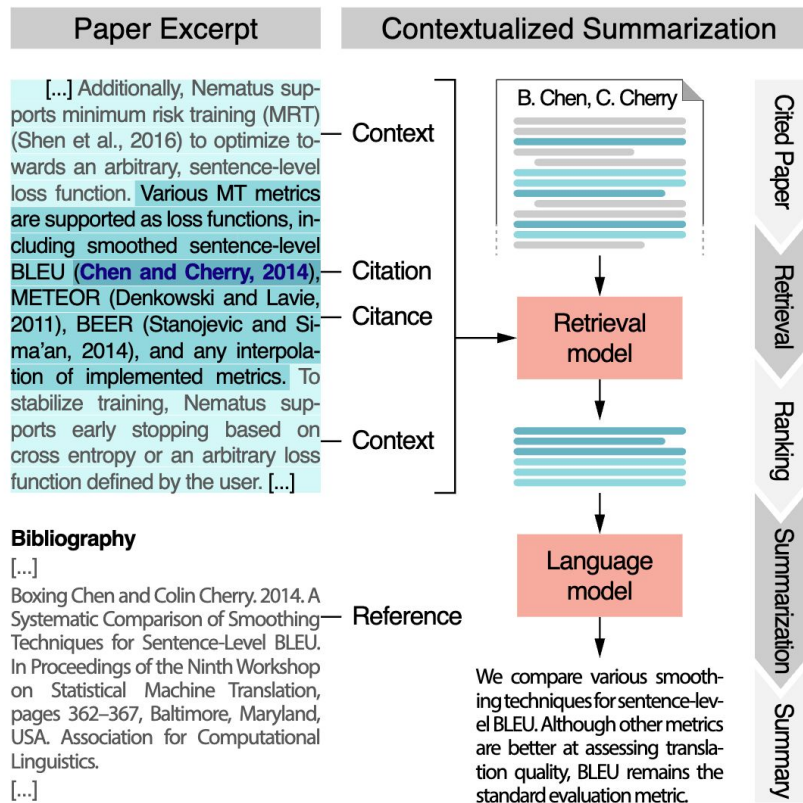
# Related Work: Limitations of Existing Summary Types

- ❑ **TL;DR** summaries mention only the key contribution.
  - ❑ **Abstracts** are incomplete and/or biased towards specific contents.
  - ❑ **Highlighting** by rhetorical role does not eliminate referencing the cited paper.
  - ❑ Citation spans may be misinterpreted by the citing authors.
  - ❑ **Aggregating** citation sentences from subsequent papers results in an extractive summary which may be incoherent.
- Independent of the citation context as papers are often cited multiple times; *different* contexts have the *same* summary.

