GUIR @ MuP 2022: Towards Generating Topic-aware Multi-perspective Summaries for Scientific Documents

Sajad Sotudeh
IRLab, Georgetown University
sajad@ir.cs.georgetown.edu

Nazli Goharian
IRLab, Georgetown University
nazli@ir.cs.georgetown.edu

Abstract

This paper presents our approach for the MuP 2022 shared task—Multi-Perspective Scientific Document Summarization, where the objective is to enable summarization models to explore methods for generating *multi-perspective* summaries for scientific papers. We explore two orthogonal ways to cope with this task. The first approach involves incorporating a neural topic model (i.e., NTM) into the state-of-the-art abstractive summarizer (LED); the second approach involves adding a two-step summarizer that extracts the salient sentences from the document and then writes abstractive summaries from those sentences. Our latter model outperformed our other submissions on the official test set. Specifically, among 10 participants (including organizers' baseline) who made their results public with 163 total runs. Our best system ranks first in ROUGE-1 (F), and second in ROUGE-1 (R), ROUGE-2 (F) and Average ROUGE (F) scores.

1 Introduction

Scientific text summarization has received growing interest over the recent years (Cohan et al., 2018; Xiao and Carenini, 2019; Zerva et al., 2020; Cachola et al., 2020; Sotudeh et al., 2021; Cui and Hu, 2021; Pang et al., 2022; Sotudeh and Goharian, 2022), although it has been studied from years before (Teufel and Moens, 2002; Qazvinian and Radev, 2008; Nenkova et al., 2011; Qazvinian et al., 2013; Cohan and Goharian, 2015). Generating scientific summaries is deemed to be a challenging task, given the specific characteristics of scientific documents such as extreme document length, presence of complex domain-specific concepts, and specific structure, where the information is framed within sections. These characteristics of scientific papers, coupled with the aim of generating shorter or longer form summaries, call for special model considerations to deal with the challenging task of summarization. Researchers have looked into various approaches of unsupervised, supervised, neural, utilizing citations, knowledge, context, etc in generating the summaries in an extractive or abstractive way.

The existing evaluation systems in scientific summarization assume one signle gold summary for each scientific paper, based on which the summary generator optimizes the generation. The motivation of the MuP shared task (Cohan et al., 2022) is to provide multiple gold summaries per document so that the generated systems can be evaluated based on how well they captured various aspects of the paper into their summary. The assumption is that a single gold summary may not include multiple aspects expressed in the paper, as the writing of a summary is subjective. Specifically, the MuP organizers introduce a novel English summarization dataset collected from scientific peer reviews to reflect multiple perspectives from reviewers' standpoints. The participating teams are then asked to produce a scientific summary that can express diverse viewpoints on a given document.

In this study, we extend the Longformer Encoder-Decoder (LED) abstractive summarization model (Beltagy et al., 2020). In our experiments, we specifically explore two distinct approaches: (1) incorporating a neural topic modeling approach (Srivastava and Sutton, 2017) to the LED summarizer; and (2) proposing a two-step LED-based summarizer that first extracts the salient sentences and then performs abstraction over the extracted sentences to produce a *multi-perspective* summary. Our intuition of these extensions is that each *perspective* of a paper may focus on *specific sets of topics* which are discussed within specific *sets of sentences*, that should be taken into account by the summarizer. To benefit from the advantages of each of these approaches, we further combine them and propose a topic-aware two-step summarizer. Our combined model achieves the best results amongst the other settings on the validation and

Proceedings of the Third Workshop on Scholarly Document Processing, pages 273–278
October 12–17, 2022.

