Comparative Graph-based Summarization of Scientific Papers Guided by Comparative Citations

Jingqiang Chen[1,†,*], **Chaoxiang Cai**[1,†], **Xiaorui Jiang**[2], **Kejia Chen**[1]
[1]School of Computer Science, Nanjing University of Posts and Telecommunications, China
[2]Centre for Computational Science and Mathematical Modelling, Coventry University, UK
{cjq,1220044905,chenkj}@njupt.edu.cn,
xiaorui.jiang@coventry.ac.uk

Abstract

With the rapid growth of scientific papers, understanding the changes and trends in a research area is rather time-consuming. The first challenge is to find related and comparable articles for the research. Comparative citations compare co-cited papers in a citation sentence and can serve as good guidance for researchers to track a research area. We thus go through comparative citations to find comparable objects and build a comparative scientific summarization corpus (CSSC). And then, we propose the comparative graph-based summarization (CGSUM) method to create comparative summaries using citations as guidance. The comparative graph is constructed using sentences as nodes and three different relationships of sentences as edges. The relationship that sentences occur in the same paper is used to calculate the salience of sentences, the relationship that sentences occur in two different papers is used to calculate the difference between sentences, and the relationship that sentences are related to citations is used to calculate the commonality of sentences. Experiments show that CGSUM outperforms comparative baselines on CSSC and performs well on DUC2006 and DUC2007.

1 Introduction

Today, the transient and rapidly evolving research areas and the numerous published research articles require researchers to orient themselves and discover the changes of the research area (Marrone, 2020). In order to reduce the burden of researchers, a solution is to find and compare related articles in the research area, and automatically create comparative summaries showing commonalities and differences of the articles where differences mean changes. The first problem is how to find related and comparable articles, and the second problem is how to create comparative summaries.

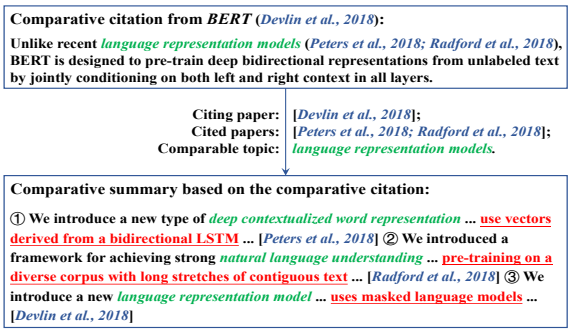


Figure 1: An example from our dataset. The comparative citation appears in the citing paper (Devlin et al., 2018) and cites two cited papers (Peters et al., 2018; Radford et al., 2018). The comparative summary is generated by summarizing the *commonality* (italicized) and difference (underlined) between three papers regarding to the comparable topic mentioned in the citation.

Fortunately, comparative citations can serve as good guidance in finding related and comparable articles and common topics. According to previous works (Teufel et al., 2006a; Hernandez-Alvarez et al., 2017), the function of comparative citations is to include shared topics between papers in the same field and reflects the comparative intent, i.e., the author intends to compare his own work with cited works. Citation function has been widely investigated. Teufel et al. (2006b) analyzed citation functions based on empirical works, and classified functions as *Contrast*, *Neutral*, *Weakness*, etc. Among all citation functions, the comparative citation is most suitable for comparative summarization as it contains most comparative information.

Given a set of comparable articles guided by comparative citations, we aim to summarize commonalities and differences between the articles and related to the comparable topics mentioned in citations. As the Figure 1 shows, the comparative citation in the upper part is captured by "unlike", where the citing and the cited papers share the comparable topic. Also, the bottom part shows a comparative summary based on mentioned comparative citation.

† The authors have contributed equally to this work.
∗ Jingqiang Chen is the corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5978–5988
October 12–17, 2022.

The summary accommodates the commonality and difference between the citing and the cited papers.

Our task is different from traditional survey generation and related work generation (Chen and Zhuge, 2016, 2019; Wang et al., 2020; Chen et al., 2021; Yuan et al., 2021) in that 1) our task utilizes the citing and the cited papers guided by comparative citations, and 2) related work and surveys focus on shared information while comparative summaries capture commonalities and differences.

We build a comparative scientific summarization corpus (CSSC) based on comparative citations. Three annotators are asked to annotate and collect comparative citations in 32 papers in the AI area using the citation function annotation scheme. We get 40 comparative citations with the corresponding citing and cited papers. For each comparative citation, annotators read through papers to generate a draft comparative summary for the comparative topic mentioned in the citation. After that, five postgraduates students specializing in works on selected 32 papers read and revise the draft comparison summary to create the ground truth.

Since our dataset is small-scaled, we propose a simple yet effective unsupervised comparative graph-based scientific summarization method (CGSUM). A comparative graph is built to represent the citation texts and papers. Each paper or citation text corresponds to a subgraph, where nodes represent sentences and weights of edges denote similarities between sentences. The salience of a sentence is computed by considering the position of the sentence within the paper. The commonality of a sentence is computed on its subgraph and the citation texts subgraph. The difference of a sentence is captured by adding negative edges between nodes from different paper subgraphs. Finally, salience, commonality and difference are linearly combined to rank and select sentences. Experiments show that CGSUM outperforms baselines on CSSC and also performs well on DUC2006 and DUC2007.

Our contributions are summarized as follows:

- We propose the task of comparative citation-guided summarization of scientific papers.
- We construct the comparative summarization dataset CSSC for scientific papers.
- We propose the comparative graph-based summarization method that considers three relationships between sentences. Experiments show the efficacy of the proposed model.

