

國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

生成科學論文之條列式貢獻

Generating Disentangled Contributions for Scientific  
Documents

劉孟寰

Meng-Huan Liu

指導教授: 陳信希 博士

Advisor: Hsin-Hsi Chen Ph.D.

中華民國 111 年 8 月

August, 2022

國立臺灣大學碩士學位論文  
口試委員會審定書

生成科學論文之條列式貢獻

Generating Disentangled Contributions for Scientific Documents

本論文係劉孟寰君（學號 R09944022）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國一百一十一年八月十八日承下列考試委員審查通過及口試及格，特此證明。

口試委員：

陳信希

（簽名）

（指導教授）

鄭卜壬

蔡銘暉

陳名身

鄭卜壬

所長：



## Acknowledgements

我的研究和論文能順利完成，首先要感謝的是指導老師陳信希教授。儘管平時日程繁忙，老師每週都仍會撥出時間和大家討論每個人的研究進度並交流想法。在我剛開始進行研究的時候，老師總是給予我充分的自由去探索感興趣的議題，在我面對困難的時候，老師也會鼓勵我並一起討論各種改進方向。自然語言處理領域的發展瞬息萬變且充滿挑戰，而老師對於議題的發想和解讀總能讓我受益良多。

再者，口試期間承蒙口試委員鄭卜壬教授、蔡銘峰教授和陳冠宇教授提供諸多寶貴的意見，使論文能夠更加完善，在此感謝老師們的蒞臨指教。

此外，我也要感謝實驗室的大家的照顧。謝謝瀚萱學長、安孜學姐、重吉學長和建宏學長每週一起參與討論；謝謝同組的法宣學長、柏君學長、柏承和宏晉，大家一起分享論文和技術細節讓我的研究更加完整和充實；謝謝網管禹廷學長、佑恩、韋霖和彥斌維護實驗室的資源；謝謝又慈打理實驗室的各種事務；謝謝同屆的宏哲、偉鋒和知遙，還有所有學長姐和學弟妹們，承蒙大家的照顧了。

最後，我要感謝我的父母和家人們，你們總是無條件支持我，讓我能持續成長。兩年的時間過得很快，初入研究所時的種種緊張和新鮮彷彿還在眼前，謝謝所有家人和朋友的陪伴，求學之路暫時告一段落，也期許自己能帶著這段充實的經歷和所學投入到未來的每個挑戰中。



## 摘要

科學論文的貢獻側重於描述其原創之處和重要價值，對於每個科學研究來說這都可以被認為是其最核心的部分。一個能精確辨認論文貢獻並將其組織為結構化摘要的系統對於輔助自動化處理科學文本和幫助讀者理解等應用具有潛在價值。雖然近期的工作開始致力於與論文貢獻相關的任務的研究中，目前仍缺少高品質的大規模資料集來輔助深度學習模型的訓練。有鑑於此，我們收集並整理了一個資料集，其中包含大約兩萬四千篇計算機科學領域的論文及其作者條列之貢獻，根據我們提出的標記框架，這些科學貢獻又被進一步分為對應的不同類別。接著我們正式定義了生成科學論文之條列式貢獻這個任務。利用大量的無監督資料和原始論文中重要語句以及生成目標所包含的貢獻類別，我們提出了一個細粒度的訓練策略。實驗結果表明我們提出的方法優於具競爭力的基線模型和其他訓練策略，證明了其有效性。我們也進行了詳細分析以研究我們所提出的資料集和任務的特性及其挑戰之處。

**關鍵字：**科學文本處理，抽象式摘要，科學貢獻生成



# Abstract

Contributions of scientific papers highlight their novelty and key values, which are essentially the core parts of every research work. Systems that are capable of identifying the contributions of the papers precisely and organizing them into well-structured summaries are valuable in aiding both automatic text processing and human comprehensions. Though recent works have focused more on tasks dealing with the contributions of the scientific documents, there is currently no large-scale dataset with high quality that can facilitate the training of modern deep learning based models. To this end, we curate a dataset consisting of 24K computer science papers with contributions explicitly listed by the authors, which are further classified into different contribution types based on our newly-introduced annotation scheme. Then we formally formulate the task of generating disentangled contributions for scientific documents. We present fine-grained post-training strategy leveraging abundant unsupervised data and the contribution types of both high-light sentences in the source documents and the generation targets. Experimental results

show that the proposed method outperforms competitive baselines and other post-training strategies, demonstrating the effectiveness of our approach. Detailed analysis is also conducted to study the characteristics and challenges of our dataset as well as the newly-proposed task.

**Keywords:** Scholarly Document Processing, Abstractive Summarization, Research Contribution Generation



# Contents

	<b>Page</b>
<b>Acknowledgements</b>	<b>i</b>
<b>摘要</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivation and Contribution . . . . .	3
1.3 Thesis Organization . . . . .	7
<b>Chapter 2 Related Work</b>	<b>8</b>
2.1 Scholarly Document Processing . . . . .	8
2.2 Abstractive Summarization . . . . .	11
<b>Chapter 3 Datasets</b>	<b>14</b>
3.1 Dataset Collection . . . . .	14
3.2 Dataset Analysis . . . . .	18
3.2.1 Dataset Statistics . . . . .	18



3.2.2	Comparisons with Existing Datasets . . . . .	19
3.2.3	Structural Alignments . . . . .	20
3.3	Contribution Type Annotation . . . . .	22
3.3.1	Annotation Scheme . . . . .	23
3.3.2	Annotation Procedure and Results . . . . .	24
3.3.3	Analysis of Contribution Types . . . . .	25
<b>Chapter 4</b>	<b>Methodology</b>	<b>28</b>
4.1	Contribution Type Classification . . . . .	28
4.2	Disentangled Contribution Generation . . . . .	29
4.2.1	Task Formulation and Model Architecture . . . . .	29
4.2.2	Finetune for Disentangled Contribution Generation . . . . .	31
4.2.3	Fine-grained Post-training . . . . .	36
<b>Chapter 5</b>	<b>Experiments</b>	<b>39</b>
5.1	Contribution Type Classification . . . . .	39
5.2	Disentangled Contribution Generation . . . . .	41
5.2.1	Experimental Setup . . . . .	41
5.2.2	Evaluation Metrics . . . . .	42
5.2.3	Main Results . . . . .	45
5.2.4	Comparisons with Other Post-training Strategies . . . . .	48
<b>Chapter 6</b>	<b>Discussion</b>	<b>50</b>
6.1	Ablation Study . . . . .	50
6.2	Experiment Results in Low Resource Setting . . . . .	51
6.3	Results Based on Different Contribution Types . . . . .	53



6.4	Analysis of Contribution-Level Evaluations . . . . .	56
<b>Chapter 7</b>	<b>Conclusion</b>	<b>60</b>
<b>References</b>		<b>62</b>





# List of Figures

1.1	An example paper with explicitly listed contributions by the authors . . .	4
3.1	Examples of common patterns in which authors list their contributions explicitly . . . . .	15
3.2	Relative positions of greedily extracted sentences . . . . .	22
3.3	Illustrations of our contribution type annotation scheme . . . . .	24
3.4	Transition matrix of different contribution types . . . . .	25
3.5	Examples of common contribution patterns . . . . .	27
4.1	Organization of input and target from an example in our dataset . . . . .	35
4.2	Example of our fine-grained sentence masking strategy . . . . .	38
5.1	Confusion matrix of contribution type classification . . . . .	40
5.2	Example of contribution-level matching . . . . .	44
6.1	Case study of contribution type coverage of LED-FP . . . . .	55
6.2	Histogram of the ratios of matched contributions generated by our model	57
6.3	Case study of contribution-level mapping of LED-FP . . . . .	59



# List of Tables

3.1	Basic statistics of our dataset . . . . .	19
3.2	Comparisons of ContributionSum with existing datasets . . . . .	20
3.3	ROUGE scores between targets and sections in our dataset . . . . .	21
5.1	Results of Summary Level Evaluation . . . . .	46
5.2	Results of Contribution Level Evaluation . . . . .	47
5.3	Comparisons with other post-training strategies incorporated with LED .	49
6.1	Ablation study of LED-FP . . . . .	50
6.2	Results of zero-shot and few-shot experiments . . . . .	52
6.3	Results based on different contribution types . . . . .	53
6.4	Comparisons of the contribution numbers in generation results and refer- ences . . . . .	57
6.5	Disentanglement scores of the references and the generation results of LED-FP . . . . .	58



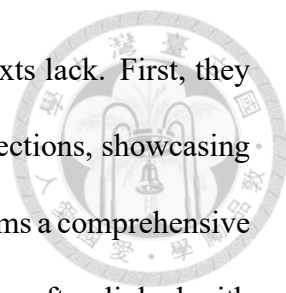
# Chapter 1 Introduction

## 1.1 Background

The volume of the scientific literatures is growing at a rapid rate, especially in those trending research fields. According to the statistics from arXiv<sup>1</sup>, there were 76,578 submissions dated back to 2011, while in 2021 the number increased to 181,630. If we focus on submissions categorized into computer science only, the margin is even bigger – almost tenfold as the number grew from 7,581 to 66,254. At such high pace, it is nearly impossible for the researchers to keep up with every latest finding. To address the issue of information overload, automatic methods to process scientific documents become compelling. With the growing popularity of digital archives providing rich resources for constructing large-scale datasets as well as the huge advances in neural network based NLP models, huge efforts have been put into the aforementioned topic recently, also known as the field of scholarly document processing. Various tasks have been explored to facilitate its development such as paper summarization, paper recommendation, citation intent classification, review generation and scientific knowledge graph construction. As the intersection of Natural Language Processing, Information Retrieval, Data Mining and Digital Libraries, scholarly document processing differs from other research topics in that scien-

---

<sup>1</sup>[https://arxiv.org/stats/monthly\\_submissions](https://arxiv.org/stats/monthly_submissions)

The logo of National Taiwan University (NTU) is located in the upper right quadrant of the page. It is a circular emblem with a central bell and the university's name in Chinese characters around the perimeter.

tific documents exhibit unique properties that common web-based texts lack. First, they are typically much longer and yet carefully organized into different sections, showcasing the importance of the underlying discourse in a scientific paper that forms a comprehensive picture of its complete research process. Second, scientific papers are often linked with their related works through inline references, the resulting network formed by hundreds of thousands of links between papers provide valuable overviews of their corresponding research fields. Last but not least, the complicated structures and layouts of scientific papers pose challenges to the construction of machine-readable corpora compared with plain texts. Section headers, paragraph breaks, footnotes, metadata, citation marks, figures, tables and their captions all need rigorous post-processing based on the formulation and the need of downstream tasks.

As an interesting and indispensable element in every scientific literature, research contributions and tasks related to it remain understudied. Contributions of a scientific research highlight the novelty and key values that make it stand out from previous works. They serve as important roles in various kinds of applications. For researchers who want to quickly grasp the key points of papers and discern ones that worth digging into, contributions are of great values especially for those who are already familiar with the backgrounds of the related research fields. For the process of paper reviewing, one may argue that the evaluation of a scientific work is essentially the evaluation of what it contributes to the research community. As a result, some venues stipulate that the reviewers should state the contributions of the reviewed target in the review-rebuttal process, and authors also start to explicitly list the contributions in the papers to highlight their work. For the field of scholarly document processing, tasks like knowledge graph construction, entity extraction and paper recommendation are likely to benefit from it since contributions contain

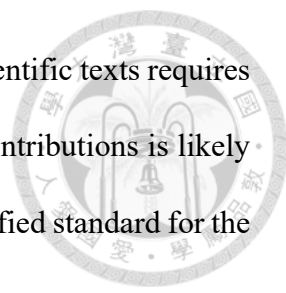
the most salient information in the papers and thus provide alternatives to the lengthy and complicated raw papers for automatic methods to build on.



## 1.2 Motivation and Contribution

Existing works in scientific paper summarization mainly consider abstracts of papers as the reference summaries, since the role of the abstract in the whole paper structure is indeed to provide a summary of the research work and it is also easily accessible in nearly every scientific document. As a result, popular large-scale benchmark datasets were built by automatically collecting papers and their abstracts from open-access digital archives [1, 2]. However, from the application’s perspective, this might seem redundant as human-written abstracts already exist. In addition, abstracts don’t necessarily contain key contents in the research work only. For instance, authors often introduce the backgrounds and closely-related works at the beginning of the abstract – information that is not likely to be considered as primarily important by domain experts and experienced researchers. Besides, abstracts may still be considered too lengthy for readers to quickly get to the main points of the papers. Instead, they might prefer well-structured summaries [3] with more concise writing styles that can hit the mark such as bullet point lists gathering and organizing the principal information of the papers.

On the other hand, contributions, as mentioned in the background section, are perfect replacements for abstracts as the reference summaries in the task of scientific paper summarization. However, due to the lack of automatic methods to acquire high-quality research contributions, there is currently no large-scale dataset available in this regard. Recent works over research contributions rely on human efforts to collect and annotate



datasets [4–6]. This imposes significant costs as the annotation of scientific texts requires knowledge from domain experts. In addition, the identification of contributions is likely to be subject to personal judgement, making it tricky to establish a unified standard for the annotation.