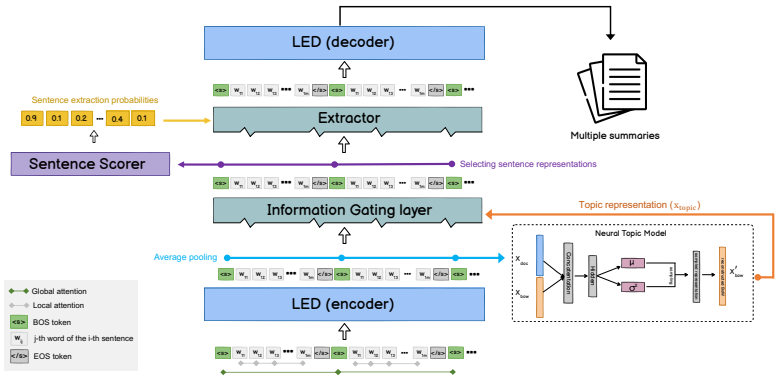

Figure 1: The overview of our proposed model. The LED encoder and decoder modules are expressed in blue boxes, the neural topic model takes in the contextualized representation of the document $\mathbf{x}_{\texttt{doc}}$ (average pooled from sentence representations), as well as the bow representation $\mathbf{x}_{\texttt{bow}}$ to generate topic representations $\mathbf{x}_{\texttt{topic}}$. The gating layer influences topic channels into the encoder outputs. The extractor picks the top sentences (in respect to the gold summaries) and passes their associated word representation to the decoder. The decoder attends to the top sentence representations for generating the summaries. In inference, we make the decoder generate only one summary.

official blind test sets. Specifically, it attains the first rank in ROUGE-1 (F), and second in ROUGE-1 (R), ROUGE-2 (F), and average ROUGE (F) scores, with 1.4% relative improvement over the baseline in terms of average ROUGE (F) scores.

2 Model

The general overview of our model is demonstrated in Figure 1. Our summarizer is composed of multiple components, including an LED encoder, a neural topic modeling layer, an information gating layer, and an extractor layer, followed by an LED decoder. In what follows, we explain the details of our proposed model.

2.1 Neural topic modeling for summarization

Topic modeling and text summarization can provide complementary features since both aim to distill salient information from a massive collection of textual data. With this intuition, we incorporate a *neural topic model* (NTM) (Miao et al., 2017; Srivastava and Sutton, 2017) into the summarization model (i.e., LED) to enrich the encoded word representations with topical information. We utilize the Combined Topic Model (Bianchi et al., 2021) as our topic modeling approach. This model is built around ProdLDA (Srivastava and Sutton, 2017), a neural topic modeling approach based on the Variational Autoencoders (VAE). VAE-based topic networks first infer a continuous latent representation $z \in \mathbb{R}^K$ (latent distribution over K topics) given the bag-of-words (bow) document representation $\mathbf{x}_{\texttt{bow}} \in \mathbb{N}^V$ (bow distribution over V distinct vocabulary). An NTM model assumes that z is generated from a prior distribution $p(z|x)$, which is estimated by the conditional distribution $q_\phi(z|x)$ modelled by a decoder ϕ. The NTM model aims to calculate the posterior $p(z|x)$, which is estimated by the variational distribution $q_\theta(z|x)$, modelled by an encoder θ. The NTM model optimizes the topic modeling network by defining the following loss criterion,

$$\mathcal{L}_{\texttt{topic}} = \max(\mathbb{E}_{q_\theta(z|x)}[\log p_\theta(x|z)] - \mathbb{KL}[q_\theta(z|x)||p(z)]). \quad (1)$$

The first term is the reconstruction error, and the second one is Kullback-Leibler (KL) divergence that regularizes $q_\theta(z|x)$. We refer the readers to (Srivastava and Sutton, 2017) for more details.