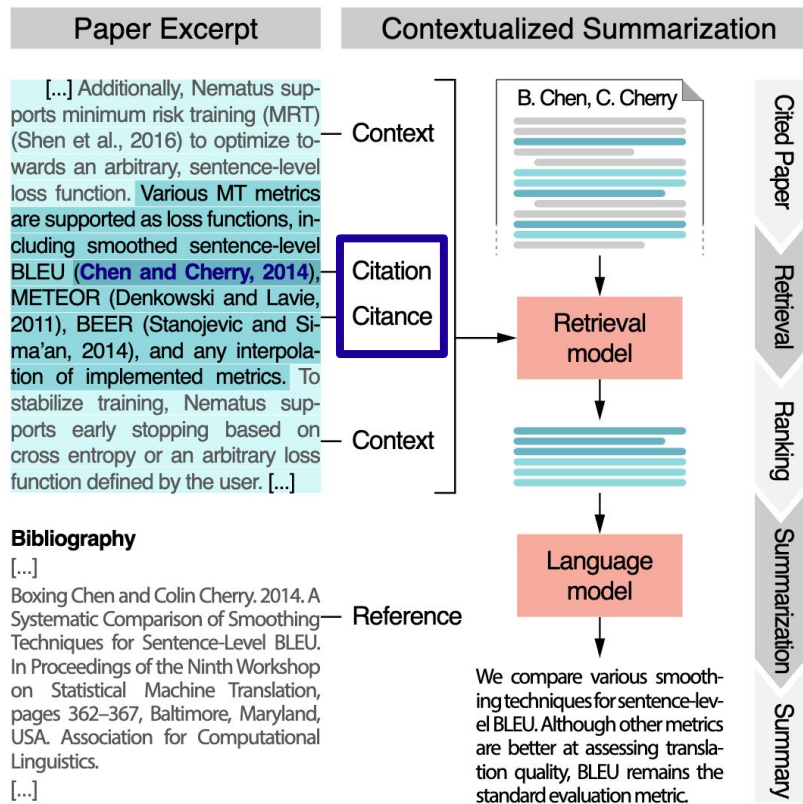
# Citance-Contextualized Summarization

- ❑ Is a form of aspect-oriented summarization.
- ❑ Respects the current citation context of a paper.
- ❑ Enables multiple summaries for the same paper if cited in different contexts.
- ❑ Provides a better reading experience.