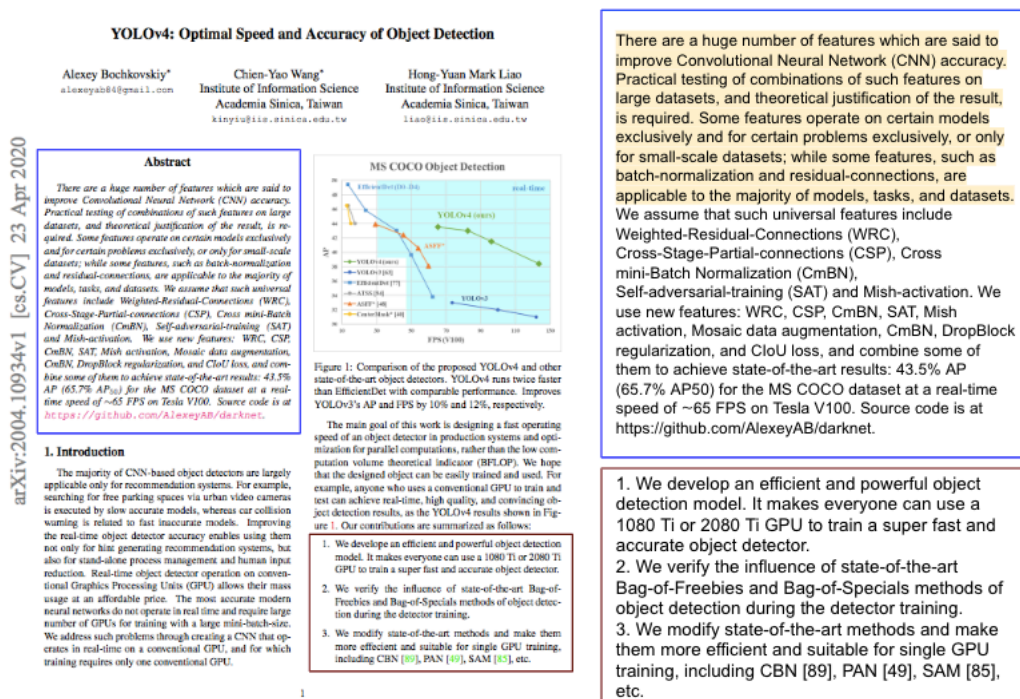
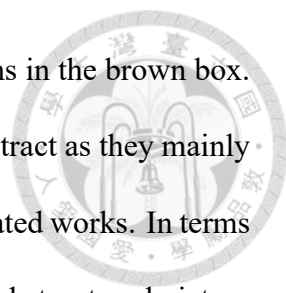


Figure 1.1: One example paper<sup>2</sup> with its abstract marked in the blue box and explicitly listed contributions by the authors in the brown box, highlighted texts in abstracts are not contribution-related.

In this thesis, we aim to build a model capable of generating disentangled contributions for scientific documents. This new task’s goal is to summarize the research paper into several key points that highlight the most important contributions made by the authors. We observe that more and more researchers start to explicitly list their contributions in the paper, especially in AI-related domains. Notably, there might be several contributions presented in a paper and they probably represent different types of contributions as well. Ideally, the desired system should be able to generate them separately yet sequentially since they are closely related to each other. Figure 1.1 shows one example paper with

<sup>2</sup>Original paper: <https://arxiv.org/pdf/2004.10934.pdf>

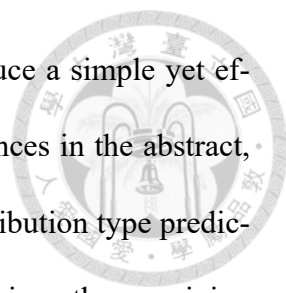


its abstract presented in the blue box and explicitly listed contributions in the brown box. These contributions serve as a better summary compared with the abstract as they mainly focus on the authors' research instead of the background and other related works. In terms of writing skills, readers are provided with a more comprehensive and structured picture as these contributions combined form a highlight story of the research process throughout the paper.

To facilitate the development of automatic method tackling the proposed task as well as other tasks related to research contributions, we present ContributionSum, a contribution summarization dataset built on arXiv papers in computer science categories. In comparison with previous works, we automatically collect contributions written by the authors in the paper so that large-scale dataset is constructed minus the significant costs of human labors as well as the potential errors result from the annotation adjudication.

Furthermore, we propose an annotation scheme for contribution type classification. The predicted contribution type of each disentangled target serves as a dual role in our work. From the perspective of potential applications, they provide further explanations about the generated results to improve the comprehension of the readers with additional clarity. The generated results along with their corresponding contribution types may also benefit downstream tasks such as scientific knowledge graph construction and entity extraction where the absence of human annotated datasets prevails. From the perspective of model designs, these contribution types provide sketch supervisions that can guide the summarization process and align with important sentences in the source documents, thus improve model performances. Based our annotation scheme, we provide human annotations of contributions in 1K papers and apply a data-driven approach to annotate all the other contributions in our dataset.



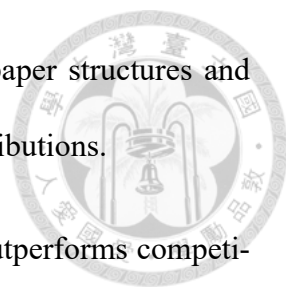


Built on the Gap Sentence Generation objective [7], we introduce a simple yet effective sentence masking strategy tailored to our task. Salient sentences in the abstract, introduction and conclusion sections are masked based on their contribution type predictions and heuristic rules, and the model is trained to generate them given the remaining parts of the papers. Based on the proposed strategy, we exploit papers without contributions written by the authors in our corpus to construct pseudo summaries and utilize them as self-supervised data in the fine-grained post-training stage. For the model architecture, we leverage the powerful transformer-based models [8] pretrained on large-scale corpus [9, 10] and incorporate them with paper structures as well as highlight contribution sentences. These models are then post-trained with our sentence masking strategy and finally finetuned on our gold dataset to generate disentangled contributions for scientific documents.

We conduct extensive experiments and the results of automatic metrics on both summary-level and individual contribution-level evaluations demonstrate the improvement of our proposed method over competitive baselines and other post-training strategies. We also perform ablation study and detailed analysis to investigate the shortcomings of the existing models, which can hopefully inspire future researches in this newly introduced task.

Our contributions in this thesis are summarized as follow:

- We introduce the task of generating disentangled contributions for scientific documents to build a better summarization system that benefits both researchers and the field of scholarly document processing.
- A large-scale dataset is constructed by automatically collecting and extracting contributions listed by the authors in computer science papers from arXiv.

- 
- We propose a novel post-training strategy and leverage the paper structures and highlight contribution sentences to generate disentangled contributions.
  - Experimental results demonstrate that our proposed method outperforms competitive baselines in both summary-level and individual contribution-level evaluations. A comprehensive analysis is also conducted to further investigate the shortcomings of our model and the characteristics of the new task.

### 1.3 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 summarizes the related works of our research. Chapter 3 introduces the collection and annotation scheme of ContributionSum. We also provide detailed analysis as well as comparisons with previous datasets. In chapter 4, we present our novel post-training strategy and generation models. Implementation details and main experimental results are shown in Chapter 5. Further discussion and analysis of our models are presented in Chapter 6. Finally, we conclude the thesis and discuss future works in Chapter 7.



## Chapter 2 Related Work

### 2.1 Scholarly Document Processing

With the increasing volume of scientific publications, the growing need of computational methods for enhancing applications such as summarization, search, and analysis of scientific documents to serve human researchers has fostered the advances in the field of scholarly document processing. On the other hand, driven by the adoptions of neural network-based models, recent methods in Natural Language Processing (NLP) often require large amounts of supervised data. To this end, several online resources with sheer amounts of scientific publications as well as other useful information like meta-data have been utilized to construct large-scale corpora suitable for downstream tasks. To name a few, papers from academic publishers and literature archives such as arXiv<sup>1</sup>, PubMed<sup>2</sup>, ACL Anthology<sup>3</sup>, Semantic Scholar<sup>4</sup> and Emerald<sup>5</sup> have been collected to develop datasets for paper summarization [1, 2, 11–13]. Online platforms for paper reviews and rebuttals like OpenReview<sup>6</sup> are used to derive corpus for review generation and argument pair extraction with further annotations [14–19]. Other resources such as Microsoft

---

<sup>1</sup><https://arxiv.org>

<sup>2</sup><https://www.ncbi.nlm.nih.gov/pmc>

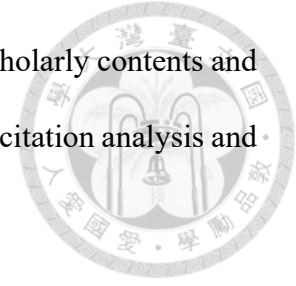
<sup>3</sup><https://www.aclweb.org/anthology>

<sup>4</sup><https://www.semanticscholar.org/>

<sup>5</sup><https://www.emerald.com>

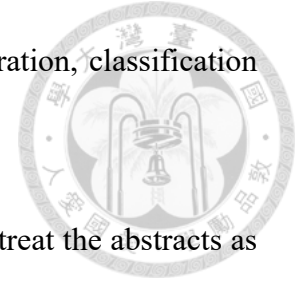
<sup>6</sup><https://openreview.net>

Academic Graph [20] and the Semantic Scholar offer fine-grained scholarly contents and enhanced contextual search results for various applications related to citation analysis and knowledge graph construction [21, 22].



In this thesis, we focus on a rather understudied element as paper summarization targets – research contributions in scientific documents. Recently, D’Souza *et al.*[4] proposed a pipeline to automatically construct knowledge graphs from NLP papers: sentences that describe the contributions of the paper are first extracted from the full text, then scientific knowledge terms and predicates in the extracted sentences need to be identified, finally, the system organizes these entities as triples to build the knowledge graph. They defined an annotation scheme under which contributions are classified into 10 categories and they annotated 442 papers in NLP domains. Similar to their work, Chen *et al.*[6] introduced a dataset with 5K sentences describing research contributions manually collected from ACL Anthology and IP&M, as well as their fine-grained annotations with six categories of contribution types. For generation tasks, Hayashi *et al.*[5] proposed to generate summaries discussing the contributions and the contexts of the papers separately. To tackle this problem, they manually labeled abstracts of 400 papers from the S2ORC corpus [12] with binary labels indicating whether a sentence is contribution-related or context-related. These gold reference summaries are then used to finetune a sentence classifier which is later applied to automatically generate reference labels for all other papers in the corpus. He *et al.*[23] also explored the task of contribution generation for scientific papers to evaluate their controllable summarization framework through zero-shot experiments. Different from the above works, we are able to construct a large-scale dataset by applying automatic methods to extract author-written contributions from computer science papers. In addition, we design a new annotation scheme for contribution type classification which is less

challenging yet reasonable. Overall, our dataset can facilitate generation, classification and other downstream tasks related to research contributions.



In addition to traditional scientific summarization datasets that treat the abstracts as the generation targets, there are several works exploring other forms of reference summaries for better application values. Cachola *et al.*[24] presented the task of TLDR generation and the associated dataset, they aimed to summarize the scientific papers in extremely short texts that highlight the key aspects concisely. Meng *et al.*[13] extended the idea of facet summarization to scientific domains, they collected papers from Emerald Publishing where summaries of the papers from four aspects – purpose, method, finding and value are directly available. Collins *et al.*[25] introduced a extractive summarization dataset CSPubSum consisting of 10K computer science papers from ScienceDirect, their generation targets are author-written highlight statements of the papers. Gidiotis *et al.*[26] proposed the task of structured summarization for scientific papers by applying a divide-and-conquer approach to generate parts of the abstracts describing certain aspects of the papers based on section matching. Similarly, Liu *et al.*[27] presented the dataset and methodology for generating structured summaries for groups of related academic documents to serve as the role of overviews or survey papers. Among all, our dataset resembles to CSPubSum as both of our generation targets provide highlights of the scientific papers. The main differences are that we focus on abstractive summarization and our reference summaries are formed by disentangled contributions of different types based on our annotations while theirs are plain texts that can be only treated as a whole. This characteristic also indicates that our dataset shares some common points with facet summarization, in which models are trained to summarize the targets from different aspects. Inspired by the facet-aware evaluation of extractive models[28], we develop

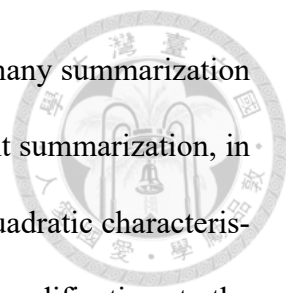
contribution-level evaluations and analysis to study the model performances in terms of each disentangled contribution.



## 2.2 Abstractive Summarization

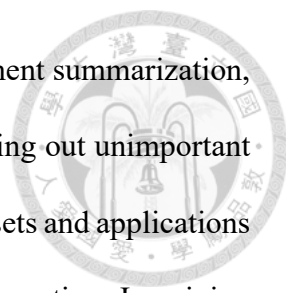
Our task of generating disentangled contributions for scientific documents is closely related to abstractive summarization. Abstractive summarization aims to condense the source documents into concise summaries. Compared with extractive summarization, it requires the model to organize important information from the source document and produce coherent texts in novel wordings instead of copying.

Neural network-based methods for abstractive summarization formulate the task as a sequence-to-sequence problem. The encoder takes the source document as input and produce its representation through computation. The decoder then outputs a token distribution for each time step autoregressively, conditioned on both the input document representation and previous generation results. Recently, transformer-based models[8] have showcased its performance superiority over other model architectures in generation tasks with the ability of effectively capturing and encoding dependency across contexts by leveraging the powerful attention mechanism [29–32]. Abstractive summarization is no exception. Lewis *et al.*[9] proposed BART, a seq2seq model pretrained with a denoising objective where input texts are corrupted based on special designs and the model is trained to reconstruct the original documents. Zhang *et al.*[7] presented a self-supervised pretraining strategy called Gap Sentences Generation (GSG) for abstractive summarization. They construct pseudo-summaries for large unsupervised corpus by selecting salient sentences that maximize the ROUGE score between themselves and the remaining of the document.



Both of these two pretrained models have achieved great results on many summarization datasets and became strong baselines in this field. For long document summarization, in order to address the issue of computational overload caused by the quadratic characteristic of self-attention mechanism, recent works have proposed various modifications to the vanilla transformer architecture. Kitaev *et al.*[33] reduced the computational complexity by using locality-sensitive hash(LSH) to compute nearest neighbors as replacements of the full self-attention. Wang *et al.* [34] approximate the self-attention mechanism by a low-rank matrix and reduce the time and space complexity to linear forms. In addition, Longformer-Encoder-Decoder (LED) [10] and BigBird [35] are two popular long document summarizers with a combination of global and local attentions as well as model weights initialized from BART and PEGASUS respectively. PRIMERA [36], on the other hand, is a multi-document summarization model built on LED that also achieves state-of-the-art performance in scientific document summarization. It is trained with a post-training strategy named Entity Pyramid Masking which selects sentences based on entity importance across multiple documents and self-ROUGE scores as pseudo summaries.

Recent approaches for abstractive summarization also explore various guidance signals to either control the generation results or improve the model performances. Dou *et al.*[37] developed a guided summarization framework that incorporates additional information with the input, including highlight sentences, keywords, triples and relevant summaries in the training set. He *et al.*[23] also proposed a generation framework that can achieve entity-centric and length-controllable summarization mainly through manipulating additional input keywords. Similarly, Narayan *et al.*[38] utilized entity chains in output targets to improve summarization performances as well as faithfulness. Mao *et al.*[39] explored the integration of keywords in the constrained decoding stage to im-



prove the factual consistency of the generation results. In long document summarization, content selection is utilized to explicitly reduce input length by filtering out unimportant texts [40]. In addition to these commonly-used guidance signals, datasets and applications in special domains also inspire the exploitation of other additional information. In opinion summarization, aspect queries are used to control the summarization of opinions towards certain targets [41]. In meta-review generation, Shen *et al.* [19] defined 9 categories of intent roles for sentences in meta reviews and used them as controllers to guide the generation process. Our work also leverages contribution types of the target contributions as guidance signals, in addition, we add contribution types to highlight sentences and the section headers in the source papers to provide the model with the alignments between the input documents and the output targets.