2.2 Information gating layer

After obtaining the topic representations, we influence the topic channels into the encoder representations that are the outputs from LED encoder layer. To this end, we design an information gating layer in which multiple linear layers are used to transform and combine topic and encoder representations and pass them along to the next stage. Formally written, let $\mathbf{x}_{\texttt{topic}}$ be the topic representation from the NTM model, and $\mathbf{x}_{\texttt{encoder}}$ be the contextualized word representations from the LED encoder. Our gating layer combines $\mathbf{x}_{\texttt{topic}}$ with $\mathbf{x}_{\texttt{encoder}}$ to implement a filtering gate, and then produces a *fused* word representation that has the information of both NTM and LED encoder,

$$\mathbf{x}'_{\texttt{topic}} = W_j \mathbf{x}_{\texttt{topic}} + b_j$$
$$g = \sigma(W_i[\mathbf{x}_{\texttt{topic}}; \mathbf{x}_{\texttt{encoder}}] + b_i) \quad (2)$$
$$\mathbf{x}_{\texttt{fused}} = (1-g)\mathbf{x}'_{\texttt{topic}} + (g)\mathbf{x}_{\texttt{encoder}}$$

where W_i, W_j, b_i, and b_j are trainable parameters, g is the filtering gate ($g \in [0-1]$), and $\mathbf{x}_{\texttt{fused}}$ is the topic-aware contextualized word representations.

2.3 Two-step summarization

After obtaining the topic-aware word representation, we aim to implement a two-step summarizer to drop the unimportant sentences and retain the salient content of the scientific document. In this sense, we ensure that the LED decoder only attends to the salient content of source information. To consider the sentential importance, we take the representations associated with the BOS token as the sentence representations and define a classification task over the document's sentences to predict summary-worthy sentences using a Sigmoid classifier. We then minimize the cross-entropy loss function as follows,

$$\mathcal{L}_{\texttt{sent}}(y, \hat{y}) = -\sum_{n=1}^{N}\sum_{i=1}^{|S|} y_i \log \hat{y}_i \quad (3)$$

in which y is the probability output from the Sigmoid classifier, $\hat{y}$ is the gold label, $|S|_d$ is the set of sentences within the scientific document, and N is the number of gold summaries for the given document. Upon obtaining sentential probabilities, we sample the representations associated with top sentences until a fixed length (e.g., 3072 tokens) is reached and then pass the resulting word representations to the decoder for summary generation. Then the model minimizes the following generation loss for a θ-parameterized model.

$$\mathcal{L}_{\texttt{gen}} = -\sum_{n=1}^{N}\sum_{t=1}^{T} \log P_\theta(\hat{y}_t)|\hat{y}_{<t}, x) \quad (4)$$

where N is the number of ground-truth summaries for a given document, and T is the length of summary in tokens. We then optimize the whole network using multi-tasking heuristics as follows,

$$\mathcal{L}_{\texttt{multi}} = \mathcal{L}_{\texttt{gen}} + (\alpha)\mathcal{L}_{\texttt{topic}} + (\beta)\mathcal{L}_{\texttt{sent}} \quad (5)$$

where $\mathcal{L}_{\texttt{gen}}$ is the cross-entropy generation loss computed from the decoder's outputs and gold summaries, and α and β are regularizing hyper-parameters for topic modeling and sentence extraction tasks, respectively.

3 Experiments

Dataset. We use the dataset introduced by the organizers and fine-tune it on our model. The MuP dataset (Cohan et al., 2022) is composed of scientific documents, each with one or more summaries that are the submitted peer reviews hosted by OpenReview platform [1]. There are 8,734 (train) and 1,060 (validation) distinct documents with a total of 26.5K summaries (with an average number of 2.57 summaries per paper), with summaries being 100.1 words long on average. The official blind test set includes 1,052 documents.

Experimental setup. We use the Huggingface Transformers library (Wolf et al., 2020) to implement our model. Specifically, we fine-tune `allenai/led-large-16384-arxiv` (an LED large model fine-tuned on arXiv scientific dataset (Cohan et al., 2018)) on the MuP dataset. The learning rate of our summarization system is set to be $1e-3$ for parameters that we train from scratch (i.e., Sigmoid classifier and topic modeling), and $3e-5$ for the rest of the parameters. α and β hyper-parameters are tuned to be 0.1, and 0.2. We train the models for 5 epochs [2], and perform evaluation in each 0.5 epoch. The checkpoint that achieves the best validation scores is further used for inference on the official test set.