2 Related Work

Citations throughout scientific papers help understand the frontiers and trends in diverse research fields. Teufel et al. (2006b) analyzed citation functions based on empirical works, which is similar to (Su et al., 2019). Whereas Dong and Schäfer (2011); Abu-Jbara et al. (2013); Hernandez-Alvarez et al. (2017); Su et al. (2019) focused on the dimensions of organic and perfunctory as well as intentions and sentiments respectively.

Generic scientific summarization uses extractive(Yang et al., 2016; An et al., 2021; Dong et al., 2021), abstractive(See et al., 2017; Cachola et al., 2020; Dangovski et al., 2021) and other (Teufel and Moens, 2002; Cohan et al., 2018; Sharma et al., 2019) methods to summarize a document. Citation generation has also been studied (Xing et al., 2020; Ge et al., 2021). Early studies were based on keywords (Hoang and Kan, 2010; Chen and Zhuge, 2016). Xing et al. (2020) considered the abstract of cited papers to generate citations. Citation generation concerns semantics, while comparative summarization concerns citation function. The citation text is often too short to describe in detail.

Related work generation and survey generation generates from multiple documents. He et al. (2016) captured hot topics in fields. Chen and Zhuge (2019); AbuRa'ed et al. (2020); AbuRa'ed and Saggion (2021) took citations into account to mine information. What's more, Wang et al. (2020); Yuan et al. (2021) generated reviews that cover more aspects. Chen et al. (2021) took abstractive method. Related work generation places emphasis on shared content in cited papers and summarizes the common information. However, it is different from our task of comparative summarization, which is guided by comparative citations and summarizes commonalities and differences.

3 Task Definition

Given the comparative citation (*Cit*), the citing paper (*CP*) where the *Cit* appears, and a set of reference papers (*RPs*) that the *Cit* cites, the task aims to create the comparative summary containing commonalities and differences between *CP* and *RPs* with regard to the comparable topic mentioned in the *Cit*. Taking the case in Figure 1, *Cit* refers to the comparative citation, *CP* refers to the (Devlin et al., 2018) and *RPs* refers to (Peters et al., 2018; Radford et al., 2018). The comparable topic mentioned in the *Cit* is *language representation models*.

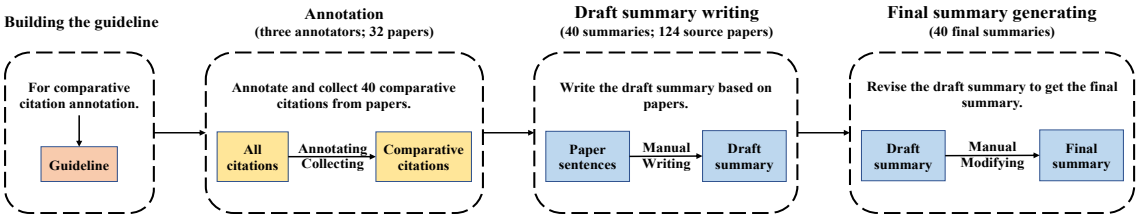


Figure 2: Overview of the dataset construction process.

Since there are no existing datasets and methods for the task, we build a comparative citation-guided dataset and propose a comparative graph-based method. To create the dataset, we annotate the function of citations of 32 papers in the AI area, and manually create comparative summaries for 40 annotated comparative citations. Then we create the comparative summary, the proposed summarization method leverages different relationships between sentences to construct a comparative graph and extract sentences from papers.

4 Dataset Construction

This section mainly contains citation annotating as well as data processing. Figure 2 depicts an overview of the data construction process.

4.1 Comparative Citation Annotating

Comparative citations provide comparable information such as related and comparable articles and comparable topics for comparative summary generation. However, previous studies (Teufel et al., 2006a; Dong and Schäfer, 2011; Jha et al., 2015; Jurgens et al., 2018; Su et al., 2019) showed that the proportion of comparative citations in scientific articles are minimal (See Appendix A.1 for details). In their annotation schemes, the comparisons are scattered in different categories and are not easily distinguished. Therefore, we propose our own annotation guideline (See Appendix A.2 for details) that is sensitive to finding comparative citations. Using the guideline, three annotators are asked to annotate comparative citations in 32 papers selected from the AI area. Each paper contains 21 to 30 citations. Each citation consists of one to five sentences and cites two to six cited papers. Finally, we obtain 40 comparative citations for building the comparative citation-guided dataset.

4.2 Data Processing and Summaries Writing

With the comparative citations we get, we collect the citing and the cited papers associated with each citation from the web. The abstract, introduction, conclusion, etc. sections from papers are used for summarizing as these sections contain dense and essential information about papers. Firstly, annotators are asked to read through papers and manually write a draft comparative summary for the same topic mentioned in each comparative citation based on the crucial sentences in the papers. Secondly, five other graduate students who are professionals in the works of selected 32 papers read and modify the draft comparative summaries until they all agree that complete information such as the commonalities and differences and the salience within papers are included. The generated comparative summaries serve as the ground truth. We end up with a dataset that includes comparative citations, the citing paper, the cited papers, and the reference summaries for each citation.

5 Comparative Graph-based Scientific Summarization

We propose the comparative graph-based summarization method (CGSUM for short). The overview of the method is shown in Figure 3. The core idea of the method is to select sentences by calculating the salience of the sentences and estimating the degrees to which the sentences reflect the commonalities and differences between papers.