## Chapter 3 Datasets

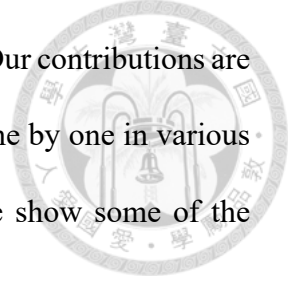
In this chapter, we introduce the newly-proposed dataset in our work. Section 3.1 describes the details of constructing our dataset. Section 3.2 presents statistics of our dataset and comparisons with other summarization datasets in scientific domains. Section 3.3 introduces our further annotation with the extracted contributions.

### 3.1 Dataset Collection

To develop models capable of generating disentangled contributions for scientific documents, a desired dataset should consist of papers and corresponding contributions summarizing their researches into several keypoints. However, due to the lack of automated methods in extracting contributions from scientific papers, existing datasets related to our work are either at limited scales or lacks disentangled contributions.

The method of constructing our dataset is inspired by a trending writing style in recent scientific publications. We observe that more and more authors start to explicitly list their contributions in the papers, especially in AI-related fields. According to our pilot study of 100 papers from top conferences in computer vision, computational linguistic, machine learning and artificial intelligence such as CVPR, ACL, ICLR and NeurIPS, 65 of them contain the contributions stated by the authors. Moreover, most contributions locate in the

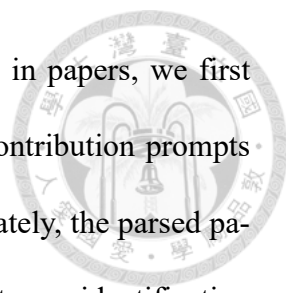
end of introduction sections. They typically start with a prompt like "Our contributions are summarized as follow:". Then the authors state their contributions one by one in various formats such as bullet point lists, hyphen lists and link words. We show some of the common patterns in Figure 3.1.



Type	Example
Arabic Numerals	<b>Contributions. This study makes the following contributions:</b> (1) We formally analyze the imperceptibility of arithmetic coding based steganography algorithms; (2) We propose SAAC, a new nearimperceptible linguistic steganography method that encodes secret messages using self-adjusting arithmetic coding with a neural LM; and (3) Extensive experiments on four datasets demonstrate our approach can on average outperform the previous state-of-the-art method by 15.3% and 38.9% in terms of bits/word and KL metrics, respectively.
Roman Numerals	<b>Contributions: This work makes the following contributions:</b> (i) We propose a unified framework DiscProReco to jointly perform CDP and DPR, and show that these two tasks can benefit each other. (ii) We construct a new large-scale dataset SPDPR (Section 4) which supports fair comparison across different methods and facilitates future research on both DPR and CDP. (iii) We present experimental results which show that DiscProReco with its joint learning mechanism realizes knowledge sharing between its CDP and DPR components and results in improvements for both tasks (Section 5).
Link Words	<b>Our contributions are summarized as follows:</b> Firstly, we develop an RL agent featured with question-guided task decomposition and action space reduction. Secondly, we design a two-phase framework to efficiently train the agent with limited data. Thirdly, we empirically validate our method' s effectiveness and robustness in complex games.
Other Symbols	<b>Our contributions include:</b> • A multi-source label aggregator CHMM with token-wise transition and emission probabilities for aggregating multiple sets of NER labels from different weak labeling sources. • An alternate-training method CHMM-ALT that trains CHMM and BERT-NER in turn utilizing each other' s outputs for multiple loops to optimize the multi-source weakly supervised NER performance. • A comprehensive evaluation on four NER benchmarks from different domains demonstrates that CHMM-ALT achieves a 4.83 average F1 score improvement over the strongest baseline models.

Figure 3.1: Examples of common patterns in which authors list their contributions explicitly

Another important issue is the collection of machine readable texts from scientific documents. Our initial attempt is to directly extract contributions from the papers in arXiv dataset [1] and the S2ORC corpus [12], both of which are well-established and widely-used resources in scientific document processing since they provide large amounts of structured full text parse. However, we find that it is unsuitable to construct our dataset



building on them. The main reason is that, to extract contributions in papers, we first identify introduction sections, then search for the aforementioned contribution prompts and specific patterns following disentangled contributions. Unfortunately, the parsed papers in these two datasets do not cater to our need in this task. For instance, identification of section headers is too noisy to effectively locate introduction sections on a large scale, not to mention symbols such as bullet points and hyphens need additional post-processing to retain. Another reason is that since our extraction rules are based on observation on papers in computer science domains, they may not adapt to papers in other domains which take up a large proportion in arXiv and S2ORC. During manual inspection of our trial experiment on S2ORC, the extracted results are quite noisy for non-CS papers as their writing styles are somewhat different. Our speculation over this issue is also advocated by a previous work[23], where the authors were only able to extract 1,018 papers with their contributions out of 67K papers from the arXiv database.

To this end, we decide to parse papers on our own instead of resorting to existing resources. We focus on papers in CS-related fields and leave the expansion of a multi-domain scientific dataset to future works. The first step is to download  $\LaTeX$ source files from arXiv using arXiv API as the availability of  $\LaTeX$ sources enables us to extract structured and high quality body texts compared with using PDF files. Besides, we are able to specifically target papers in computer science categories through the API. In detail, we query papers in the following domains: cs.AI, cs.LG, cs.CV, cs.LR, cs.IR, cs.SI, cs.DL, cs.RO, and cs.HC. For more information about the definition of these categories, please refer to the arXiv website<sup>1</sup>. Following previous works[42], we then convert  $\LaTeX$ sources into XML files and then extract structured information from them. Notably, several cus-

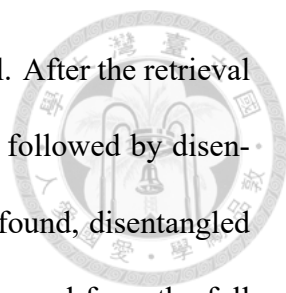
---

<sup>1</sup>[https://arxiv.org/category\\_taxonomy](https://arxiv.org/category_taxonomy)

tomized cleaning rules are implemented to obtain high-quality paper texts, we list the main ones as follow:



- Tables, figures and their captions are all removed since they are not important in our task and it is extremely hard to arrange them in the plain text with reasonable positions. Complicated formulas are also removed except for those containing only numbers.
- Texts are retrieved while maintaining sections and subsections (only first order) as we explicitly target introduction sections in the papers to extract contributions, also, these structures serve as important roles in analyzing the discourse of the scientific literature.
- Different from other works in which the parsed papers are used for citation-related tasks so that reference resolution is necessary, we simply replace citation spans with a special marker [REF]. Based on our hypothesized application scenario where the desired summarization systems help readers quickly consume latest works, there is supposed to be very few inbound citations. On the other hand, outbound citations are also fairly irrelevant to our task since our goal is to summarize contributions eschewing related works and contexts.
- We explicitly process texts in list forms of  $\text{L}^{\text{T}}\text{E}^{\text{X}}$  syntax such as items and enumerates because we find that many authors list their contributions in this way.
- Contents after conclusions such as appendixes and acknowledgements are omitted, we also filter out parsed papers with outlying length or missing all of abstract, introduction and conclusion sections.



Following these steps, we are able to collect 110K papers in total. After the retrieval of full texts, we target contribution prompts in introduction sections followed by disentangled contributions based on the previously discussed patterns. If found, disentangled contributions are extracted as our target summaries and they are removed from the full texts. Meanwhile, papers that we are unable to find contributions in are also retained as unlabelled data, which will be leveraged to construct pseudo-summaries in the post-training stage. Last but not least, we further clean the extracted contributions by filtering out noisy ones based on hand-crafted rules such as length limits, websites and the existence of non-readable symbols.

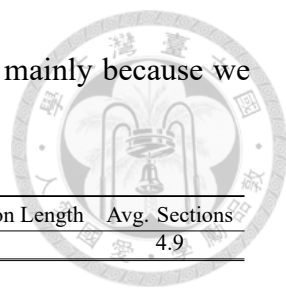
Overall, the above procedures yield 24K papers and their extracted contributions composing our final dataset. The ratio in our dataset between papers that explicitly state their contributions and total papers are significantly lower than that in the pilot study. We speculate that our rather strict cleaning rule is one of the cause, as the ratio before and after final cleaning is 30% and 22% respectively. Also, papers submitted to top conferences might have a higher tendency to list their contributions for benefiting the peer review process.

## **3.2 Dataset Analysis**

### **3.2.1 Dataset Statistics**

Table 3.1 shows the basic statistics of our dataset. On average, each paper in our dataset has 3.2 contributions with a length of 29. This roughly coincides with our pilot study that most papers write 3 or 4 short contributions. Note that our dataset has smaller

paper lengths and section numbers compared with the arXiv dataset mainly because we omit extremely long papers and remove tables as well as figures.



# Papers	Avg. Contribution	Avg. Paper Length	Avg. Summary Length	Avg. Contribution Length	Avg. Sections
24130	3.2	3632	91	29	4.9

Table 3.1: Basic statistics of our dataset

### 3.2.2 Comparisons with Existing Datasets

Since our task is closely related to scientific paper summarization, we provide comparisons between ContributionSum and the following datasets.

**arXiv** [1] is a benchmark dataset for scientific paper summarization and long document summarization.

**SCITLDR** [24] is a extreme summarization dataset containing both author-written and expert-derived TLDRs for scientific papers

**FacetSum** [13] is a facet summarization benchmark consisting of articles from Emerald journals, they provide paper summaries from four aspects: purpose, method, finding and value, we report the statistics in value aspect as it resembles to our task based on their definition.

**DisentangledSum** [5] is extended from the S2ORC corpus by adding contribution-related and context-related reference labels to sentences in abstracts, we report the statistics of contribution-related sentences in their work.

Following previous work[24], we analyze the text length, compression ratio (source length divided by target length) and percentage of novel words (words appear in the reference yet not in the source) of each dataset. As shown in Table 3.2, the summary length

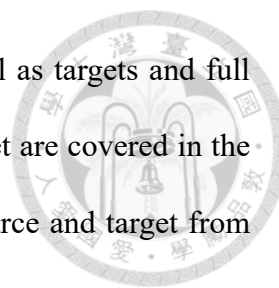
and the compression ratio of ContributionSum lies between arXiv and other scientific summarization dataset, suggesting that our dataset provide an alternative for the community other than abstract or extreme summarization. It is worth mentioning that although both DisentangledSum and our dataset aim to generate contributions for scientific papers, their average summary length is longer since they perform sentence-level extraction in abstract to obtain the target and their dataset is mainly composed of silver summaries that are likely to introduce noise, while ours consists of author-written contributions which are supposed to be more abstractive.

Dataset	# Papers	Avg. Paper Length	Avg. Summary Length	Compression Ratio	% Novel Words
arXiv	215K	4.9K	220	22.5	8.3
SCITLDR	3.2K	5K	21	238.1	15.2
FacetSum	60K	6.8K	47	144.7	NA
DisentangledSum	400	6.3K	136	46.3	NA
Ours	24K	3.6K	91	39.6	9.9

Table 3.2: Comparisons of ContributionSum with existing datasets, some statistics are not available due to the absence of open-access datasets

### 3.2.3 Structural Alignments

Structure information is important in long document processing, especially for scientific papers where the discourse is well-organized into sections as conventions. Among all the possible sections, previous studies have found that the most salient information in a paper for writing a summary is often found in the abstract, introduction, and conclusion sections [43]. To analyze the importance of these three sections in our task, we first leverage ROUGE-score[44] to study the lexical overlaps between the target contributions and the texts in the abstract, introduction and conclusion sections. Specifically, we randomly sample 1,000 papers that possess all three sections from our dataset, then calculate



ROUGE-recall and ROUGE-f1 between targets and sections as well as targets and full texts. A high ROUGE-recall indicate that more N-grams in the target are covered in the source, and a high ROUGE-f1 suggests high similarity between source and target from the lexical perspective.

As presented in Table 3.3, we can see that abstracts, introductions and conclusions all have significantly more overlapped N-grams with the target summary than other sections based on the results of ROUGE-recall. Also, there is no surprise that they have much higher ROUGE-f1 since they are already highly-summarized texts compared with other sections. Among these three, abstract and conclusion are much shorter and closer to our targets while introduction might elaborate more on the backgrounds, motivations and technical details of the papers.

Section	Source Length	ROUGE-1/2 recall	ROUGE-1/2 f1
Abstract	172	65.17/26.42	43.01/17.31
Introduction	613	76.1/29.97	21/7.8
Conclusion	183	58.27/22.18	41.61/16.02
Others	681	54/17.91	15.01/4.73
All (full paper)	3579	83.71/59.97	5.75/3.46

Table 3.3: ROUGE scores between targets and sections in our dataset

We further investigate the sentence-level alignments in these sections. We greedily select sentences in the paper that maximize the sum of ROUGE-1 and ROUGE-2 f1 scores with each disentangled contribution in the references. We plot the relative positions of these selected sentences in their corresponding sections and present the result of abstract, introduction, conclusion and other sections in Figure 3.2. For selected sentences in abstracts and introductions, they tend to be positioned towards the end, since authors might discuss the backgrounds and the motivations first when writing these two sections. On the other hand, those in conclusions skew towards the beginnings, mostly because au-



thors often quickly summarize their work first in conclusions and then state future works later. In other sections, there is no clear pattern observed except for the potential lead bias.