# Citance-Contextualized Summarization

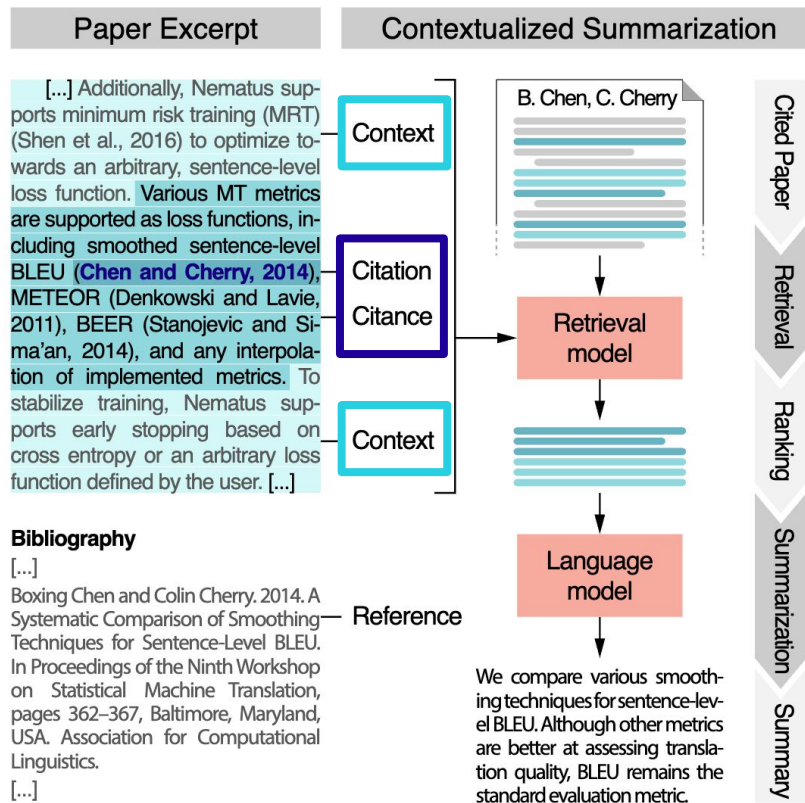


# Citance-Contextualized Summarization



1. Given a **Citation**, mark the constituent citation sentence as the **Citance**.

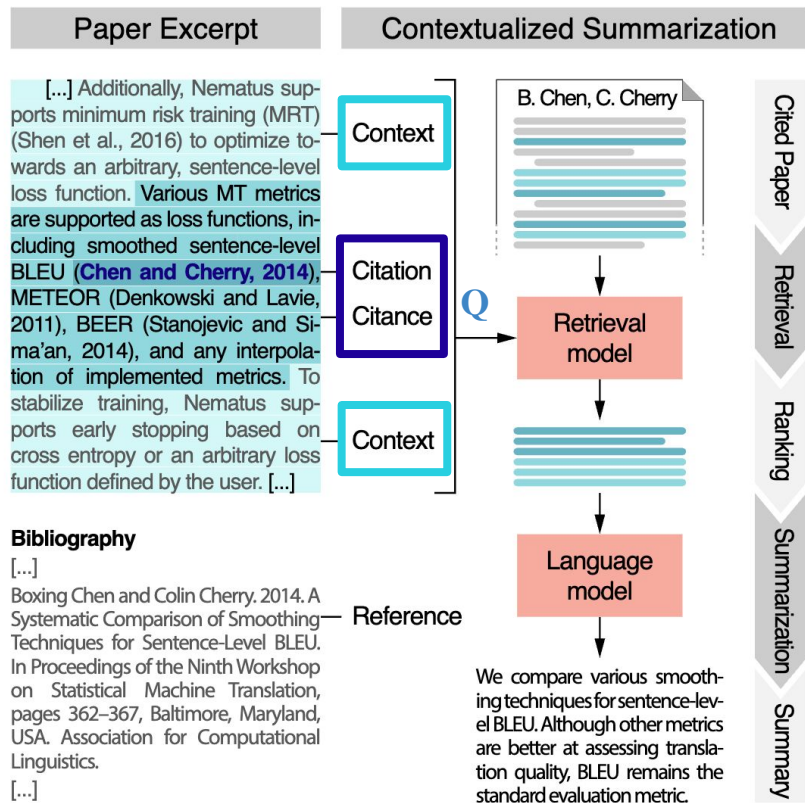
# Citance-Contextualized Summarization



1. Given a **Citation**, mark the constituent citation sentence as the **Citance**.
2. Compose a **Citance Context** e.g., a fixed window of surrounding sentences.