5.1 Construction of Comparative Graph

In the same document, a sentence receives positive influence from sentences that correlate to it, whereas in the different documents, a sentence receives negative influence from sentences that correlate to it (Li et al., 2008). And citations contain common topics between papers. For our task, there are three different relationships between two sentences: two sentences occurring in a same paper (Intra-paper Relationship); two sentences occurring in two different papers (Inter-papers Relationship); and the sentence related to citation texts (Citation-text Relationship). All three relationships are used.

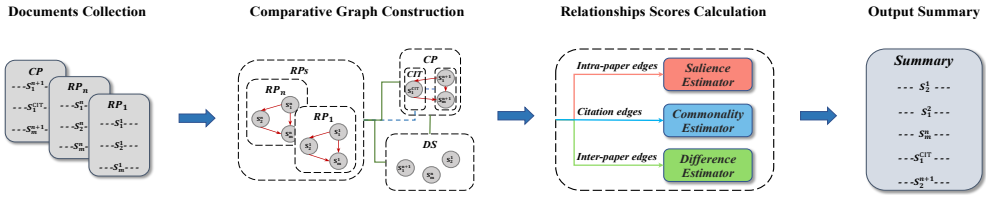


Figure 3: Overview of the comparative graph-based scientific summarization method CGSUM.

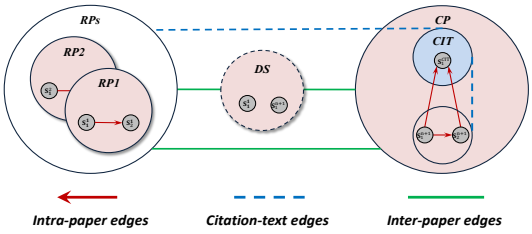


Figure 4: Example of a comparative graph that considers three different relationships between sentences. In this example, the graph contains five subgraphs {*CIT*, *CP*, *RP1*, *RP2*, *DS*}, where *CIT* denotes a comparative citation, *CP* denotes a citing paper, *RP1* and *RP2* denote two reference papers, and *DS* denotes the dynamic summary which consists of summary sentences generated so far and is initially empty and updated iteratively.

The above three relationships are used to construct the comparative graph for papers and citations. As is shown in Figure 4, the graph consists of the citing paper subgraph (the right part of the figure), the reference paper subgraphs (the left part of the figure) and the citation subgraph (the blue circle in the right part). To avoid redundancy, we introduce the dynamic summary subgraph (the circle in the middle part) which consists of summary sentences generated so far, and compare candidate sentences with the dynamic summary.

Intra-paper edges (directed solid edges) correspond to Intra-paper relationships. These edges are directed because sentences in a paper are sequentially ordered. If one sentence occurs before another sentence in a paper, the direction of the Intra-paper edge is from the former to the latter. The weights of these edges are set as similarities between sentences. These edges can be used to compute the salience of sentences.

Inter-paper edges (undirected solid edges) correspond to Inter-paper relationships. The weights of these edges are set as negative similarities between sentences. These edges can be used to compute differences between sentences.

Citation-text edges (undirected dotted edges) correspond to Citation-text relationships. Weights of these edges are similarities between the sentences in the citation and papers, reflecting the common topic in papers and citation.

5.2 Sentence Ranking and Selecting

The ranking scores of sentences are supposed to reflect the salience of sentences within papers, the degree to which sentences capture commonalities between the papers and citation, and the degree to which sentences capture differences between papers and between papers and the dynamic summary. As is shown in Task Definition, each paper is extracted to produce the comparative summary, where the sentences are salient within the paper that they belong. Sentences extracted from each paper are related to citation and reflect the commonality between papers. Besides, the extracted sentences are different from those extracted from other papers, which captures the difference between papers. Three estimators that calculate the scores of salience, commonalities and differences of sentences, are proposed on the comparative graph.

Salience estimator calculates the salience score of a sentence node by summing up the weights of its outcoming Intra-paper edges and subtracting the weights of its incoming Intra-paper edges. The contributions of any two sentences to their respective centrality are influenced by their relative positions in a document. The sentences before are central, while the sentences after supplementing them. Specifically, a sentence is salient if it has many similar sentences after the sentence. Otherwise, a sentence is redundant if it has many similar sentences before it. In the constructed graph, Intra-paper edges are directed from the sentences before (OUT) to the sentences after (IN). Therefore, for a sentence s_p, the outcoming Intra-paper edges contribute positively to its salience while the incoming Intra-paper edges contribute negatively to its salience. Equation 1 is for calculation of the salience score, where $\alpha, \beta \in [0,1]$, and $\alpha+\beta=1$.

$$SAL(s_p) = \alpha \sum_{s_o \in OUT} sim_{p,o} - \beta \sum_{s_i \in IN} sim_{p,i}, \quad (1)$$

Commonality estimator calculates the commonality score of a sentence node by summing up the weights of its Citation-text edges. The citation bridges the citing paper and the reference papers and contains the commonality of the topic shared by the papers. It is reasonable to believe that the more similar the sentences with the citation, the more common information the sentences contain. Equation 2 is for calculations of commonality scores, where the sentences s_p are in papers and the sentences s_{cit} are in citation.