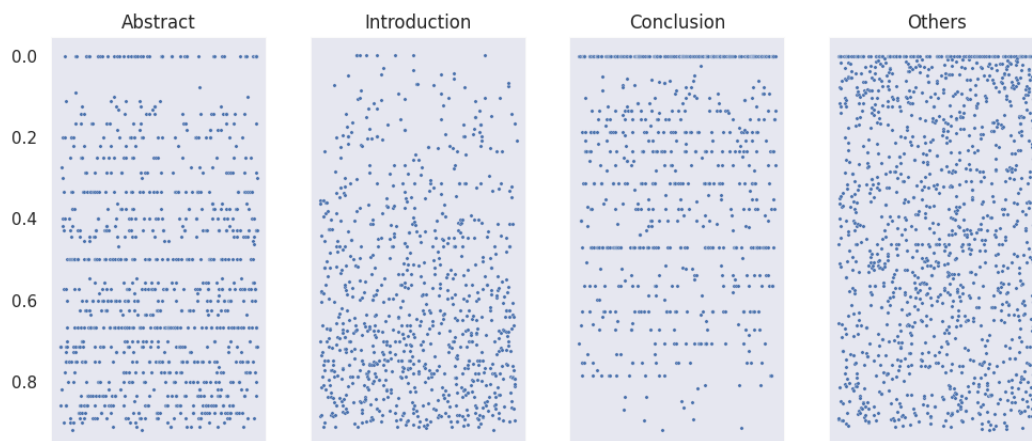


Figure 3.2: Relative positions in the associated sections of greedily extracted sentences that maximize the ROUGE scores with the target contributions

### 3.3 Contribution Type Annotation

In this section, we describe our further annotation of contributions types in our dataset. Although our main task is to generate disentangled contributions, which does not necessarily requires contribution types, we argue that our annotation is still valuable. First, recent works have studied the topic of contribution extraction and contribution type classification, yet due to the huge cost of manual annotation in scientific domain, they are unable to construct large-scale datasets. On the other hand, our dataset provide significant amounts of disentangled contributions, should they be paired with their corresponding contribution types, it would benefit the aforementioned tasks a lot. Second, if we are able to generate disentangled contributions with their contribution types, our generation results can serve as valuable resources for downstream tasks such as scientific entity extraction and knowledge graph construction. Last but not least, the presence of contribution types can provide sketch supervisions to our main task, guiding the generation process to focus

on certain types of contributions as well as aligning with highlight sentences in the source documents. We will elaborate more on this point in Chapter 4.



### 3.3.1 Annotation Scheme

Previous works in contribution type classification have already proposed annotation schemes for this task. However, their annotation schemes are too challenging to develop automatic methods [4, 6]. As we want to apply a data driven approach to train a contribution type classifier for automatic annotation of our full dataset, their designs are not suitable for our work. To this end, we present a new annotation scheme tailored to our dataset. We classify the contributions into the following four categories, detailed explanations and examples are showed in Figure 3.3:

**Approach and Method:** Proposal of new methodology to an existing research problem, since our dataset focuses on papers in computer science domain, this mostly refers to models, systems, frameworks, algorithms and strategies.

**Theory, Analysis and Finding:** Detailed theoretical or empirical analysis of existing works as motivation for future improvements or the proposed methodology such as ablation studies.

**Experiment Result:** Evaluation of the proposed methodology, frequently accompanied with comparisons with existing works.

**New Topic or New Resource:** Creation of new research topic, new task, or new datasets.

Contribution Type	Percentage in Our Annotation	Common Keywords	Explanation	Example
Approach and Method	48%	propose, model, novel, framework, method, learning, new, training, introduce	Models, systems, algorithms and frameworks tackling existing problems	We design aspect and opinion propagation decoder so that the model has a comprehensive understanding of the whole context, and thus it results in better prediction of the polarity.
			Detailed components such as training strategies built on existing frameworks	To accomplish the contextual alignment, we design the Directional Contrastive Loss, which applies the contrastive learning in a pixel-wise manner. Also, two effective sampling strategies are proposed to further improve performance.
Theory, Analysis and Finding	20.7%	show, performance, different, analysis, study, provide	Analysis and deriving of theory based on existing works, possibly serve as motivation for future improvement	Through in-depth analysis, we point out that different orders of branch expansion are suitable for handling different multi-branch AST nodes, and thus dynamic selection of branch expansion orders has the potential to improve conventional Seq2Tree models.
			Analysis of the proposed methodology or resource	We show that the proposed model can capture long-distance interactions between entities. Our further analysis statistically demonstrates the proposed gating mechanism is able to aggregate the structured information selectively.
Experiment Result	22.8%	state-of-the-art, show, performance, result, proposed, dataset, experiment, demonstrate, outperform	Evaluation of the proposed methodology, frequently accompanied with comparisons with existing works	Extensive experiments show that our approach significantly outperforms previous state-of-the-art models on both MSMARCO and Natural Questions datasets.
New Topic or New Resource	8.5%	dataset, task, available, release, research, website, code, benchmark, publicly	Creation of new research topic	We introduce a sequence-level self-supervised task called Replaced Mask Detection to distinguish between different transformations applied to a text.
			Creation of new resource	To the best of our knowledge this is the first work which utilizes meta pages, i.e., talk pages as an additional signal for this task. All code, sample data and image embeddings related to the paper are made available to promote reproducible research.

Figure 3.3: Illustrations of our contribution type annotation scheme

### 3.3.2 Annotation Procedure and Results

We manually annotate 1K papers randomly sampled from our dataset. This results in a total of 3,798 contributions. During the annotation process, we found that around 10% of the contributions (339/3798) contains more than one type of contribution based on our annotation scheme. Since the percentage is not very high, we omit these contributions and leave the extension of multi-label contribution classification to future works. To obtain contribution type labels for all of our dataset, we then finetune a sentence classifier using our annotated data and utilize it to generate reference labels for other papers in our dataset.



### 3.3.3 Analysis of Contribution Types

While the contributions types are labeled for each disentangled contribution, they are not independent to each other. For instance, as most people would expect, contributions about experiment results are very likely to appear after those describing the methods. To study the relations between the contribution types, we present their transition matrix in Figure 3.4.

Among all, contributions talking about new methods are the most likely to locate at the beginning while experiment results are often positioned at the end. Notably, though frequently followed by experiment results as expected, we observe that contributions about approach and method can also appear in segments where authors divide their method into components and list them separately. This also happened to contributions about analysis and findings.

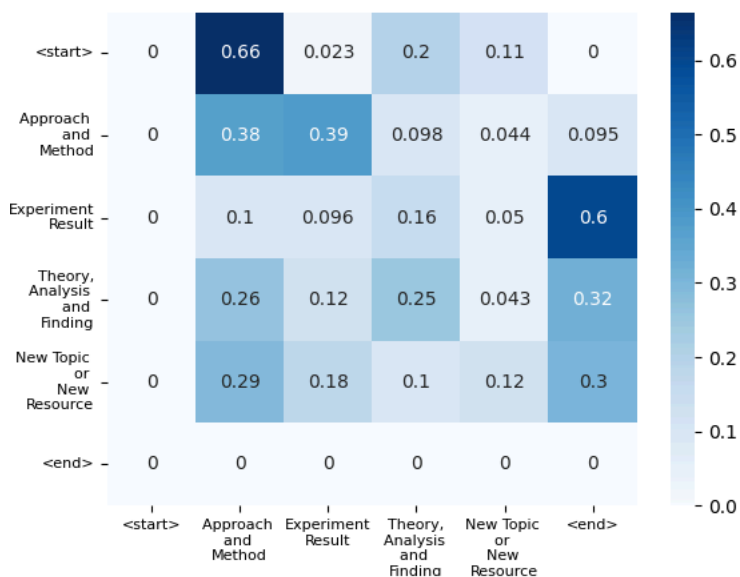
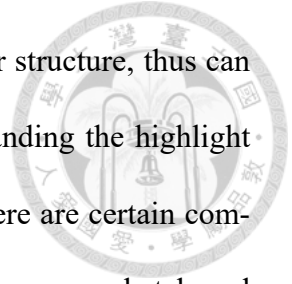


Figure 3.4: Transition matrix of different contribution types

During our annotation, we also observe some common patterns of papers in listing their contributions as showed in Figure 3.5. This phenomenon suggests that the combi-

nation of disentangled contributions is closely aligned with the paper structure, thus can provide a more comprehensive summary for the readers in understanding the highlight and story-line throughout the research process. In addition, since there are certain common templates, the presence of contribution types can serve as a summary sketch and further guides the summarization process.



Pattern	Example
Approach and Method – Experiment Result	<p>First, we propose Hierarchical Evidence Set Modeling, which consists of document retriever, multi-hop evidence retriever, and claim verification.</p> <p>Second, our multi-hop evidence retriever retrieves evidence sentences and combines them as evidence sets. Our claim verification component conducts the hierarchical verification based on each evidence set individually and then based on all the evidence sets.</p> <p>Finally, our experimental results show that our model outperforms 7 state-of-the-art baselines in both the evidence retrieval and claim verification.</p>
Approach and Method – Experiment Result – Theory, Analysis and Finding	<p>First, We propose a simple and robust Syn-LSTM model to better incorporate the structured information conveyed by dependency trees. The output of the Syn-LSTM cell is jointly determined by both contextual and structured information. We adopt the classic conditional random fields (CRF) (Lafferty et al., 2001) on top of the Syn-LSTM for NER.</p> <p>Second, we conduct extensive experiments on several standard datasets across four languages. The proposed model significantly outperforms previous approaches on these datasets.</p> <p>Finally, we show that the proposed model can capture long-distance interactions between entities. Our further analysis statistically demonstrates the proposed gating mechanism is able to aggregate the structured information selectively</p>
Theory, Analysis and Finding – Approach and Method – Experiment Result	<p>First, we analyze partially correct predictions of a SOTA English reader model, revealing a distribution over three broad categories of errors.</p> <p>Second, we show that an Answer Corrector model can be trained to correct errors in all three categories given the question and the original prediction in context.</p> <p>Finally, we further show that our approach generalizes to other languages: our proposed answer corrector yields statistically significant improvements over strong RoBERTa and Multilingual BERT (mBERT) (Devlin et al., 2019) baselines on both monolingual and multilingual benchmarks.</p>
Only Theory, Analysis and Finding	<p>First, we examine the assumptions Baker et al. use to operationalize economic policy uncertainty via keyword-matching of newspaper articles. We demonstrate that using keywords collapses some rich linguistic phenomena such as semantic uncertainty.</p> <p>Second, we also examine the causal assumptions of Baker et al. through the lens of structural causal models (Pearl, 2009) and argue that readers' perceptions of economic policy uncertainty may be important to capture.</p> <p>Third, we conduct an annotation experiment by re-annotating documents from Baker et al.. We find preliminary evidence that disagreements in annotation could be attributed to inherent ambiguity in the language that expresses EPU.</p> <p>Finally, we replicate and extend Baker et al.'s data pipeline with numerous measurement sensitivity extensions: filtering to US-only news, keyword-matching versus supervised document classifiers, and prevalence estimation approaches. We demonstrate that a measure of external predictive validity, i.e., correlations with a stock-market volatility index (VIX), is particularly sensitive to these decisions.</p>

Figure 3.5: Examples of common contribution patterns



## Chapter 4 Methodology

In this chapter, we introduce our overall framework for generating disentangled contributions. Our method can be divided into three parts: contribution type classification, fine-tuning and post-training for disentangled contribution generation. In Section 4.1 we describe how we leverage a pretrained language model to finetune a contribution type classifier which will later be utilized in our generation model. In Section 4.2 we present our generation model built on existing encoder-decoder architectures and a fine-grained sentence masking strategy tailored to our task for the post-training stage so that we can exploit self-supervised learning methods.

### 4.1 Contribution Type Classification

For contribution type classification, we fine-tune SciBERT [45] on our annotated classification dataset which contains 3.3K contributions. SciBERT is a language model pretrained on a large corpus of 1.14M scientific papers. While following the same architecture as BERT[46] which is a transformer encoder, it uses a different vocabulary constructed on their own corpus. To train a text classifier leveraging SciBERT, we take the final hidden state  $h$  of the first token [CLS] as the representation of the whole sequence

and add a feedforward layer on top of it for classification:

$$y = \text{softmax}(W * h_{[CLS]} + b) \quad (4.1)$$



where  $h_{[CLS]}$  is the final hidden state of the [CLS] token,  $W \in \mathbb{R}^{C*d}$  and  $b \in \mathbb{R}^C$  are trainable parameters of the additional classification layer,  $C$  is the number of class which is four in our work and  $d$  is the dimension of final output of SciBERT. The model is then optimized with a cross-entropy loss.

After training, we apply the model to predict the contribution type labels for the target contributions in all other papers without annotation in our dataset, which we will use in the following generation model.

## 4.2 Disentangled Contribution Generation

### 4.2.1 Task Formulation and Model Architecture

We formulate our task of generating disentangled contributions for scientific documents as an abstractive summarization problem. The goal is to develop a model that takes a scientific document as input and generate its disentangled contributions sequentially as output:

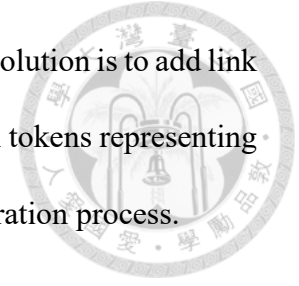
$$\begin{aligned} S &\leftarrow g(D) \\ S &= \text{concat}(C_{1:n}) \end{aligned} \quad (4.2)$$

where  $g$  is the mapping inferred from the model,  $D$  is the input document,  $C_{1:n}$  represents  $n$  disentangled contributions and the concatenation of them forms the desired output  $S$ .

Note that in order to achieve disentangled generation, the concatenation operation should



add additional tokens to separate the contributions. The most simple solution is to add link words such as first, second and finally. In our method, we use special tokens representing different contribution types to serve as guidance signals for the generation process.



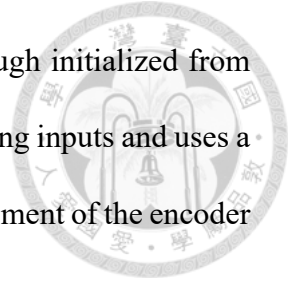
Recent advances in neural abstractive summarization mostly adopts a sequence-to-sequence architecture to generate summaries. The encoder takes the document  $D$  as input and produces the input representation  $\mathbf{Z}$ . The decoder then outputs a token distribution  $p(y_t)$  for each time step  $t$ , conditioned on both  $\mathbf{Z}$  and previous generation results  $y_{1:t-1}$ . Finally, the model is trained using a cross-entropy loss to minimize the negative log-likelihood of the tokens in the reference summary  $\hat{y}$  in a auto-regressive way using teacher forcing:

$$\begin{aligned}\mathbf{Z} &= \text{Encoder}(D) \\ p(y_t) &= \text{Decoder}(y_{1:t-1}, \mathbf{Z}) \\ \mathcal{L}_{gen} &= -\sum_t \hat{y}_t \log p(y_t)\end{aligned}\tag{4.3}$$

Among all model architectures, transformer-based approaches have achieved state-of-the-art performances in many summarization tasks leveraging the power of attention mechanisms. In a vanilla transformer model, the encoder uses self-attention to capture the relations between input tokens and enrich their representations. On the decoder side, in addition to self-attention, cross-attention is utilized to draw global dependencies between input and output.