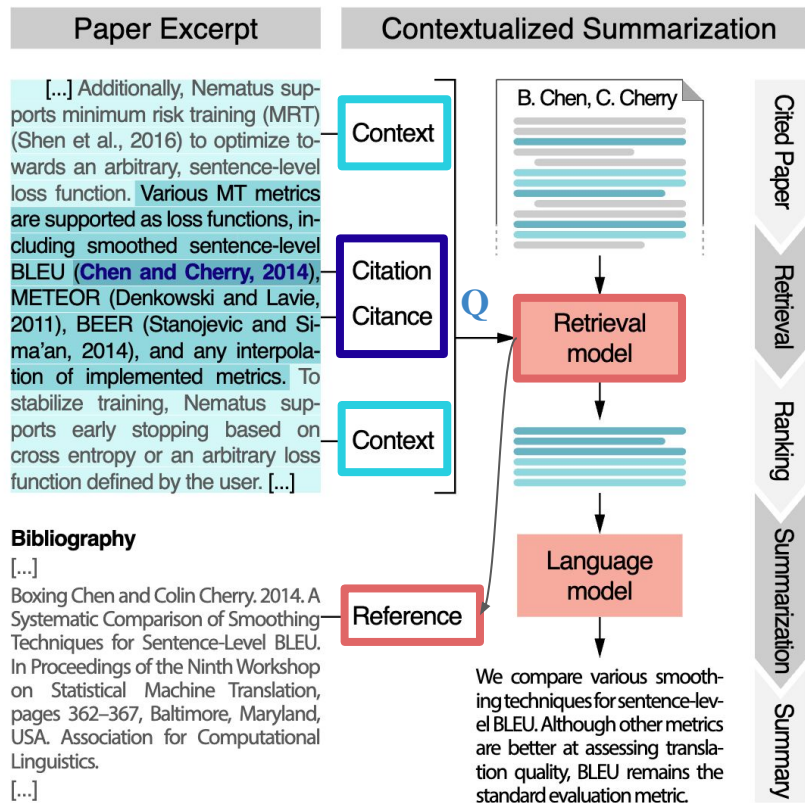


# Citance-Contextualized Summarization



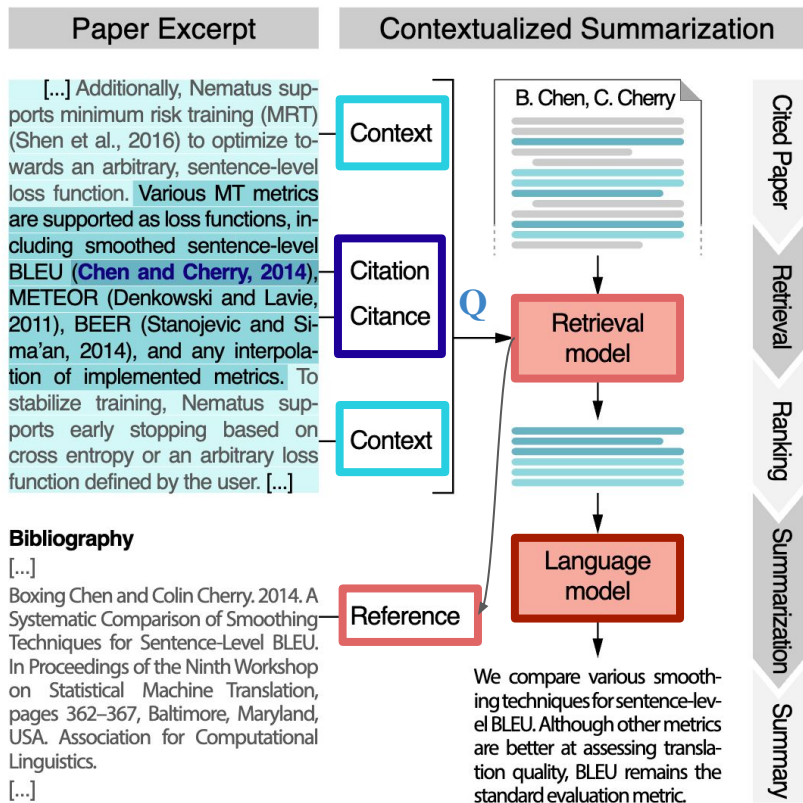
1. Given a **Citation**, mark the constituent citation sentence as the **Citance**.
2. Compose a **Citance Context** e.g., a fixed window of surrounding sentences.
3. Construct one or more **queries** from the context.

# Citance-Contextualized Summarization



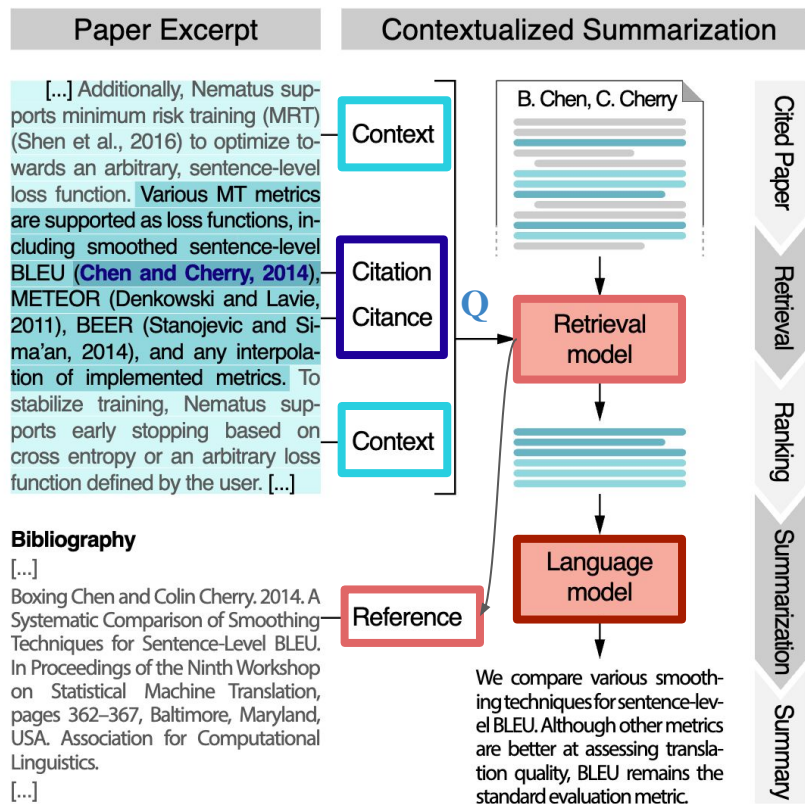
1. Given a **Citation**, mark the constituent citation sentence as the **Citance**.
2. Compose a **Citance Context** e.g., a fixed window of surrounding sentences.
3. Construct one or more **queries** from the context.
4. **Retrieve** relevant contents from the cited paper.