$$COM(s_p) = \sum_{s_{cit} \in CIT} sim_{p,cit}, \quad (2)$$

Difference estimator calculates the difference score of a sentence node by summing up the weights of Inter-paper edges of the sentences. Sentences from different papers introduce the common topic from different aspects. Avoiding redundancy brings more differences. Our extractive method is iterative, which generates summaries by selecting sentences from papers in order of publication time. To avoid redundancy, we add an extra paper named dynamic summary. Dynamic summary is a dynamic paper consisting of the sentences of summary generated so far. It is initially empty and ends up being a comparative summary. The negative influence between the dynamic summary and papers waiting to be summarized is used to calculate the difference score of a sentence. Equation 3 is for calculations of difference scores, where s_{cp} is a sentence in the citing paper, s_{rp} is a sentence in the reference papers and s_{ds} is a sentence in the dynamic summary. Weights of Inter-paper edges are set as negative similarity values, and the difference scores are also of negative values. The higher difference score of the sentence is, the more different the sentence is from sentences in other papers.

$$DIF(s_{cp}) = -sim_{cp,ds} - \sum_{s_{rp} \in RPs} sim_{cp,rp}, \quad (3)$$

The salience score of a sentence reflects the salience of the sentence within the paper, the commonality score reflects the commonality of topic information contained in the sentence, and the difference score reflects different aspects of knowledge of the topic discussed in the citation. Equation 4 linearly interpolates the three scores as the final ranking score of the sentence, where $\lambda_1, \lambda_2, \lambda_3 \in [0,1]$, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

$$Score(s) = \lambda_1 SAL(s) + \lambda_2 COM(s) + \lambda_3 DIF(s), \quad (4)$$

Algorithm 1 CGSUM

Require: *RPs*, *CIT*, *CP*
Ensure: Comparative Summary
  for RP in RPs **do**
    for sent in RP **do**
      score = λ_1SAL(sent)+λ_2COM(sent)+λ_3DIF(sent)
      Add (sent, score) into SenScore
    end for
    Rank and select sent into Comparative Summary
    Clear SenScore
  end for
  Select CIT into Summary
  for sent in CP **do**
    score = λ_1SAL(sent)+λ_2COM(sent)+λ_3DIF(sent)
    Add (sent, score) into SenScore
  end for
  Rank and select sent into Summary
  Return Comparative Summary

With the sentences of each paper ranked by the final ranking scores, we select sentences to generate summaries. As is shown in the Pseudo code. To ensure that every papers can be summarized, we select from each and use the citation to bridge the reference papers and citing paper.

6 Experiments

6.1 Datasets

Dataset	CSSC	DUC2006	DUC2007
Domain	Sci	News	News
Query	Long	Long	Short
Clusters	40	50	45
Documents	3-5	25	25

Table 1: Statistics of the three datasets.

Experiments on comparative scientific summarization are carried out on CSSC. Additional experiments are carried out on DUC2006 and DUC2007 to show the generalization of the proposed method to multi-document summarization. As is shown in Table 1, CSSC contains citations over 40 clusters with three to five scientific papers each. DUC2006 and DUC2007 contain long queries over 50 clusters and short queries over 45 clusters, respectively.

6.2 Comparing methods

There are four kinds of comparing methods:

The first four methods: (1) ORACLE returns a extractive sentences subset with the highest ROUGE scores. (2) SIM2GOLD and (3) SIM2CIT, respectively, select three sentences which are most similar to the reference summary and the citation texts from each paper. (4) LEAD returns lead sentences (up to three) of each paper.

Models	CSSC				CSSC (concatenated)			
	R-1	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**
ORACLE (extractive)	57.4	39.1	55.3	40.7	57.4	39.1	55.3	40.7
SIM2GOLD	54.1	36.9	52.6	35.9	53.4	36.5	50.1	33.8
SIM2CIT	43.9	22.3	41.1	24.2	38.0	19.5	40.4	23.0
LEAD	43.5	21.7	40.7	23.6	33.3	14.4	30.9	16.3
Heuristic								
RANDOM (motivated by (Xing et al., 2020))	22.7	4.1	15.6	4.0	23.3	4.8	15.8	3.9
COPY-CIT (motivated by (Xing et al., 2020))	33.6	18.0	30.1	18.3	33.6	18.0	30.1	18.3
Multi-document								
Summpip (Zhao et al., 2020) (reproduce)	43.4	18.6	38.6	21.3	36.4	15.6	31.6	18.4
QUERYSUM (Xu and Lapata, 2020) (reproduce)	42.2	18.5	38.4	21.7	—	—	—	—
TIF-IDF-Sum (Lamsiyah et al., 2021)	42.8	19.2	38.1	22.3	—	—	—	—
Graph-based								
TextRank (Mihalcea and Tarau, 2004) (reproduce)	41.2	15.0	37.4	16.6	32.4	12.3	31.2	13.2
LexRank (Erkan and Radev, 2004) (reproduce)	42.1	18.0	37.5	18.9	36.3	11.5	31.8	15.0
PACSUM (Zheng and Lapata, 2019) (reproduce)	42.3	19.1	37.5	21.6	37.9	14.2	33.3	17.3
HIPORANK (Dong et al., 2021) (reproduce)	42.0	17.9	37.5	15.3	—	—	—	—
Ours								
CGSUM-TF-IDF	47.2	25.5	43.5	27.1	41.6	17.6	36.4	19.9
CGSUM-BERT	**48.6**	**28.5**	**45.3**	**28.7**	**42.1**	**20.4**	**40.8**	**21.3**

Table 2: Automatic evaluation results on CSSC and CSSC(concatenated). **Bold** indicates the best result.