To leverage powerful pretrained transformer models, in this thesis, we use BART and Longformer-Encoder-Decoder (LED) as our underlying architecture. BART is pre-trained with a denoising objective where additional noise such as shuffling and masking of spans of texts are imposed and the model learns to reconstruct the original text. On the other

hand, LED is designed for generation task of long documents. Though initialized from the parameters of BART, it extends the positional embeddings to fit long inputs and uses a combination of sparse local and global attention mechanism in replacement of the encoder self-attention for improving computational efficiency.

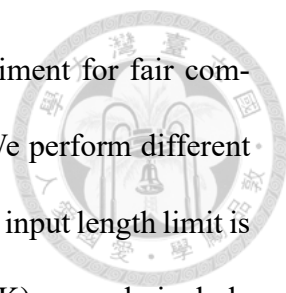


#### 4.2.2 Finetune for Disentangled Contribution Generation

Built on existing encoder-decoder architectures, we focus on organizing input documents and output targets to finetune for disentangled contribution generation. In order to provide guidance signals and structure alignments for the generation process, we create special tokens representing different contribution types in both input documents and output targets. Recall that we define four types of contributions that appear in our dataset: Approach and Method, Theory, Analysis and Finding, Experiment Result and New Topic or New Resource. We add a new token to the tokenizer representing each contribution type. In the remaining part of this thesis, we abbreviated them to <Methodology>, <Analysis>, <Result> and <Resource> respectively.

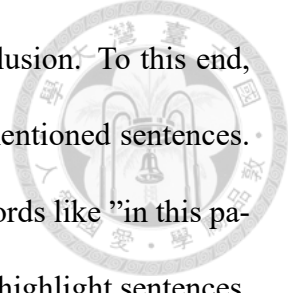
Furthermore, following previous works [36, 47], we add a special token <Doc-Sep> to separate the sections, adding structural signals of the paper to our model. We also assign global attentions to these tokens which the model can use to share information across different sections. The <Doc-Sep> token is randomly initialized while contribution type tokens are initialized as the average of the token embeddings of the common keywords in each contribution type.

To organize the scientific papers for model input, we first perform truncation to fit in the input length limit of the model architectures. BART has an input length limit of



1024, while LED can scale up to 16384, we use 4096 in our experiment for fair comparisons with previous works as well as computational efficiency. We perform different truncation strategies for two model architectures. For BART, since the input length limit is significantly smaller than the average paper length in our dataset (3.6K), we only include title, abstract, introduction and conclusion sections of the papers in the input sequences based on our previous analysis that indicates the importance of these sections in generating research contributions. Besides, results from previous works [13, 24] in scientific paper summarization have proven that treating AIC (abstract+introduction+conclusion) as the model input space can substantially reduce computational costs without sacrificing performances. For LED, we retain title, abstract, introduction and conclusion sections as in BART. Moreover, for every other section, we allocate input length for it based on the remaining input length limit (original input length limit minus length of title and AIC) multiplied by the ratio of current section length and total length of the remaining sections. Analogously, if subsection structures exist in a section, we allocate length limit for each subsection and perform truncation based on our allocation.

In addition to the truncation strategy, we add special tokens to highlight important sentences and inform the model of high-level structural information of the paper discourse. Our highlight sentences extraction algorithm is based on heuristic rules from our observation and the analysis of the dataset. Specifically, we first target sentences in abstract, introduction and conclusion sections as they are of primary importance compared with other sections. Then we use a heuristic rule to filter out sentences in these sections that are not likely to discuss the contributions of the paper. Based on our analysis in section 3.2 and manual inspections, sentences at the beginning of abstract and introduction usually describes the background and the related works, whereas authors might list some



directions for future works or acknowledgements at the end of conclusion. To this end, we use a simple keyword matching approach to filter out the aforementioned sentences. For abstract and introduction, we search for a predefined set of keywords like "in this paper" as starting signals and add all sentences after it as our candidate highlight sentences. Similarly, for conclusion section we add sentences from the beginning sequentially to the candidate list until keywords like "future work" are detected. If no keyword is matched, we use a length threshold to filter out sentences instead.

After the filtering stage, we then feed each candidate sentence into our contribution type classification model described in section 4.1. Sentences that have prediction scores above a threshold  $\tau$  is selected as the highlight sentences since our classification model is more confident that they are discussing certain type of contribution. Algorithm 1 shows the overall procedure of highlight sentence extraction. The selected highlight sentences will be prepended with the special token representing their contribution types.

---

**Algorithm 1:** Highlight Sentence Extraction

---

```

Input : Sentence sets for abstract  $S_a$ , introduction  $S_i$  and conclusion  $S_c$ 
Input : A contribution type classification model  $g$ 
Input : Predefined keyword sets  $K_s$  and  $K_e$ 
Input : Length threshold  $l_s$  and  $l_e$ , prediction score threshold  $\tau$ 
Output: Sets of highlight sentences  $H$ 
1  $H = []$ 
2 for section in [a,i,c] do
3   start,end=True, False
4   if section in [a,i] then
5     start=False
6   for idx,sentence in  $S_{section}$  do
7     if not start and (any( $K_s$  in sentence) or  $idx/\text{len}(S_{section}) > l_s$ ) then
8       start=True
9     if section is c and not end and (any( $K_e$  in sentence) or  $idx/\text{len}(S_{section}) > l_e$ ) then
10      end=True
11    if start and not end and  $\text{predict\_score}_g(\text{sentence}) > \tau$  then
12       $H.append(\text{sentence})$ 

```

---

In addition to highlight sentences, we also match each section to certain contribution

type and add corresponding special tokens in front of the section header. Similar to existing work [48], the matching is done by heuristic keyword mappings of some common keywords in the section header. Finally, we concatenate all the sections based on their original sequences after the truncation and the additions of special tokens.

To incorporate contribution types in the output target, we simply add the corresponding special token in front of each disentangled contribution based on their corresponding contribution type. One example of our model input and output is showed in Figure 4.1. In this example, the special tokens serve as additional guidance to the generation process and they provide clear alignments between each of the reference target and their closely-related highlight sentences in the source document extracted based on our algorithm.



---

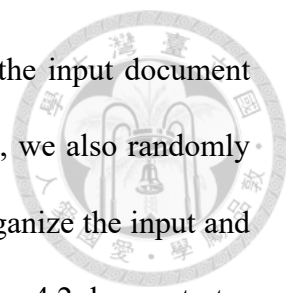
<b>Input</b>	<p>&lt;DOC-SEP&gt; title: VBridge: Connecting the Dots Between Features, Explanations, and Data for Healthcare Models</p> <p>&lt;DOC-SEP&gt; abstract: ..... &lt;METHODOLOGY&gt; Following an iterative design process, we further designed and developed VBridge, a visual analytics tool that seamlessly incorporates ML explanations into clinicians' decision-making workflow ..... &lt;RESULT&gt; We demonstrated the effectiveness of VBridge through two case studies and expert interviews with four clinicians, showing that visually associating model explanations with patients' situational records can help clinicians better interpret and use model predictions when making clinician decisions .....&lt;/p&gt;&lt;DOC-SEP&gt; Introduction: ..... &lt;ANALYSIS&gt; We derived seven design requirements from a pilot study with these clinicians; then, by observing their interactions with our early-staged system, we summarized two workflows - forward analysis and backward analysis - preferred by clinicians with different levels of expertise. &lt;METHODOLOGY&gt; These requirements and workflows guided the overall design and development of VBridge, a Visualization system that Bridges the gap between clinicians and ML models with tailored feature explanation algorithms and novel interaction and visualization techniques ..... &lt;RESULT&gt; The system was evaluated through two case studies and an expert interview with four clinicians, and results showed that our system is capable of supporting clinical decision-making.&lt;/p&gt;&lt;DOC-SEP&gt; .....&lt;/p&gt;&lt;DOC-SEP&gt; &lt;RESULT&gt; Evaluation: In this section, we first introduce two case studies conducted with two clinicians (P1, P5) for evaluating whether VBridge and our proposed workflows can support clinical decision-making.....&lt;/p&gt;&lt;DOC-SEP&gt; &lt;ANALYSIS&gt; Discussion: These case studies suggest that VBridge is helpful to clinicians and can support them in their decision-making. In addition to the case studies, we conducted semi-structured interviews with P4 and P6 by showing them the case study results and encouraging them to freely explore the system to collect additional feedback.....&lt;/p&gt;&lt;DOC-SEP&gt; Conclusion: &lt;ANALYSIS&gt; In this work, we identified three key challenges limiting the use of ML in clinical settings, including clinicians' unfamiliarity with ML features, lack of contextual information, and the need for cohort-level evidence. &lt;METHODOLOGY&gt; We then introduced VBridge - a visual analytics system designed according to the requirements identified in a pilot study - to support clinicians using ML to make decisions with both forward and backward analysis workflows. &lt;ANALYSIS&gt; We conducted two case studies and expert interviews with four clinicians. Their positive feedback and in-depth insights demonstrate the usefulness and effectiveness of the system.....&lt;/p&gt;&lt;/td&gt;&lt;/tr&gt;&lt;tr&gt;&lt;td style="vertical-align: top; padding-right: 20px;"&gt;<b>Target</b>&lt;/td&gt;&lt;td&gt;&lt;ANALYSIS&gt; A summary of seven design requirements facilitating the interpretation of ML predictions to clinicians; and the identification of two workflows describing how they work with ML models with feature-level explanations and needed context information.&lt;/p&gt;&lt;METHODOLOGY&gt; A visual analytics system that integrates novel explanation algorithms and visualization and interaction techniques, to connect the dots between ML features, explanations, and health records for an improved clinicians' decision-making workflow.&lt;/p&gt;&lt;RESULT&gt; Two case studies and an expert interview demonstrating the usefulness and efficiency of our system.&lt;/p&gt;&lt;/td&gt;&lt;/tr&gt;&lt;/table&gt;&lt;hr&gt;&lt;/div&gt;&lt;div data-bbox="191 813 806 833" data-label="Caption"&gt;&lt;p&gt;Figure 4.1: Organization of input and target from an example in our dataset&lt;/p&gt;&lt;/div&gt;&lt;div data-bbox="483 950 511 968" data-label="Page-Footer"&gt;&lt;p&gt;35&lt;/p&gt;&lt;/div&gt;&lt;div data-bbox="676 944 963 963" data-label="Page-Footer"&gt;&lt;p&gt;doi:10.6342/NTU202203034&lt;/p&gt;&lt;/div&gt;</p>
--------------	---



### 4.2.3 Fine-grained Post-training

Recall that in our dataset construction process, we first download computer science papers from arXiv and then extract papers with contributions explicitly listed by the authors from the downloaded papers. This leaves a total of 90K papers treated as unsupervised data in our task. To leverage this huge amount of in-domain data, we present a novel post-training strategy tailored to our task.

In our fine-grained post-training stage, we propose to construct pseudo-summaries similar to the concatenation of disentangled contributions as previously described. Again, we resort to the highlight sentences in abstract, introduction and conclusion sections as in the finetuning stage. To construct coherent pseudo-summaries where each disentangled contribution discuss the efforts of the authors from a unique aspect yet combining them provides a comprehensive picture of the complete research progress, we first choose candidate highlight sentences as backbones from either abstract or conclusion since they are closer to the reference summaries based our analysis. We perform a simple check on the sum of the length of the candidate highlight sentences to exclude outliers. If both of them are qualified, we randomly choose one as the backbone. Next, for each sentence in the backbone summary, we greedily select a sentence in the candidate set that shares the same contribution type and maximize the ROUGE-1 and ROUGE-2 f1 with it. If this maximum score exceeds certain replacement score threshold  $\phi$ , we replace the target sentence in the backbone with this new candidate sentence. As a result, the usage of highlight sentences in one section as backbone ensures that the pseudo-summary is coherent, yet the replacements based on ROUGE-score matching encourages the model to reconstruct important sentences in all of abstract, introduction and conclusion sections given the body text.



After the selection, we mask out all the selected sentences in the input document with a special token <mask>. Inspired from previous works [7, 36], we also randomly masked out the remaining highlight sentences in AIC. Finally, we organize the input and output as the same in the finetuning stage described previously. Figure 4.2 demonstrates one example from our post-training dataset. In this example, the abstract is chose as the backbone and extracted sentences are highlighted. Based on our sentence masking strategy, the sentences highlighted in brown and green are replaced with similar statements in conclusion and introduction respectively. In this way, the masked targets are not limited to the backbone and the model is trained to leverage important contexts across multiple sections without sacrificing the coherence of the pseudo-summary.

Overall, our generation method consists of a post-training stage and a finetuning stage. We denote our method built on BART and LED as BART-FP and LED-FP (Fine-grained Post-training) respectively.





---

<b>Abstract (Backbone)</b>	<p>Music, speech, and acoustic scene sound are often handled separately in the audio domain because of their different signal characteristics. However, as the image domain grows rapidly by versatile image classification models, it is necessary to study extensible classification models in the audio domain as well. <b>In this study, we approach this problem using two types of sample-level deep convolutional neural networks that take raw waveforms as input and uses filters with small granularity. One is a basic model that consists of convolution and pooling layers. The other is an improved model that additionally has residual connections, squeeze-and-excitation modules and multi-level concatenation. We show that the sample-level models reach state-of-the-art performance levels for the three different categories of sound. Also, we visualize the filters along layers and compare the characteristics of learned filters.</b></p>
<b>Pseudo-Summary</b>	<p>&lt;METHODOLOGY&gt; We presented the two sample-level CNN models that directly take raw waveforms as input and have filters with small granularity. &lt;METHODOLOGY&gt; One is a basic model that consists of convolution and pooling layers. The other is an improved model that additionally has residual connections, squeeze-and-excitation modules and multi-level concatenation. &lt;RESULT&gt; We show that the sample-level models reach state-of-the-art performance levels for the three different categories of sound. &lt;ANALYSIS&gt; Furthermore, we visualize hierarchically learned filters for each dataset in the waveform-based model to explain how they process sound differently.</p>

---

Figure 4.2: Example of our fine-grained sentence masking strategy



# Chapter 5 Experiments

In this chapter we present the experiment results in our work. Section 5.1 describes the training of our contribution type classifier and the classification results. Section 5.2 introduces the evaluation of disentangled contribution generation including experimental setups, evaluation metrics and main results.