Automatic results. Table 1 reports the system performances in terms of ROUGE (Recall and F) metrics, as well as the average ROUGE (F) on validation and official test sets. Our best system (i.e., LED (topic-aware $\oplus$ two-step)) achieves the first rank on ROUGE-1 (F), and second in ROUGE-1 (R), ROUGE-2 (F) and average ROUGE. We also see a similar trend of model performance on the validation set. It is also clear that the addition of two-step summarizer results in a promising performance boost, indicating that the extractor can efficiently ease the information flow from the encoder to decoder for generating improved summaries grounded on the most important sentences of the document. Considering the performance of the BART baseline, it appears that feeding first 1024 tokens of the document to the summarizer leads to a promising performance in ROUGE Recall metrics, but degrades the performance in terms of ROUGE precision metrics as we see a large decrease in ROUGE (F) scores. Our best model improves upon the baseline by 1.4% relative improvement.

[1] https://openreview.net/

[2] Empirically determined.

	Recall			F-measure			
	R-1(%)	R-2(%)	R-L(%)	R-1(%)	R-2(%)	R-L(%)	Avg. RG-F (%)
Other systems							
guneetAI	42.96	**13.98**	26.62	41.08	**13.29**	**25.36**	**26.58**
ashokurlana	40.13	12.33	24.74	40.68	12.47	24.99	26.04
MuP baseline	**44.20**	13.50	**26.81**	40.80	12.33	24.48	25.87
sandeep.kumar82945	42.02	11.98	24.26	40.37	11.98	24.26	25.54
prachuryanath	35.83	10.88	22.43	38.74	11.73	24.21	24.89
This work							
LED (topic-aware)	42.15	12.46	25.21	40.62	11.96	24.18	25.59
LED (topic-aware ⊕ two-step)	43.29	13.20	26.21	**41.36**	12.52	24.83	26.24

(a) Top 5 Participating teams' (on Avg. ROUGE (F)) system performance on official blind test set.

	Recall			F-measure			
	R-1(%)	R-2(%)	R-L(%)	R-1(%)	R-2(%)	R-L(%)	Avg. RG-F (%)
LED	40.17	11.97	24.61	39.97	11.79	23.76	25.38
LED (topic-aware)	42.19	12.70	24.39	40.70	12.15	24.07	26.03
LED (topic-aware ⊕ two-step)	**42.82**	**12.80**	**25.86**	**41.05**	**12.18**	**24.61**	**26.55**

(b) Our systems' and LED baseline's (Beltagy et al., 2020) performance on validation set.

Table 1: ROUGE (F1) results of (a) our submissions compared to the other top 5 participating teams on the official blind test set of MuP challenge, and (b) our system's results on validation set. **Bold** scores show the top scores (in (a) and (b)), and underlined scores are the second top (in (a)). The table is sorted by the average RG-F score (last column). The MuP baseline is the BART (Lewis et al., 2020) summarizer, submitted by the challenge organizers.

Analysis. To explore the qualities and limitations of each system, we further perform a qualitative analysis over a random set of 15 test papers, comparing LED baseline with our submitted models. The percentage rate of our observations is also presented in parentheses. We found that: (1) in outperformed cases, detected topics by the NTM component fairly align with those discussed in gold summaries (i.e., gold topics); hence, the summarizer is guided to pick up on the paper information around the gold topics (47%), (2) addition of two-step summarizer has the most effect on refining the paper in terms of dropping unimportant/irrelevant information (66%), (3) in underperformed cases, our topic-guided summarizers focus more on the topics that are frequently mentioned in the paper; missing those topics that are less mentioned despite their saliency in gold summaries (72%). This might be addressed in future work by some heuristics such as saliency-aware (Zou et al., 2021), and hierarchical (Jin et al., 2021) topic-modeling.