Heuristic: (1) RANDOM randomly selects three sentences of each paper. (2) COPY-CIT treats the citation texts as the output.

Multi-document: (1) Summpip (Zhao et al., 2020) is an unsupervised graph-based method for multi-document summarization. (2) QUERYSUM (Xu and Lapata, 2020) is a query-focused framework for estimating relevant text segments, and (3) TF-IDF-Sum (Lamsiyah et al., 2021), which estimates relevant sentences for query-focused multi-document summarization.

Graph-based: (1) TextRank (Mihalcea and Tarau, 2004) and (2) LexRank (Erkan and Radev, 2004) are unsupervised methods based on Markov random walks. (3) PACSUM (Zheng and Lapata, 2019) and (4) HIPORANK (Dong et al., 2021) are directional graph-based methods considering the relative position and the hierarchy, respectively.

6.3 Results on CSSC

6.3.1 Automatic evaluations

We evaluate our models with ROUGE (Lin, 2004), reporting the F1 scores for ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4. The test results are shown in Table 2, where CSSC (concatenated) concatenates the citing and the cited papers into one for each. The test results on CSSC (concatenated) are used to verify the hypothesis that it is necessary to select from each document.

Our models outperform comparative baselines on CSSC. The results of COPY-CIT show the particularity of the task in this paper and deduct that citation generation does not match our task. Our proposed models, CGSUM-TF-IDF and CGSUM-BERT, obviously outperform the baseline models. This result proves the effectiveness of our comparative graph-based summarizer, which considers different relationships between sentences. It requires the model to capture the critical content of the cited paper and to capture the attitude of the citing paper to the cited paper. The model not only needs to generate fluent and informative text but also needs to ensure contextual coherence. The results on CSSC are all higher than the results on CSSC (concatenated), which means it is necessary to select from each document because it ensures that the generated summary can reflect the salience, commonality, and difference of each document, which avoids information miss.

6.3.2 Human evaluations

We adopt the QA paradigm and the similarity between the gold summaries and system summaries to evaluate summaries quality. For the QA paradigm, reviewers create questions (e.g. salient content of each paper, common topic in the citation, and different aspects concerning the common topics) based on gold summaries. They examine whether system summaries can answer these questions. The more detailed questions the system summaries can answer, the better they are. For the similarity, reviewers assess the degree to which system summaries retain the salience of papers and the commonality and difference between papers. Specifically, the

Models	CSSC				
	SAL	COM	DIF	COH	ALL
ORACLE	4.50	4.78	4.36	4.06	4.44
SIM2GOLD	4.38	4.36	4.20	3.96	4.24
LEAD	3.56	3.16	3.68	3.24	3.42
Summpip (2020)	3.68	3.80	3.58	3.42	3.68
QUERYSUM (2020)	3.58	4.06	3.36	3.24	3.56
PACSUM(2019)	3.62	3.82	3.50	3.38	3.58
HIPORANK(2021)	3.72	3.78	3.52	3.28	3.58
CGSUM-BERT	**4.12**	**4.02**	**3.98**	**3.86**	**4.00**

Table 3: Human evaluation results on CSSC. **SAL**ience, **COM**monality, **DIF**ference, **COH**erence, **ALL** is the average across all scores. **Bold** indicates the best result.

salience score is assessed by comparing the similarities between system summaries and abstracts of each paper. The commonality is assessed by comparing the similarities between system summaries and citations. The difference is assessed by comparing the dissimilarity between system summaries and abstracts of each paper on the common topic. The coherence of system summaries is also taken by assessing their readability. After two stages of review, each reviewer gives each human evaluation metric a score of 0.0-5.0 based on the questions they created and the similarity they assessed. These scores will be averaged to obtain a final score for the system summary.

As is shown in Table 3, our models outperform the baseline models. The COH score and the ALL score of our models are especially higher than that of the baselines. This result further demonstrates the efficacy of our proposed models. Using the salience, commonality and difference estimators, CGSUM captures salience within papers and commonalities and differences between papers. Summaries created by CGSUM are also more coherent by using citations to join the contents of papers.

6.4 Results on DUC2006 and DUC2007

The results on DUC2006 and DUC2007 are summarized in Table 4. GRSum (Wan, 2008) integrated query-relevance into a Graph Ranking algorithm. C-Attention (Li et al., 2017) compresses multi-document summarization. The results show that our models perform well on the DUC2006 and DUC2007 datasets. Compared to the results ORACLE gets, the results our models get mean that our extractive models are almost close to the mostly perfect extractive summaries at the sentence level. The exciting conclusion shows that our comparative graph-based models are promising to be applied for the multi-document summarization task.

Models	DUC2006			DUC2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
ORACLE	40.6	9.1	14.8	41.8	10.4	16.0
LEAD	32.1	5.3	10.4	33.4	6.5	11.3
Graph-based						
LexRank	34.2	6.4	11.4	35.8	7.7	12.7
GRSUM	38.4	7.0	12.8	42.0	10.3	15.6
TF-IDF-Sum	39.0	7.9	13.8	40.1	10.1	15.2
Compress-based						
C-Attention	39.3	8.7	14.1	42.3	10.7	16.1
QUERYSUM	**41.1**	**9.6**	**15.1**	**42.9**	**11.6**	**16.7**
Ours						
CGSUM-TF-IDF	39.8	8.2	14.0	41.0	9.8	15.5
CGSUM-BERT	40.1	8.4	14.3	41.2	10.3	15.7

Table 4: Automatic evaluation results on DUC2006 and DUC2007. **Bold** indicates the best result overall. Underline denotes the best sentence-extractive results.