## 5.1 Contribution Type Classification

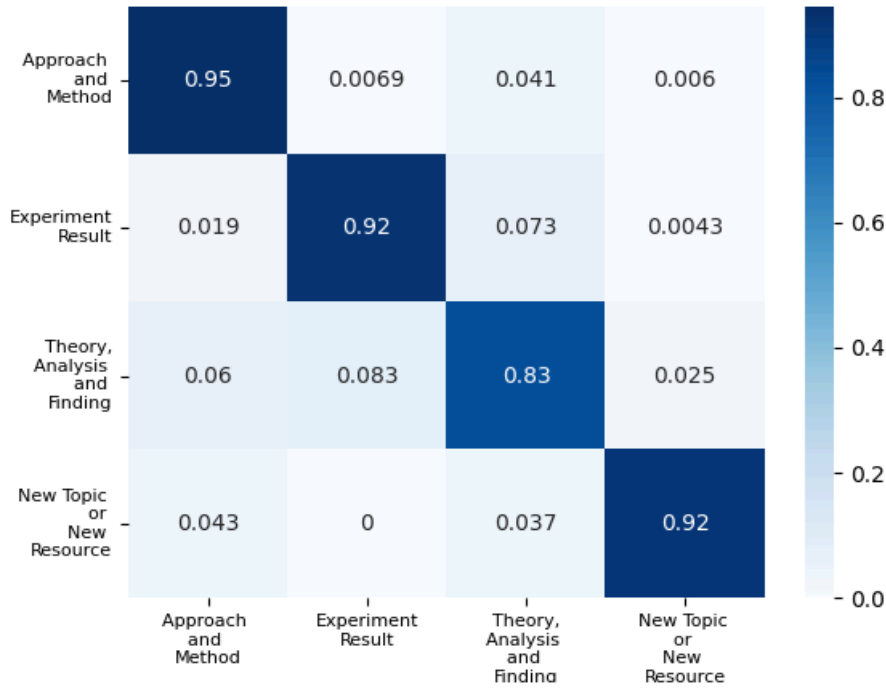
We first split our annotated dataset into a training set and a held-out test set of size 2750 and 709. The distribution of different contribution types is explicitly ensured to be balanced across the training set and the test set. We finetune SciBERT for classification. Specifically, we use the model checkpoint `allenai/scibert_scivocab_uncased` in Huggingface library<sup>1</sup>. The learning rate is set to  $2e-5$  and we train the model with a batch size of 64 for 5 epochs. After fine-tuning, our model achieves 89.66 Macro-F1, 90.5 Micro-F1 on the held-out test set. The resulting confusion matrix is presented in Figure 5.1.

Compared with other annotation schemes for contribution type classification, the overall results in our work are much better, indicating that our annotation scheme is less challenging yet reasonable to develop automatic methods. The relative high classification

---

<sup>1</sup><https://huggingface.co/models>

Figure 5.1: Confusion matrix of contribution type classification



accuracy also enables us to obtain high-quality labels for those unannotated contributions in our dataset by applying the trained classifier. From the confusion matrix in Figure 5.1 we can see that the performance of our model in Theory, Analysis and Finding is the poorest among all categories mostly because the model cannot classify it with Approach and Method or Experiment Result perfectly. We attribute this to the diversity of the semantic meanings in this category where authors might discuss the analysis of their proposed approach, findings from their experiment results other than common evaluation metrics or even the methods they utilize to perform the analysis, all posing challenges to the classification.

## 5.2 Disentangled Contribution Generation



### 5.2.1 Experimental Setup

For training and evaluation, we split our dataset to a train/validation/test set with size 19434/2302/2393. The distribution of papers in different categories is explicitly ensured to be balanced across the three sets. We also assign the 1K papers with manually annotated contribution types to the validation set for further analysis.

We build our method on BART and LED. The input length limit for BART and LED are set to 1024 and 4096 respectively, and the output length limit is 400 for both models. We use the sentence tokenizer provided by the Natural Language Toolkit [49]. We first post-train the models on our self-supervised dataset for 4 epochs then finetune on the gold training set for another 4 epochs. With a warm-up ratio of 0.1, the learning rates for BART and LED are  $7e-5$  and  $4e-5$ , in addition, the effective batch size<sup>2</sup> is set to 64. At inference time, we decode using beam search with a beam size of 5. The total training time (training, validation and model saving) using 2 Nvidia V100 is roughly 4 days and 20 hours for post-training and finetuning with LED, or 1 day and 6 hours with BART.

Recall that in our highlight sentence extraction algorithm, we have several hyper-parameters, namely length threshold  $l_s$  and  $l_e$ , prediction score threshold  $\tau$ . In our experiments,  $l_s$  and  $l_e$  are set to 0.4 and 0.8, which is roughly the relative position of 90 percentile among the greedily extracted sentences in their associated sections as demonstrated in Figure 3.2. The prediction score threshold  $\tau$  is set to 0.92 by a grid search on the validation set. This results in 9.8 highlight sentences for each source document on

---

<sup>2</sup>batch size per device \* number of devices \* gradient accumulation steps

average. In the process of pseudo-summary construction, the replacement score threshold  $\phi$  is 0.6. The average number of sentences in the pseudo-summary is 5.1 and that of the replaced sentences by our post-training strategy is 2.2. We also randomly mask out other unselected highlight sentences with a probability of 0.3.

We compare our methods with the following well-known and competitive baselines in document summarization and their corresponding initial model checkpoints in the Huggingface library:

- BART: facebook/bart-large
- PEGASUS: google/pegasus-large
- LED (Longformer Encoder Decoder): allenai/led-large-16384
- PRIMERA: allenai/PRIMERA

In addition, we also perform oracle extraction [50] which can be treated as the upper-bound of extractive methods.

### 5.2.2 Evaluation Metrics

Our evaluation is based on four automatic metrics: ROUGE-1 f1, ROUGE-2 f1, ROUGE-L f1 and BERTScore [51] f1<sup>3</sup>. Rouge scores measure the lexical similarity between the reference and the generation result by calculating unigram, bigram and longest common subsequence overlaps while BERTScore measures the semantic similarity using the embeddings of pre-trained language models. As recommended in the original paper, we use microsoft/deberta-xlarge-mnli for our BERTScore calculation since it has the best

---

<sup>3</sup>Hash code: microsoft/deberta-xlarge-mnli\_L40\_no-idf\_version=0.3.10(hug\_trans=4.13.0.dev0)

correlation with human evaluations. For better visualization, our reported scores in the following tables are original scores multiplied by 100.



Built on these automatic evaluation metrics, we perform our evaluation on two granularity: summary-level and contribution-level. In summary-level evaluation, the references and the generation results are both treated as a whole by concatenating all the disentangled contributions. In this setting, we evaluate the overall quality of the generation, as we hypothesize that the combination of all contributions should serve as a well-structured and comprehensive summary for the scientific paper, this should includes the reasonable and coherent organization of the disentangled contributions, possibly aligned with the paper structures. On the other hand, inspired by the facet evaluation of extractive summarization models [28], we evaluate the generation results on a contribution level since ideally each generated contribution should discuss certain aspect of the paper and match with exactly one contribution in the reference. Hence, for every test instance, we split the generation result into disentangled contributions  $G = \{G_1, G_2, \dots, G_n\}$  and map each contribution in the reference  $R = \{R_1, R_2, \dots, R_m\}$  to a generated contribution by maximizing ROUGE recall and BERTScore recall, where  $n$  and  $m$  are the number of contributions in the generation result and the reference respectively:

$$R_i \mapsto G_j \mid j = \arg \max_{k \in [1:n]} R1_{recall}(G_k, R_i) + R2_{recall}(G_k, R_i) + BS_{recall}(G_k, R_i) \quad (5.1)$$

by this mean, every reference is matched with the generated one that covers itself the most while there is no guarantee on the opposite side, which depends on the generation quality. Finally, we calculate the aforementioned automatic metrics between the mapped pairs. We choose to perform the matching based on the references because we are only confident that the contents in the references are correct. While there might exist cases

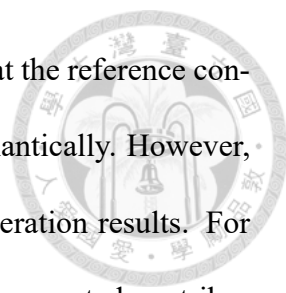
where the generated contribution is also factually correct but not presented in the reference, we are unable to evaluate its quality given our current dataset. Thus in our contribution-level evaluation, they are ignored. One potential drawback of our evaluation is that our matching algorithm cannot penalize redundant generated contributions (not mapped to any reference contribution) that are also not factually correct, yet we show in section 6.4 that these cases are not frequent. Figure 5.2 shows an example of the contribution-level mapping.

---

<b>Reference</b>	<ol style="list-style-type: none"><li>1. We formalize a new problem of open-set annotation (OSA) for real-world large-scale annotation tasks. It raises a practical challenge on how to maintain a higher recall to find the examples of known classes from a large unlabeled open-set for target model training.</li><li>2. We propose a new active learning framework LfOSA to address the OSA problem. It iteratively trains two networks, one for distinguishing the known and unknown classes, while the other one for classification of target task. To the best of our knowledge, this is the first work on active learning for open-set annotation.</li><li>3. The experimental results validate that the proposed approach can significantly improve the selection precision and recall, while achieving higher classification accuracy with lower annotation cost than state-of-the-art active learning methods.</li></ol>
<b>Generation Result</b>	<ol style="list-style-type: none"><li>1. We formulate a new open-set annotation (OSA) problem for real-world large-scale annotation tasks. It introduces a practical challenge on how to maintain a high recall in identifying the examples of known classes for target model training from a massive unlabeled open-set.</li><li>2. We propose an active learning framework called LfOSA to precisely select examples of known classes by decoupling detection and classification. By minimizing low-temperature cross-entropy loss, it exploits both known and unknown supervision to train a detector, whose activation values will be fed into a mixture Gaussian model to estimate the per-example max activation value (MAV) distribution. Based on MAV distribution, we can distinguish examples of known classes against unknown classes in unlabeled data to build a query set for annotation.</li><li>3. Experiments are conducted on multiple datasets with different mismatch ratios of known and unknown classes. The experimental results show that the proposed method can significantly improve the selection quality of known classes, and achieve higher classification accuracy with lower annotation cost than state-of-the-art active learning methods.</li></ol>

---

Figure 5.2: Example of contribution-level matching



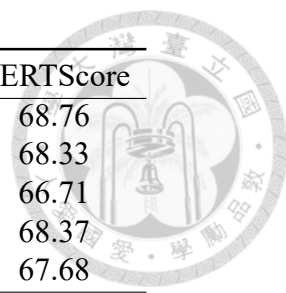
We highlight the matched pairs in the same color. We can see that the reference contribution and the generation result align well both sequentially and semantically. However, in other cases there might be noisy mapping resulted from poor generation results. For example, several reference contributions might mapped to the same generated contribution if the generation result is less disentangled and covers many different points in one contribution.

### 5.2.3 Main Results

Table 5.1 presents the results of summary level evaluation. Our methods outperform competitive baselines across all automatic metrics. Since our methods are built on BART and LED, we can directly observe the improvements brought by our finetuning leveraging contribution types and paper structures as well as the fine-grained post-training strategy. Compared with BART and LED, both BART-FP and LED-FP achieve improvements of over 1 average ROUGE score. Our methods also work well in terms of BERTScore. This indicates that our method is model-agnostic and can be incorporated with both the vanilla transformer and the long transformer. The performance margin of BART is slightly larger than LED, we hypothesize that the important contexts needed for generating research contributions still mostly lie in abstracts, introductions and conclusions. As we focus on adding special tokens for highlight sentences in these sections, the ability of encoding dependencies between texts in other sections by the long transformers might not be of significant values. Yet this may inspire future works to incorporate long transformers with methods that can exploit valuable guidance signals in the body text other than AIC.

Table 5.2 presents the results of contribution level evaluation. The relative comparisons between different models are consistent with Table 5.1. However, the performances



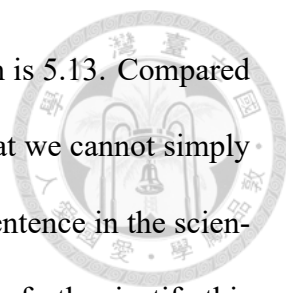


Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
EXT-ORACLE	54.1	29.18	34.09	68.76
BART	48.45	20.16	30.46	68.33
PEGASUS	46.85	20.16	30.87	66.71
LED	48.61	21.21	31.78	68.37
PRIMERA	48.39	20.99	31.05	67.68
BART-FP (ours)	<b>50.06</b>	21.25	31.73	68.75
LED-FP (ours)	49.87	<b>22.22</b>	<b>32.68</b>	<b>68.9</b>

Table 5.1: Results of Summary Level Evaluation

of all models are significantly worse than that in summary level evaluation. This is no surprise since that the reference become one contribution discussing specific aspect of the research paper instead of concatenations of contributions that are much similar to a summary covering all aspects. Therefore, ideal generation results should be able to separate and organize different contributions. Based on our experiment results, there is still room for improvements for models to achieve better performances in disentangled generation. Notably, PRIMERA performs worse than LED on our dataset. This is not expected as it is further post-trained from the LED checkpoint and also achieves state-of-the-art results in multi-document summarization and scientific paper summarization. The main reason is that the generation results of PRIMERA are about 10% longer than the reference targets as well as the results of other models, both in summary level and contribution level evaluations.