# Citance-Contextualized Summarization



1. Given a **Citation**, mark the constituent citation sentence as the **Citance**.
2. Compose a **Citance Context** e.g., a fixed window of surrounding sentences.
3. Construct one or more **queries** from the context.
4. **Retrieve** relevant contents from the cited paper.
5. **Summarize** top-k contents to create an abstractive, citance-contextualized summary.

# Citance-Contextualized Summarization



1. Given a **Citation**, mark the constituent citation sentence as the **Citance**.
2. Compose a **Citance Context** e.g., a fixed window of surrounding sentences.
3. Construct one or more **queries** from the context.
4. **Retrieve** relevant contents from the cited paper.
5. **Summarize** top-k contents to create an abstractive, citance-contextualized summary.

Completely unsupervised & modular pipeline

# Context Formulation and Content Retrieval

- ❑ Given a citation in the source paper we compose three types of context:
  - ❑ **citance**: the citation sentence itself
  - ❑ **neighbors**: citance and its immediate neighbors
  - ❑ **similar**: citance and the top-2 semantically similar sentences
- ❑ These contexts are used as queries to fetch relevant content from the cited paper.
- ❑ Additionally, we also experimented with only keywords of the contexts as queries.

# Context Formulation and Content Retrieval

- ❑ Given a citation in the source paper we compose three types of context:
  - ❑ **citance**: the citation sentence itself
  - ❑ **neighbors**: citance and its immediate neighbors
  - ❑ **similar**: citance and the top-2 semantically similar sentences
- ❑ These contexts are used as queries to fetch relevant content from the cited paper.
- ❑ Additionally, we also experimented with only keywords of the contexts as queries.
- ❑ Two types of content are retrieved from the target paper:
  - ❑ Top-5 sentences (for higher coverage)
  - ❑ Top-2 paragraphs (for higher focus)
- ❑ We employed BM25 (shallow retrieval) and SciBERT (dense retrieval) for retrieval.  
Index Size: 151M sentences, 40M paragraphs

# Abstractive Summarization

- ❑ We used instruction-tuned LLMs for summarizing the retrieved content.  
Alpaca 7b, Vicuna 13b, LLaMA-CoT, Falcon Instruct
- ❑ We **paraphrased** the top-5 sentences and **summarised** the top-2 paragraphs.
- ❑ Summaries are no longer than five sentences.

## Paraphrasing Prompt

### ### Instruction:

A chat between a curious human and an artificial intelligence assistant. The assistant knows how to paraphrase scientific text and the user will provide the scientific text for the assistant to paraphrase.

### ### Input:

Generate a coherent paraphrased text for the following scientific text: *{input}*.

### ### Output:

## Summarization Prompt

### ### Instruction:

A chat between a curious human and an artificial intelligence assistant. The assistant knows how to summarize scientific text and the user will provide the scientific text for the assistant to summarize.