Models	CSSC			
	R-1	R-2	R-L	R-SU4
CGSUM-TF-IDF	**47.2**	**25.5**	**43.5**	**27.1**
w/o Salience	↓43.7	↓19.8	↓39.5	↓22.4
w/o Commonality	↓41.5	↓17.8	↓37.8	↓20.7
w/o Difference	↓45.3	↓23.4	↓41.3	↓25.3
CGSUM-BERT	**48.6**	**28.5**	**45.3**	**28.7**
w/o Salience	↓42.2	↓18.4	↓38.6	↓20.9
w/o Commonality	↓44.8	↓22.3	↓41.5	↓24.3
w/o Difference	↓43.7	↓19.8	↓40.0	↓22.3

Table 5: Ablation results on CSSC. ↓ denotes decrease.

6.5 Ablation Studies

Ablation studies in Table 5 are carried out to show effects of two representations and three estimators.

- **Representations** include TF-IDF and BERT. BERT performs better than TF-IDF.
- **w/o Salience** represents CGSUM without the salience estimator, and it performs worse than CGSUM with the salience estimator, indicating that salience estimator is effective in capturing salient information within papers.
- **w/o Commonality** represents CGSUM without the commonality estimator, and it performs worse than CGSUM with the commonality estimator, indicating that commonality estimator is effective because the estimator can find the commonality between papers.
- **w/o Difference** represents CGSUM without the difference estimator, and it performs not as well as CGSUM with the difference estimator, implying that difference estimator is effective.

Removing each estimator leads to a drop of the performance of CGSUM. Meaning estimators capture different information to produce summaries.

Comparative Citation from (*Devlin et al., 2018*):

Unlike recent ***language representation models*** (*Peters et al., 2018; Radford et al., 2018*), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

GOLD:

All three works belong to the field of Natural Language Processing and are all about language representation models. [*Peters et al., 2018*] introduced a type of deep contextualized word representation that models both complex characteristics of word use, and how these uses vary across linguistic contexts. The vectors are derived from a bidirectional LSTM that is trained with a coupled language model objective on a large text corpus. [*Radford et al., 2018*] explored a semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning. The approach introduces a framework for achieving strong natural language understanding with a single task-agnostic model through generative pre-training and discriminative fine-tuning. [*Devlin et al., 2018*] introduced a language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers and is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers.

CGSUM:

[*Peters et al., 2018*]: We introduce a new type of ***deep contextualized word representation*** that models both (1) complex characteristics of word use, and (2) how these uses vary across linguistic contexts. We use vectors derived from a bidirectional LSTM that is trained with a coupled language model objective on a large text corpus. Unlike previous approaches for learning ***contextualized word vectors*** (Peters et al., 2017; McCann et al., 2017), ELMo representations are deep, in the sense that they are a function of all of the internal layers of the biLM. [*Radford et al., 2018*]: We demonstrate that large gains on these tasks can be realized by generative pre-training of a ***language model*** on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task. We introduced a framework for achieving strong ***natural language understanding*** with a single task-agnostic model through generative pre-training and discriminative fine-tuning. By pre-training on a diverse corpus with long stretches of contiguous text our model acquires significant world knowledge and ability to process long-range dependencies which are then successfully transferred to solving discriminative tasks such as question answering. [*Devlin et al., 2018*]: Unlike recent ***language representation models*** (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The major limitation is that standard ***language models*** are unidirectional, and this limits the choice of architectures that can be used during pre-training. We introduce a new ***language representation*** model called BERT, which uses masked language models to enable pre-trained deep bidirectional representations.

Comparative Citation from (*Tran et al., 2020*):

To ***encode*** the article text we use RoBERTa. Unlike GloVe (*Pennington et al., 2014*) and word2vec (*Mikolov et al., 2013*) ***embeddings***, RoBERTa uses BPE which can encode any word made from Unicode characters.

GOLD:

All three studies are about embeddings for text. By subsampling of the frequent words. [*Mikolov et al., 2013*] obtained significant speedup and also learn more regular word representations. The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document. [*Pennington et al., 2014*] used their insights to construct a new model for word representation which they called GloVe. They constructed a model that utilizes the benefit of count data while simultaneously capturing the meaningful linear substructures prevalent in recent log-bilinear prediction-based methods like word2vec. To encode the article text, [*Tran et al., 2020*] used RoBERTa, a recent improvement over the popular BERT model. RoBERTa is a pre-trained language representation model providing contextual embeddings for text. It consists of 24 layers of bidirectional transformer blocks.

CGSUM:

[*Mikolov et al., 2013*]: The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed ***vector representations*** that capture a large number of ***precise syntactic and semantic word relationships***. By subsampling of the frequent words we obtain significant speedup and also learn more regular ***word representations***. We also describe a simple alternative to the hierarchical softmax called negative sampling, and show how to train distributed ***representations of words and phrases*** with the Skip-gram model and demonstrate that these representations exhibit linear structure that makes precise analogical reasoning possible. [*Pennington et al., 2014*]: Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a ***word vector space*** with meaningful sub-structure utilizes the main benefit of count data while simultaneously capturing the meaningful linear substructures prevalent in recent log-bilinear prediction-based methods like word2vec. The result, GloVe, is a new global log-bilinear regression model for the unsupervised learning of ***word representations*** that outperforms other models on word analogy, word similarity, and named entity recognition tasks. [*Tran et al., 2020*]: Unlike GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) ***embeddings***, RoBERTa uses BPE which can ***encode any word*** made from Unicode characters. In BPE, each word is first broken down into a sequence of bytes. Common byte sequences are then merged using a greedy algorithm.