Though the performance of oracle extraction (which should be seen as the upper-bound of extractive methods) is better than that of the abstractive summarization models. We argue that extractive methods are not suited for our task. The most principal reason is that our task aims to generate disentangled contributions, while current extractive methods perform sentence-level extraction to decide whether each sentence in the paper should be included in the summary. This results in structure-less plain texts. Based on the ora-



cle extraction result, the average number of sentence in the prediction is 5.13. Compared with the number of contribution (3.2) in our dataset, this indicates that we cannot simply treat each extracted sentence as a contribution since more than one sentence in the scientific document are needed to cover the content of one contribution. To further justify this point, we perform another oracle extraction by treating each disentangled contribution as a reference. The average number of extracted sentence for each contribution is 1.94. In addition, the average number of section where the extracted sentences locate in for each contribution is 1.67. This shows that a divide-and-conquer approach by simply extracting from each section and combine them as a resulting contribution might not work well. Another reason is we find that some keywords addressing the contributions of the paper is not presented in the source input, which is likely to be missed by the extractive methods. According to the statistics of novel words (words that exist in the reference target but not in the source input), the most frequent (top-10) ones are: novel, propose, extensive, demonstrate, introduce, new, develop, best, present and provide. Based on this observation, extractive methods might not be enough to conclude and emphasize the contributions made by the authors instead of mainly coping narratives and statements in the paper.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
EXT-ORACLE	40.23	20.42	31.4	68.39
BART	35.27	13.73	27.14	66.59
PEGASUS	35.18	14.29	27.6	66.26
LED	36.01	14.9	28.09	66.92
PRIMERA	34.99	14.61	26.77	65.89
BART-FP (ours)	36.02	14.93	28.08	66.98
LED-FP (ours)	<b>36.69</b>	<b>15.65</b>	<b>28.48</b>	<b>67.04</b>

Table 5.2: Results of Contribution Level Evaluation

## 5.2.4 Comparisons with Other Post-training Strategies



We also compare our methods with other post-training strategies. For a fair comparison, we start from the LED checkpoint and train on our self-supervised dataset with the following post-training objectives.

- Gap Sentence Generation (GSG)[7]: a post-training objective tailored to abstractive summarization proposed in PEGASUS.
- Pyramid Sentence Masking (PSM)[36]: a post-training objective built on GSG that takes entity importance across multiple documents into account.
- Abstract: we explore a simple post-training strategy that simply trains the model to generate the abstract given the paper.

The results are presented in Table 5.3. In both summary level and contribution level settings, LED-FP outperforms all other post-training strategies. The margin in contribution level evaluations is larger since our method explicitly adds contribution type special tokens in both source documents and reference targets to provide alignments between them when generating certain types of contributions. In addition, the simple post-training objective of generating abstracts actually perform on par with other tailored post-training strategies.



Model	Summary-Level				Contribution-Level			
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
Ours	<b>49.87</b>	<b>22.22</b>	<b>32.68</b>	<b>68.9</b>	<b>36.69</b>	<b>15.65</b>	<b>28.48</b>	<b>67.04</b>
GSG	49.54	21.85	32.12	68.44	35.9	15.18	27.43	66.4
PSM	49.63	21.91	32.17	68.5	35.87	15.15	27.51	66.39
Abstract	49.83	22	31.99	68.54	36.09	15.34	27.61	66.46
LED	48.61	21.21	31.78	68.37	36.01	14.9	28.09	66.92

Table 5.3: Comparisons with other post-training strategies incorporated with LED



## Chapter 6 Discussion

In this chapter, we conduct further analysis and discussions of our methods. In section 6.1 we perform ablation study to investigate the components in our methodology. Section 6.2 presents the experiment results in low resource settings. Furthermore, We study the experiment results grouped by contribution types in section 6.3. Finally, we elaborate on the challenges of contribution-level generation performances in section 6.4.

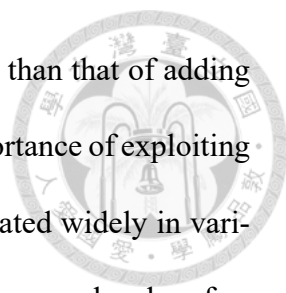
### 6.1 Ablation Study

Our method consists of two main components: the utilization of contribution type special tokens and the fine-grained post-training strategy. To investigate the effect of them in terms of model performances, we conduct ablation study on LED-FP and the results are presented in Table 6.1.

Model	Summary-Level				Contribution-Level			
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
Full model	<b>49.87</b>	<b>22.22</b>	<b>32.68</b>	<b>68.9</b>	<b>36.69</b>	<b>15.65</b>	<b>28.48</b>	67.04
w/o post-training	48.52	21.57	32.26	68.39	36.16	15.28	28.36	<b>67.22</b>
w/o special tokens	49.5	21.79	31.72	68.31	35.76	15.21	27.43	66.27
LED	48.61	21.21	31.78	68.37	36.01	14.9	28.09	66.92

Table 6.1: Ablation study of LED-FP

Based on the results, though both of our components contribute positively to the



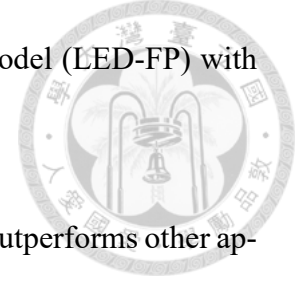
model performances, the effect of fine-grained post-training is larger than that of adding special tokens in summary-level evaluations. This showcases the importance of exploiting unlabeled in-domain data for self-supervised learning, also demonstrated widely in various NLP tasks including abstractive summarization where general summary-level performances are improved. [52–54]. On the other hand, for contribution-level performances, the benefit brought by adding special tokens are more significant, indicating the effectiveness of providing proper guidance signals in both source documents and reference targets to encourage the model to focus on generating well-structured salient points instead of flat general summaries. In contrast, our post-trained model without leveraging contribution type special tokens performs worse than vanilla LED in contribution-level evaluations. Last but not least, even without post-training, our finetuning method leveraging contribution type special tokens still outperforms the LED model with vanilla finetuning.

## 6.2 Experiment Results in Low Resource Setting

As discussed in previous sections, significant costs of annotations from domain experts have made it more difficult to acquire high-quality supervised datasets for scientific documents than common web based texts. In this regard, models that generalize well under low resource limitations become more important in scholarly document processing.

In this section, we present evaluations in zero-shot and few-shot settings where the model is provided with 0 and 100 training examples. Since few-shot results might vary according to the randomly sampled training data, we run the experiment five times (each time the training data is shared among all models) and report the average scores. The hyper-parameters are the same as that in fully-supervised experiments, except we train for

20 epochs for each run in the few-shot settings. We compare our model (LED-FP) with the previously-mentioned post-training strategies.



As showed in Table 6.2, our fine-grained post-training strategy outperforms other approaches significantly, since our method is tailored to our task, the improvement is larger than that in the fully-supervised setting. In addition, we observe that the margin between zero-shot or few-shot results and fully-supervised results are not as large as those in other summarization datasets. We attribute this phenomenon to two causes. First, our post-training is directly performed on in-domain data, namely computer science papers. The extremely similar distribution between self-supervised data and fully-supervised data enables our model to achieve surprisingly good results in low resource settings. Second, we speculate that there is a performance upper-bound for the models due to writing style variance in our dataset. Some authors prefer to write long statements describing their contributions in detail while others tend to use concise wordings such as noun phrases. This imposes challenges to the evaluation of our task as well as the generalization of the models to the fully-supervised dataset.

Model	Zero-Shot				Few-Shot			
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
Ours	<b>46.73</b>	<b>19.02</b>	<b>28.44</b>	<b>65.9</b>	<b>47.93</b>	<b>20.5</b>	<b>31.38</b>	<b>67.66</b>
GSG	43.08	16.35	24.77	62.89	47.27	20.15	31.03	67.29
PSM	44.86	17.53	26.57	64.11	47.84	20.32	30.98	67.38
Abstract	42.87	16.47	24.19	63.57	47.86	20.43	30.96	67.6

Table 6.2: Results of zero-shot and few-shot experiments



### 6.3 Results Based on Different Contribution Types

As our desired model should identify salient points in the paper and generate disentangled contributions of different types based on our annotation scheme. We study the performance of our model grouped by different contribution types. To ensure that the contribution types in the references are of high qualities, we focus on the papers manually annotated with their contribution types in the validation set. We then gather both the generation results and the references according to their contribution types, if there are multiple ones of the same type, we concatenate them. We calculate the coverage of contribution types between references and predictions using precision, recall and F1 scores. On top of that, we evaluate the automatic metrics for the grouped pairs. In the case where contributions of a certain type are not presented in the generation results, we simply omit the pairs. The results are presented in Table 6.3.

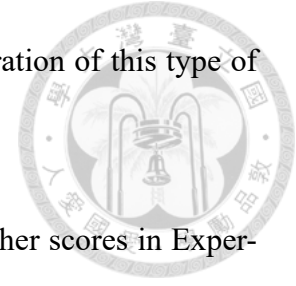
Contribution Type	Automatic Evaluation				Coverage		
	R-1	R-2	R-L	BS	P	R	F1
Approach and Method	45.02	21.55	33.37	68.7	94.8	95.08	94.94
Experiment Result	43.31	22.35	34.22	70.9	75.23	79.24	77.18
Theory, Analysis and Finding	37.47	15.66	28.01	65.52	76.32	53.7	63.04
New Topic or New Resource	44.78	23.27	37.56	70.5	81.9	60.56	69.64

Table 6.3: Results based on different contribution types

Among all contribution types, the coverage of Approach and Method is the best since it indeed takes up a large proportion in our annotation. In contrast, the generation results of our model does not cover contributions in Theory, Analysis and Finding very well. Similar to the case in contribution type classification, this type of contribution is rather complicated and challenging. Besides, as the classification performances on Theory, Analysis and Finding are already the worst among all the contribution types, the corresponding



predicted special tokens are likely to bring more noises to the generation of this type of contribution.



In terms of automatic metrics, surprisingly, model achieves higher scores in Experiment Result and New Topic or New Resource. We attribute this to the observations that the narratives of these two categories are more general than those in the other two categories. The key contents of them are more straightforward and easier to be identified in the paper. For example, contributions discussing experiment results usually mention their improvements over baselines or the state-of-the-art performances they achieve on certain datasets. Those describing new topics or new resources might simply state the task or dataset they present. On the other hand, contribution introducing methods or analysis requires detailed elaborations on domain knowledge with diverse expressions, as well as the ability of understanding and inferring to organize salient information to some extent. This is clearly much more challenging for the model to comprehend and further summarize.

We demonstrate one example from our generation results in Figure 6.1. We can see that the generation result does not state the contributions of type Analysis and Finding and New Topic or New Resource. For the former one, our model actually mentions it partially yet fails to distinguish it from the contribution of Experiment Result type. In addition, the generation results discussing methodologies are directly copied from the paper and lacks detailed descriptions.



---

## Reference

1. <METHODOLOGY> We tackle the keyword mapping problem as a sequence tagging problem and borrow state-of-the-art deep learning approaches tailored for well-known NLP tasks.
2. <METHODOLOGY> We extend the neural structure for sequence tagging, by utilizing multi-task learning and cross-skip connections to exploit the observation we made in natural language query logs of databases, that is, schema tags of keywords are highly correlated with POS tags.
3. <RESOURCE> We manually annotate query logs from three publicly available relational databases, and five different schemas belonging to Spider dataset.
4. <RESULT> We evaluate DBTagger, with above-mentioned query logs in two different setups. First, we compare DBTagger with unsupervised baselines preferred in state-of-the-art NLDBs. In the latter, we evaluate DBTagger architecture by comparing with different supervised neural architectures. We report new state-of-the-art accuracy results for keyword mapping in all datasets.
5. <ANALYSIS> We provide comprehensive run time and memory usage analysis over the existing keyword mapping approaches. Our results show that, DBTagger is the most efficient and scalable approach for both metrics.

---

## Our Generation Result

1. <METHODOLOGY> We propose DBTagger, a novel deep sequence tagger architecture to solve the problem of keyword mapping in NLDBs.
2. <METHODOLOGY> DBTagger is an end-to-end and schema independent solution, which makes it practical for various relational databases.
3. <RESULT> We evaluate our approach on eight different datasets, and report new state-of-the-art accuracy results, on the average. Our results also indicate that DBTagger is faster than its counterparts up to and scalable for bigger databases.

---

Figure 6.1: An example of our generation result, contribution type special tokens are retained for illustrations and they are removed in evaluation

## 6.4 Analysis of Contribution-Level Evaluations



In this section, we present further analysis to elaborate on the challenges of contribution-level generation performances of our model. Recall that in our contribution-level evaluations the matching process is based on the reference contributions, that is, every reference contribution is guaranteed to be matched with one generated contribution that covers its most contents. On the other hand, this also leaves possible generated contributions that are not matched with any reference and those that are matched with more than one references.

To study the deficiency of our model in terms of contribution-level evaluations, we first compare the number of disentangled contributions in each pair of our generation results  $N_{generated}$  and the test sets  $N_{reference}$  in Table 6.4. Overall, our model tends to generate less contributions than the references. This indicates that there is still space for improvements in regards of the model's ability of identifying and organizing salient points in the paper. The shortcoming becomes more obvious when we investigate the cases of matching errors. We calculate the ratio between generated contributions that are mapped to at least one reference contribution and the total number of generated contributions for each instance in the test set, the resulting histogram is presented in Figure 6.2. Ideally, as demonstrated in Figure 5.2, a good generation result should contain contributions that can be evenly matched to the reference contributions and cover all of them if possible. However, the actual cases are far from satisfaction as showed in the histogram. On average, the ratio of matched contributions is 63.59% and only 22.55% of our generation results are perfectly matched. Considering that the numbers of generated contributions are already smaller, there are still significant amounts of redundant ones that are not mapped to any reference in the generation results of our model, not to mention the quality of those

repeatedly matched ones in our contribution-level evaluations.