### ### Input:

Generate a coherent summary for the following scientific text in not more than 5 sentences: *{input}*.

### ### Output:

# Dataset: Webis-Context-SciSumm-2023

A large scale dataset for contextualized summarization of scientific papers

- ❑ Derived from the S2ORC corpus (Lo et al., 2019) of 12M papers with full text.
- ❑ We focused on the subset of 870K computer science papers.
- ❑ 540K papers contain citations to other papers.  
(346K papers have multiple citations to the same target paper)
- ❑ Extracted 4.6M citation sentences from all sections of the papers.
- ❑ Given a citance context, we retrieved the top 5 sentences and top 2 paragraphs from the cited paper.  
<citation context, top-5 sentences, top-2 paragraphs>
- ❑ Summarized the retrieved content using 4 prompt-based LLMs.  
(Due to resource constraints the released dataset has summaries for only 25K examples from the best models as per our manual evaluation)



# Evaluation: Content Retrieval

- ❑ We have **12 retrieval scenarios** in total.  
(3 contexts + 3 contexts-to-keywords) X 2 IR models
- ❑ Based on 10 random citances we retrieved the top-5 sentences per scenario: 600 sentences to be judged for relevance.
- ❑ One of the authors assessed relevance on a 3-point scale: “relevant”, “somewhat relevant”, and “non-relevant”

# Evaluation: Content Retrieval

- ❑ We have **12 retrieval scenarios** in total.  
(3 contexts + 3 contexts-to-keywords) X 2 IR models
- ❑ Based on 10 random citances we retrieved the top-5 sentences per scenario: 600 sentences to be judged for relevance.
- ❑ One of the authors assessed relevance on a 3-point scale: “relevant”, “somewhat relevant”, and “non-relevant”
- ❑ We computed NDCG@5 to select the best context type for shallow and dense retrieval.
- ❑ <similar context, BM25> and <citance, SciBERT> were chosen for the summarization evaluation.

BM25 (Shallow)		SciBERT (Dense)	
Query	Mean nDCG@5	Query	Mean nDCG@5
citance	0.943	citance	<b>0.943</b>
similar	<b>0.958</b>	similar	0.918
neighbors	0.898	neighbors	0.801
<hr/>		<hr/>	
citance-keywords	0.914	citance-keywords	0.617
similar-keywords	0.944	similar-keywords	0.650
neighbors-keywords	0.928	neighbors-keywords	0.706

Table 3: Evaluation of 12 retrieval setups as combinations of a shallow and a dense retrieval model with citance contexts as queries to extract relevant content from cited papers. We report mean nDCG@5 for 600 relevance judgments. The best combination (in bold) has been selected for the summarization step.

# Evaluation: Summarization

- ❑ Randomly sampled 15 papers from ACL anthology (2016-2020)
- ❑ 363 citances in total from which 25 citances were randomly sampled.
- ❑ Retrieved **100 documents** in total based on the best `<context, IR model>` from **previous evaluation** (a document is either the top-5 sentences or top-2 paragraphs)
  - ❑ 25 citances  $\rightarrow$  `<similar, BM25>`  $\rightarrow$  top-5 sentences, top-2 paragraphs
  - ❑ 25 citances  $\rightarrow$  `<citance, SciBERT>`  $\rightarrow$  top-5 sentences, top-2 paragraphs
- ❑ These 100 documents were paraphrased/summarized using GPT4 (zero shot).
- ❑ Each summary was manually checked for faithfulness and corrected if necessary. This forms our **ground truth** for automatic and manual evaluation.

# Evaluation: Summarization (quantitative)

- ❑ Each of the 100 retrieved documents was paraphrased/summarized using four models: Alpaca 7b, Vicuna 13b, LLaMA-CoT, Falcon Instruct.
- ❑ We used ROUGE and BERTScore for automatic evaluation.