Figure 5: Case studies on examples taken from the CSSC dataset.

6.6 Case Study

Case studies in Figure 5 are carried out on examples taken from CSSC. All comparable works related to the corresponding comparative citation are marked as [*author, year*]. The commonality is marked in green and italicized while difference is marked in green and underlined. As the Figure 5 shows, summaries created by CGSUM cover detailed salience within papers and commonalities and differences between papers, and are also quite coherent.

7 Conclusion

This paper proposes the novel task of comparative citation-guided summarization of scientific papers, which aims to summarize commonalities and differences between the articles and related to the comparable topic mentioned in comparative citations. The CSSC dataset for the task is constructed, which contains 40 groups of comparable scientific papers and corresponding reference summaries by annotating and collecting comparative citations. The unsupervised comparative graph-based summarization CGSUM method is proposed to generate comparative summaries. It utilizes three different relationships of sentences to build a comparative graph and calculates the scores of salience, commonality and difference without large-scaled data. Experiments on CSSC show that CGSUM outperforms baselines. Experiments on DUC2006 and DUC2007 demonstrate that CGSUM can be generalized to multi-document summarization tasks. In the future, we would like to study more types of relationships between documents and research the comparative scientific summarization cross the fields.

Acknowledgements

This research was sponsored by the National Natural Science Foundation of China (No.61806101). We also thank Wenwen Fan and Tong Liu for annotating the citations, and Feng Xie, Yirui Huang, Lingyun Jin, Xianzhe Xu and Qingsen Bao for performing the ground truth and human evaluation.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proc. of NAACL*.

Ahmed Ghassan Tawfiq AbuRa'ed and Horacio Saggion. 2021. A select and rewrite approach to the generation of related work reports. In *Proc. of CEUR Workshop*.

Ahmed AbuRa'ed, Horacio Saggion, Alexander Shvets, and Àlex Bravo. 2020. Automatic related work section generation: experiments in scientific document abstracting. *Scientometrics*.

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. In *Proc. of AAAI*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Proc. of EMNLP Findings*.

Jingqiang Chen and Hai Zhuge. 2016. Summarization of related work through citations. In *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*.

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 *Proc. of ACL*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proc. of NAACL*.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakova, and Marin Soljacic. 2021. We can explain your research in layman's terms: Towards automating science journalism at scale. In *Proc. of AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proc. of IJCNLP*.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proc. of EACL*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*.

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 *Proc. of ACL*.

Lei He, Wei Li, and Hai Zhuge. 2016. Exploring differential topic models for comparative summarization of scientific papers. In *Proc. of COLING*.

Myriam Hernandez-Alvarez, José M. Gomez Soriano, and Patricio Martínez-Barco. 2017. Citation function, polarity and influence classification. *Natural Language Engineering*.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proc. of COLING*.

Rahul Jha, Reed Coke, and Dragomir Radev. 2015. Surveyor: A system for generating coherent survey articles for scientific topics. In *Proc. of AAAI*.

Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R. Radev. 2017. Nlp-driven citation analysis for scientometrics. *Natural Language Engineering*.

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*.

Salima Lamsiyah, Abdelkader El Mahdaouy, Said Ouatik El Alaoui, and Bernard Espinasse. 2021. Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, bm25 model, and maximal marginal relevance criterion. *Journal of Ambient Intelligence and Humanized Computing*.

Piji Li, Wai Lam, Lidong Bing, Weiwei Guo, and Hang Li. 2017. Cascaded attention based unsupervised information distillation for compressive summarization. In *Proc. of EMNLP*.

Wenjie Li, Furu Wei, Qin Lu, and Yanxiang He. 2008. Pnr2: Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proc. of COLING*.

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

Mauricio Marrone. 2020. Application of entity linking to identify research fronts and trends. *Scientometrics*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proc. of EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of ACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL*.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proc. of ACL*.

Xuan Su, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama. 2019. Neural multi-task learning for citation function and provenance. In *Proc. of JCDL*.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006a. An annotation scheme for citation function. In *Proc. of SIGDIAL*.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006b. Automatic classification of citation function. In *Proc. of EMNLP*.

Xiaojun Wan. 2008. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval*.

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

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proc. of ACL*.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proc. of EMNLP*.

Shansong Yang, Weiming Lu, Zhanjiang Zhang, Baogang Wei, and Wenjia An. 2016. Amplifying scientific paper's abstract by leveraging data-weighted reconstruction. *Information Processing & Management*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proc. of SIGIR*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proc. of ACL*.

A Appendix

A.1 Citation Function Annotation Proportion of Scientific Papers

Table 6 shows the proportions of *Neutral* and *Comparative* citations annotated by some annotation schemes (Teufel et al., 2006a; Dong and Schäfer, 2011; Jha et al., 2017; Jurgens et al., 2018; Su et al., 2019). From the results, we can find that *Neutral* always has the highest percentage, while *Comparative* always has a low percentage. In conclusion, comparative citations are always challenging to discover.