Case	Percentage
$N_{generated} > N_{reference}$	14.77%
$N_{generated} = N_{reference}$	51.39%
$N_{generated} < N_{reference}$	33.84%

Table 6.4: Comparisons of the contribution numbers in generation results and references

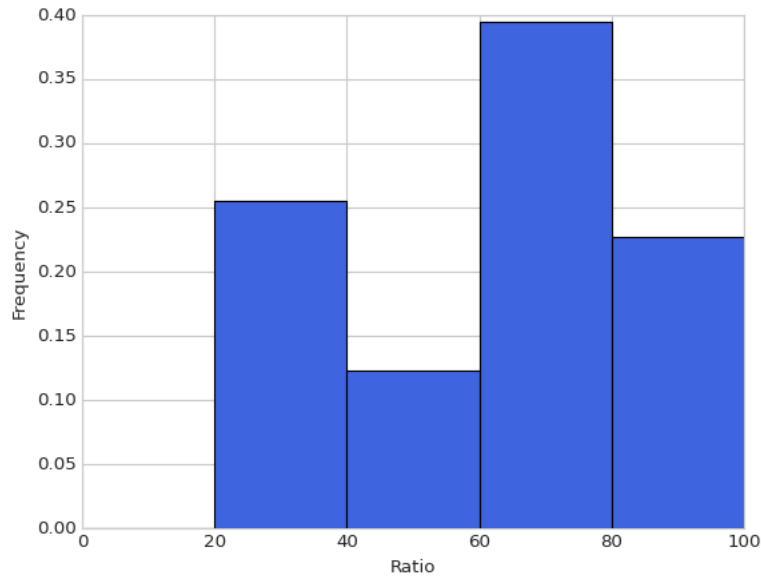


Figure 6.2: Histogram of the ratios of matched contributions generated by our model

In addition, we also study the disentanglement of our generation results and the reference targets. Since each disentangled contribution is supposed to cover information separately, it should have minimal overlap both lexically and semantically compared with other ones. Following Hayashi *et al.* [5], we calculate the DisROUGE and the DisBERTScore for each generated contribution. Specifically, for each instance, we pair every generated(reference) contribution  $C_i \in \{C_1, C_2, \dots, C_n\}$  with another one  $C_j$  that is most similar to it by maximizing the sum of ROUGE(1,2 and L) F1 and BERTScore F1. We report

the complement scores of each automatic metrics:



$$\langle C_i, C_j \rangle \mid j = \arg \max_{k \in [1:n]} \text{Score}((C_i, C_k)) \forall i \in [1 : n]$$

$$\text{Score}((C_i, C_k)) = R1_{F1}(C_k, C_i) + R2_{F1}(C_k, C_i) + RL_{F1}(C_k, C_i) + BS_{F1}(C_k, C_i)$$

$$D-i = 100 - \text{ROUGE-i}_{f1}(C_i, C_j)$$

$$D-BS = 100 - BS_{f1}(C_i, C_j)$$
(6.1)

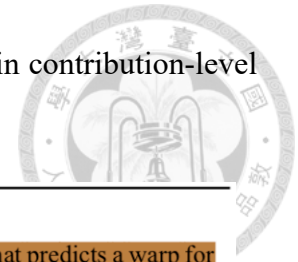
The results are presented in Table 6.5. Compared with author-written references, the generation results of our model are less disentangled across all evaluation metrics. This is consistent with our previous finding which indicates that our model still lacks human-level capability to comprehend salient information covering different aspects of the scientific researches and summarize them separately, which is crucial in generating structured summaries.

Model/Reference	D-1	D-2	D-L	D-BS
Reference	72.69	94.3	80.13	36.47
Ours	71.39	92.1	77.8	35.79

Table 6.5: Disentanglement scores of the references and the generation results of LED-FP

On top of numeric analysis, we provide an example in our test set to illustrate the aforementioned issues in Figure 6.3. From the summary-level perspective, the generation result is relatively descent as main points in the reference are mostly covered. However, the model fails to organize the contributions concisely as the original authors did. Specifically, the first generated contribution is actually the combination of the first two reference contributions. In contrast, the second and the third generated contributions are both parts of the third reference contribution, as each of them states more details. As a result, there are duplicate mappings for the first generated contribution while the third one is not

mapped to any reference. This results in performance deterioration in contribution-level evaluations.



---

#### Reference

1. A novel deep learning architecture , WarpNet, that predicts a warp for establishing correspondences between two input images across category and pose variations.
2. A novel exemplar-driven mechanism to train WarpNet without requiring supervised keypoint annotations.
3. An approach to unsupervised single-view object reconstruction that exploits the structure of the fine-grained dataset to yield reconstructions of birds nearly on par with the method that uses supervised part annotations.

---

#### Our Generation Result

1. We introduce a novel deep learning architecture, WarpNet, that learns to warp points on one object into corresponding ones on another (from a possibly different category or pose), without requiring supervised annotations.
2. We exploit the structure of the fine-grained dataset to create artificial correspondences for training this network in an unsupervised-discriminative learning approach. The output of the network acts as a spatial prior that allows generalization at test time to match real images across variations in appearance, viewpoint and articulation.
3. We demonstrate reconstructions that are nearly as good as those obtained using supervised annotations and better than those from appearance-only CNNs or unsupervised baselines such as deformable spatial pyramids.

---

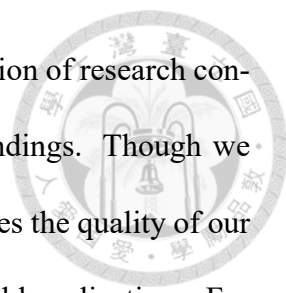
Figure 6.3: An example of our generation result, matched pairs are highlighted in the same color



## Chapter 7 Conclusion

In this thesis, we introduce the task of generating disentangled research contributions for scientific documents. To tackle data scarcity and facilitate the development of other tasks related to research contributions, we present ContributionSum, a contribution summarization dataset built on arXiv papers in computer science categories with research contributions explicitly listed by the authors. Furthermore, we design a new annotation scheme for contribution type classification. Based our annotation scheme, we provide human annotations of contributions in 1K papers and apply a data-driven approach to annotate all the contributions in our dataset. To build summarization systems tackling our task, we propose a simple yet effective sentence masking strategy tailored to our task for fine-grained post-training. We leverage existing pretrained models and incorporate them with paper structures as well as highlight contribution sentences in both source documents and reference targets. We conduct extensive experiments and the results of automatic evaluation metrics on both summary-level and contribution-level demonstrate the effectiveness of our proposed method as it outperforms competitive baselines and other post-training strategies.

In light of our work, there are several future directions for extensions or improvements. First, our dataset is solely made up of papers in computer science domains. Methods of automatically extracting contributions of papers in other research fields can be

The logo of National Taiwan University (NTU) is located in the upper right quadrant of the page. It is a circular emblem with a grey border. Inside the circle, there are traditional Chinese characters: '國立台灣大學' (National Taiwan University) around the top and '愛國愛校' (Love the country, love the school) around the bottom. In the center of the emblem, there is a bell and a book, symbolizing education and scholarship.

explored to extend our dataset for broader usages. Second, the definition of research contributions is subject to researchers' personal opinions and understandings. Though we focus on contributions written by the authors themselves which ensures the quality of our dataset, contributions from other sources are also valuable in real-world applications. For example, the contributions stated by the reviewers can be leveraged to study the review and rebuttal processes as authors and reviewers are likely to have different opinions on the values and the contributions over one review target. Last but not least, while our method built on existing pretrained models outperforms commonly-used baselines, future works can focus on designing better model architectures tailored to the task of disentangled generation in order to improve the results in both summary level and contribution level evaluations.

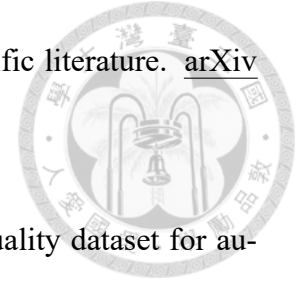




## References

- [1] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 615–621. Association for Computational Linguistics, 2018.
- [2] Arman Cohan and Nazli Goharian. Scientific article summarization using citation-context and article’s discourse structure. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 390–400, 2015.
- [3] James Hartley and Matthew R. Sydes. Are structured abstracts easier to read than traditional ones. Journal of Research in Reading, 20:122–136, 1997.
- [4] Jennifer D’Souza, Sören Auer, and Ted Pedersen. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021.
- [5] Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caim-

ing Xiong. What's new? summarizing contributions in scientific literature. [arXiv preprint arXiv:2011.03161](#), 2020.



[6] H Chen, H Nguyen, and A Alghamdi. Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles. [Scientometrics](#), 2022.

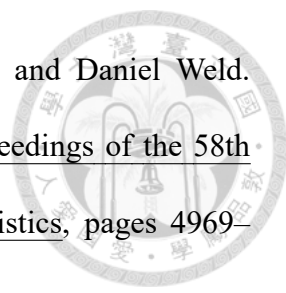
[7] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In [International Conference on Machine Learning](#), pages 11328–11339. PMLR, 2020.

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

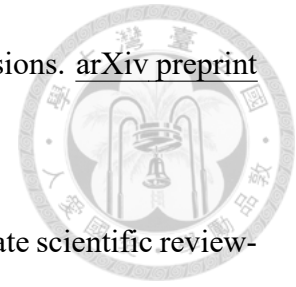
[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. [arXiv preprint arXiv:1910.13461](#), 2019.

[10] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. [arXiv preprint arXiv:2004.05150](#), 2020.

[11] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In [Proceedings of the Sixth International Conference on Language Resources and Evaluation \(LREC'08\)](#), 2008.

- 
- [12] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983, 2020.
- [13] Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 1080–1089, 2021.
- [14] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. arXiv preprint arXiv:1903.10104, 2019.
- [15] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. Argument pair extraction from peer review and rebuttal via multi-task learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7000–7011, 2020.
- [16] Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. Argument pair extraction via attention-guided multi-layer multi-cross encoding. 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 6341–6353, 2021.
- [17] Neha Nayak Kennard, Tim O’Gorman, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Rajarshi Das, Hamed Zamani, and Andrew Mc-

Callum. A dataset for discourse structure in peer review discussions. arXiv preprint arXiv:2110.08520, 2021.



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

[19] Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. MR<sub>ED</sub>: A meta-review dataset for structure-controllable text generation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2521–2535. Association for Computational Linguistics, May 2022.

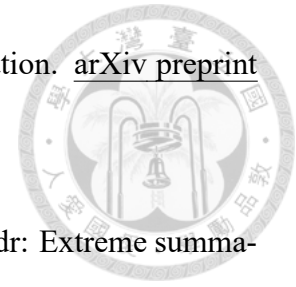
[20] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. Quantitative Science Studies, 1(1):396–413, 2020.

[21] Zhihong Shen, Hao Ma, and Kuansan Wang. A web-scale system for scientific knowledge exploration. In Proceedings of ACL 2018, System Demonstrations, pages 87–92, 2018.

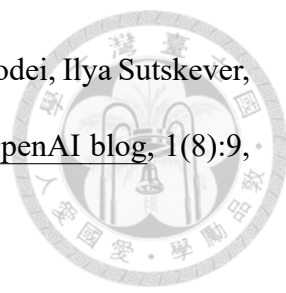
[22] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pages 84–91, 2018.

[23] Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming

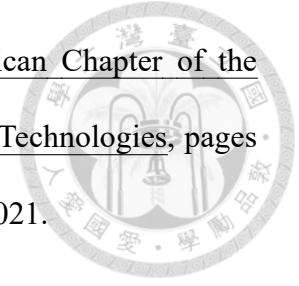
Xiong. Ctrlsum: Towards generic controllable text summarization. [arXiv preprint arXiv:2012.04281](#), 2020.




- [24] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. Tldr: Extreme summarization of scientific documents. [arXiv preprint arXiv:2004.15011](#), 2020.
- [25] Ed Collins, Isabelle Augenstein, and Sebastian Riedel. A supervised approach to extractive summarisation of scientific papers. In [Proceedings of the 21st Conference on Computational Natural Language Learning \(CoNLL 2017\)](#), pages 195–205. Association for Computational Linguistics, August 2017.
- [26] Alexios Gidiotis and Grigorios Tsoumakas. Structured summarization of academic publications. In [Joint European Conference on Machine Learning and Knowledge Discovery in Databases](#), pages 636–645. Springer, 2019.
- [27] Shuaiqi LIU, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. Generating a structured summary of numerous academic papers: Dataset and method. In [Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22](#), pages 4259–4265. International Joint Conferences on Artificial Intelligence Organization, 7 2022.
- [28] Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. Facet-aware evaluation for extractive summarization. In [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics](#), pages 4941–4957. Association for Computational Linguistics, July 2020.
- [29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. [arXiv preprint arXiv:1910.10683](#), 2019.

- 
- [30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.
- [31] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197, 2019.
- [32] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- [33] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451, 2020.
- [34] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768, 2020.
- [35] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In NeurIPS, 2020.
- [36] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. Primer: Pyramid-based masked sentence pre-training for multi-document summarization. arXiv preprint arXiv:2110.08499, 2021.
- [37] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. GSum: A general framework for guided neural abstractive summarization.

In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4830–4842. Association for Computational Linguistics, June 2021.

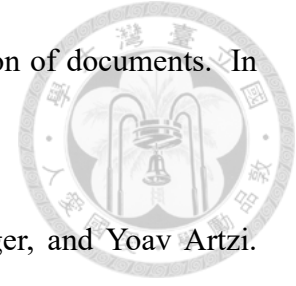


- [38] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. Planning with learned entity prompts for abstractive summarization. Transactions of the Association for Computational Linguistics, 9:1475–1492, 2021.
- [39] Yuning Mao, Wenchang Ma, Deren Lei, Jiawei Han, and Xiang Ren. Extract, denoise and enforce: Evaluating and improving concept preservation for text-to-text generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5063–5074. Association for Computational Linguistics, 2021.
- [40] Potsawee Manakul and Mark Gales. Long-span summarization via local attention and content selection. 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 6026–6041. Association for Computational Linguistics, 2021.
- [41] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. Aspect-controllable opinion summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6578–6593. Association for Computational Linguistics, 2021.
- [42] T Saier and M Färber. unarxive: a large scholarly data set with publications’ full-text, annotated in-text citations, and links to metadata. Scientometrics, 2020.

- 
- [43] Eva Sharma, Chen Li, and Lu Wang. Bigpatent: A large-scale dataset for abstractive and coherent summarization. [arXiv preprint arXiv:1906.03741](https://arxiv.org/abs/1906.03741), 2019.
- [44] Lin CY ROUGE. A package for automatic evaluation of summaries. In Proceedings of Workshop on Text Summarization of ACL, Spain, 2004.
- [45] Iz Beltagy, Kyle Lo, and Arman Cohan. 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. Association for Computational Linguistics, November 2019.
- [46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. [arXiv preprint arXiv:1810.04805](https://arxiv.org/abs/1810.04805), 2018.
- [47] Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2648–2662. Association for Computational Linguistics, November 2021.
- [48] Alexios Gidiotis and Grigorios Tsoumakas. A divide-and-conquer approach to the summarization of academic articles. [ArXiv, abs/2004.06190](https://arxiv.org/abs/2004.06190), 2020.
- [49] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”, 2009.
- [50] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neu-



ral network based sequence model for extractive summarization of documents. In Thirty-first AAAI conference on artificial intelligence, 2017.



- [51] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.
- [52] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360. Association for Computational Linguistics, July 2020.
- [53] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 704–717. Association for Computational Linguistics, June 2021.
- [54] Amir Soleimani, Vassilina Nikoulina, Benoit Favre, and Salah Ait Mokhtar. Zero-shot aspect-based scientific document summarization using self-supervised pre-training. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 49–62. Association for Computational Linguistics, May 2022.