Model	BERTScore	ROUGE		
		R-1	R-2	R-L
<b>top-2 paragraphs</b>				
<i>similar-BM25</i>				
Alpaca	0.343	47.3	25.5	44.9
Falcon	0.401	48.2	27.1	45.0
LLaMA-CoT	0.448	53.0	31.9	50.5
Vicuna	0.465	<b>58.7</b>	<b>35.4</b>	<b>55.8</b>
<i>citance-SciBERT</i>				
Alpaca	0.390	54.3	32.2	52.0
Falcon	0.413	52.1	29.6	48.9
LLaMA-CoT	<b>0.497</b>	54.7	32.9	52.5
Vicuna	0.431	56.7	34.2	53.9
<b>top-5 sentences</b>				
<i>similar-BM25</i>				
Alpaca	0.616	56.2	35.4	54.8
Falcon	0.649	57.5	35.6	55.2
LLaMA-CoT	0.707	61.2	38.6	60.0
Vicuna	0.551	57.2	34.3	54.9
<i>citance-SciBERT</i>				
Alpaca	0.595	56.6	34.7	55.1
Falcon	0.656	56.8	36.2	55.3
LLaMA-CoT	<b>0.748</b>	<b>62.9</b>	<b>40.6</b>	<b>60.9</b>
Vicuna	0.607	58.8	36.0	56.6

# Evaluation: Summarization (quantitative)

- Each of the 100 retrieved documents was paraphrased/summarized using four models: Alpaca 7b, Vicuna 13b, LLaMA-CoT, Falcon Instruct.
- We used ROUGE and BERTScore for automatic evaluation.
- We chose the best setting from each summary type for manual evaluation:
  - Vicuna performed best for summarizing top-2 paragraphs
  - LLaMA-CoT performed best for paraphrasing top-5 sentences

Model	BERTScore	ROUGE		
		R-1	R-2	R-L
<b>top-2 paragraphs</b>				
<i>similar-BM25</i>				
Alpaca	0.343	47.3	25.5	44.9
Falcon	0.401	48.2	27.1	45.0
LLaMA-CoT	0.448	53.0	31.9	50.5
Vicuna	0.465	<b>58.7</b>	<b>35.4</b>	<b>55.8</b>
<i>citance-SciBERT</i>				
Alpaca	0.390	54.3	32.2	52.0
Falcon	0.413	52.1	29.6	48.9
LLaMA-CoT	<b>0.497</b>	54.7	32.9	52.5
Vicuna	0.431	56.7	34.2	53.9
<b>top-5 sentences</b>				
<i>similar-BM25</i>				
Alpaca	0.616	56.2	35.4	54.8
Falcon	0.649	57.5	35.6	55.2
LLaMA-CoT	0.707	61.2	38.6	60.0
Vicuna	0.551	57.2	34.3	54.9
<i>citance-SciBERT</i>				
Alpaca	0.595	56.6	34.7	55.1
Falcon	0.656	56.8	36.2	55.3
LLaMA-CoT	<b>0.748</b>	<b>62.9</b>	<b>40.6</b>	<b>60.9</b>
Vicuna	0.607	58.8	36.0	56.6

# Evaluation: Summarization (qualitative)

- ❑ For human evaluation, we asked 3 domain experts to assess the “usefulness” of the contextualized summaries according to their *coverage* and *focus*.
- ❑ We employed a 5-point Likert-scale (worst to best).
- ❑ **Coverage:** “how well the summary captures essential information from the cited paper that is relevant to the current citance?”
- ❑ **Focus:** “how coherent is the summary?”
- ❑ A total of **125 summaries** were manually evaluated.  
25 citances, 5 summaries (Vicuna top-2 paragraphs, LLaMA-CoT top-5 sentences, Abstract, GPT4 top-2 paragraphs, GPT4 top-5 sentences)

# Evaluation: Summarization (qualitative)

- ❑ Abstract had the highest coverage and focus.
- ❑ Vicuna had slightly better coverage than GPT4, while the latter had better focus.
- ❑ Summarizing top-2 relevant paragraphs from the cited paper was better than paraphrasing the top-5 relevant sentences (for both coverage and focus).