Schemes	Key Categories	Proportion
Teufel et al. (2006a)	Neut	59.62%
	CoCoGM	4.65%
	CoCoR0	1.27%
	CoCo-	1.54%
	CoCoXY	3.11%
Dong and Schäfer (2011)	Background	65.04%
	Comparison	3.97%
Jha et al. (2017)	Neutral	61.15%
	Comparison	5.82%
Jurgens et al. (2018)	BACKGROUND	51.13%
	COMPARISON	18.07%
Su et al. (2019)	Neutral	70.83%
	Compare	6.42%

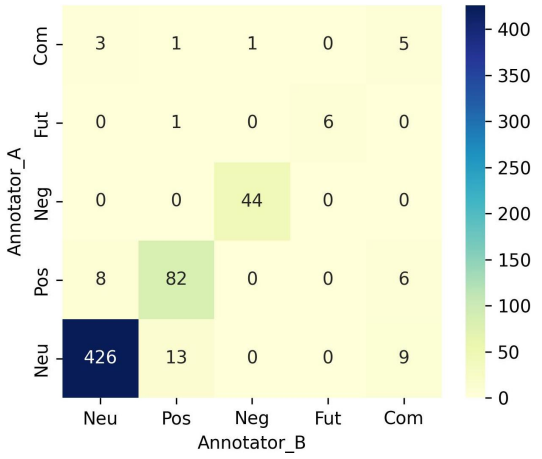
Table 6: The proportions of key citations.

A.2 Guideline for Annotating Citation

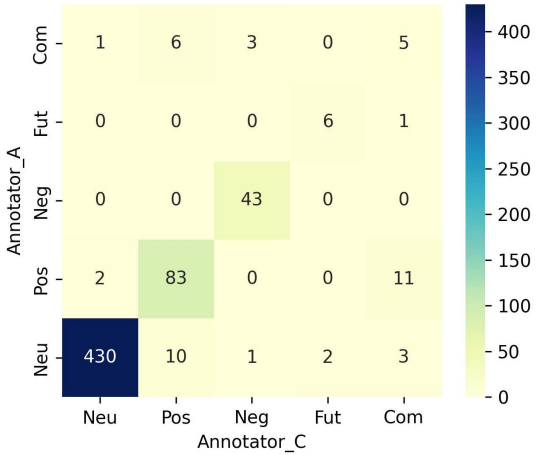
Our guideline is similar to the scheme of (Teufel et al., 2006a) but with different classifications. Taking the functions of *PModi* and *PBas* as examples, they belong to *Positive* in Teufel's. However, alterations accompany modifications and bases and we thus classify them as *Comparative*. Besides, we add the *Future* function for it is also crucial and unique in researches.

A.3 Annotating Citations of Surveys

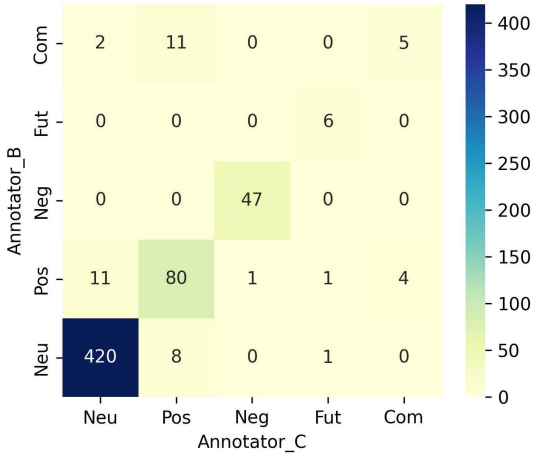
We annotate citations in surveys. Specifically, we collect eight scientific surveys from the Artificial



(a) Annotations between A&B.



(b) Annotations between A&C.



(c) Annotations between B&C.

Figure 5: Consistency of annotations of surveys between three annotators. Neu, Pos, Neg, Fut, and Com stand for the Neural, Positive, Negative, Future, and Comparative, respectively.

Categories	Description
Neutral	Normal descriptions of the cited works, or not enough textual evidence for other categories.
Positive	Authors agree with the cited works, their work and the cited works support each other.
Negative	Authors disagree with the cited works, their work is the opposite of the cited works.
Future	Authors show some hypothesis or feasible future works based on the cited works.
Comparative	Comparisons/Alterations between the works.

Table 7: Our annotating guideline.

	Neutral	Positive	Negative	Future	Comparative
A	73.81%	15.82%	7.25%	1.15%	1.98%
B	72.32%	15.98%	7.41%	0.99%	3.29%
C	71.33%	16.31%	7.91%	1.32%	3.13%
Kappa(n=5; N=607; k=3)=0.8353; **Macro-F**=0.7868					

Table 8: Citations proportions with Kappa and Macro-F.

Intelligence area and extract 607 citations from surveys. Three graduate students use our guideline to annotate 607 citations. The annotation results are shown in Table 8. Figure 5 and the values of Kappa and Macro-F in Table 8 also indicate that the annotations are of high consistency. It can be seen from Table 8 that in scientific surveys, the proportion of citations in the *Comparative* is much lower than that in other categories and is close to *Future*, which rarely appears. In comparison, the proportion of citations in the *Neutral* function is the highest. In conclusion, there are usually little comparisons in surveys. Therefore, the tasks of related work generation and survey generation are not suitable for generating differences in scientific papers, while comparative summaries capture commonalities and differences.

A.4 Implementation Details

We utilize TF-IDF and BERT (Devlin et al., 2018) to get sentence representation. The hyperparameters (α, β) are set as (0.9, 0.1), (λ_1, λ_2, λ_3) are set as (0.33, 0.33, 0.33).

https://github.com/google-research/bert