<b>Summary</b>	<b>Human Eval.</b>	
	Cov.	Focus
Abstract	<b>3.67</b>	<b>4.50</b>
<i>similar-BM25, top-2 paragraphs</i>		
GPT4 (Reference)	2.92	3.83
Vicuna	3.01	3.56
<i>citance-SciBERT, top-5 sentences</i>		
GPT4 (Reference)	2.45	2.99
LLaMA-CoT	2.33	2.33

# Evaluation: Summarization (qualitative)

- ❑ Abstract had the highest coverage and focus.
- ❑ Vicuna had slightly better coverage than GPT4, while the latter had better focus.
- ❑ Summarizing top-2 relevant paragraphs from the cited paper was better than paraphrasing the top-5 relevant sentences (for both coverage and focus).
- ❑ Results are mostly consistent with those from G-Eval which employs LLMs with chain of thought reasoning to automatically score the relevance, coverage, and focus.

Summary	Human Eval.		G-Eval		
	Cov.	Focus	Cov.	Focus	Rel.
Abstract	<b>3.67</b>	<b>4.50</b>	<b>3.12</b>	<b>3.80</b>	<b>3.23</b>
<i>similar-BM25, top-2 paragraphs</i>					
GPT4 (Reference)	2.92	3.83	2.60	3.46	3.10
Vicuna	3.01	3.56	2.96	3.20	2.80
<i>citance-SciBERT, top-5 sentences</i>					
GPT4 (Reference)	2.45	2.99	2.35	3.14	2.56
LLaMA-CoT	2.33	2.33	2.40	3.12	2.63



# Post-evaluation Survey

- ❑ While the abstract performed best, we discussed the possible strengths of our summaries with the annotators as a post evaluation survey.
- ❑ All annotators mentioned that in situations where the citance was ambiguous or not self-contained, contextualized summaries were better than abstracts.

# Post-evaluation Survey

- ❑ While the abstract performed best, we discussed the possible strengths of our summaries with the annotators as a post evaluation survey.
- ❑ All annotators mentioned that in situations where the citance was ambiguous or not self-contained, contextualized summaries were better than abstracts.

**Note for Future Work:** The sampled citances for evaluation should have been annotated for ambiguity/self-containment and the results grouped accordingly.

# Conclusion

- ❑ We presented a modular unsupervised approach to contextualized summarization of scientific papers.
- ❑ We created Webis Context SciSumm 2023, a large-scale dataset suited for citance-contextualized summarization.
- ❑ Evaluation shows that our summaries are better for understanding ambiguous citation contexts compared to the abstract.

# Conclusion

- ❑ We presented a modular unsupervised approach to contextualized summarization of scientific papers.
- ❑ We created Webis Context SciSumm 2023, a large-scale dataset suited for citance-contextualized summarization.
- ❑ Evaluation shows that our summaries are better for understanding ambiguous citation contexts compared to the abstract.
- ❑ **Future work:** fine-grained analysis of citations and their contexts, leveraging argumentation zones for targeted retrieval, evaluation on a larger scale with more participants, developing better reading interfaces augmented by contextualized summarization.

# Conclusion

- ❑ We presented a modular unsupervised approach to contextualized summarization of scientific papers.
- ❑ We created Webis Context SciSumm 2023, a large-scale dataset suited for citance-contextualized summarization.
- ❑ Evaluation shows that our summaries are better for understanding ambiguous citation contexts compared to the abstract.
- ❑ **Future work:** fine-grained analysis of citations and their contexts, leveraging argumentation zones for targeted retrieval, evaluation on a larger scale with more participants, developing better reading interfaces augmented by contextualized summarization.

**Thanks for listening!**