4 Related work

While scientific document summarization has a long history, it has recently gained increasing attention from research communities. Previous works have approached this problem by either generating regular-length summaries, such as (Qazvinian et al., 2013; Cohan et al., 2018) among many, or very recently so-called extended summaries (Chandrasekaran et al., 2020; Sotudeh et al., 2020; Ghosh Roy et al., 2020; Gidiotis et al., 2020). These attempts include hierarchical sequence modeling (Xiao and Carenini, 2019; Rohde et al., 2021; Pang et al., 2022; Ruan et al., 2022), citation-context based approaches (Qazvinian and Radev, 2008; Cohan and Goharian, 2015; Zerva et al., 2020; An et al., 2021), using documents' structural information as saliency signals (Cohan et al., 2018; Sotudeh et al., 2020, 2021; Sotudeh and Goharian, 2022), two-phase summarization models (Ghosh Roy et al., 2020; Gidiotis and Tsoumakas, 2020). Up to recently, majority of existing work in scientific domain has evaluated the systems assuming that there is only one gold summary per paper. MuP challenge is the first attempt toward evaluation of summarization systems given multiple gold summaries, each of which captures a specific aspect of the paper.

5 Conclusion

In this study, we explore two summarization approaches to tackle the multi-perspective summary generation task, organized by the MuP challenge. Our first model learns a latent topic distribution using neural topic modeling (NTM) in the fine-tuning stage, and the knowledge is shared between the topic modeling and text summarization task for summary generation. Next, as our second model, we further incorporate a two-step summarization framework into the summarization model for yielding even more improvements. Our best submission ranks first in ROUGE-1 (F); and second in ROUGE-1 (R), ROUGE-2 (F), and average ROUGE (F) scores.

References

Chen An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. *ArXiv*, abs/2104.03057.

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

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. 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 759–766, Online. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

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

Arman Cohan, Guy Feigenblat, Tirthankar Ghosal, and Michal Shmueli-Scheuer. 2022. Overview of the first shared task on multi-perspective scientific document summarization (mup). In *Proceedings of the Third Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2015. 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, Lisbon, Portugal. Association for Computational Linguistics.

Peng Cui and Le Hu. 2021. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5881–5891, Online. Association for Computational Linguistics.

Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta, and Vasudeva Varma. 2020. Summaformers @ LaySumm 20, LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 336–343, Online. Association for Computational Linguistics.

Alexios Gidiotis, Stefanos Stefanidis, and Grigorios Tsoumakas. 2020. AUTH @ CLSciSumm 20, LaySumm 20, LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 251–260, Online. Association for Computational Linguistics.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Yuan Jin, He Zhao, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Neural attention-aware hierarchical topic model. In *EMNLP*.

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

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

Ani Nenkova, Sameer Maskey, and Yang Liu. 2011. Automatic summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, HLT '11, USA. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wojciech Kryscinski, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2022. Long document summarization with top-down and bottom-up inference. *ArXiv preprint*, abs/2203.07586.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696, Manchester, UK. Coling 2008 Organizing Committee.

Vahed Qazvinian, Dragomir R. Radev, Saif M. Mohammad, Bonnie Dorr, David Zajic, Michael Whidby, and Taesun Moon. 2013. Generating extractive summaries of scientific paradigms. *J. Artif. Int. Res.*, 46(1):165–201.

Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. 2021. Hierarchical learning for generation with long source sequences. *ArXiv preprint*, abs/2104.07545.

Qian Ruan, Malte Ostendorff, and Georg Rehm. 2022. HiStruct+: Improving extractive text summarization with hierarchical structure information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1292–1308, Dublin, Ireland. Association for Computational Linguistics.

Sajad Sotudeh, Arman Cohan, and Nazli Goharian. 2020. GUIR @ LongSumm 2020: Learning to generate long summaries from scientific documents. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 356–361, Online. Association for Computational Linguistics.

Sajad Sotudeh, Arman Cohan, and Nazli Goharian. 2021. On generating extended summaries of long documents. *The AAAI-21 Workshop on Scientific Document Understanding (SDU)*.

Sajad Sotudeh and Nazli Goharian. 2022. TSTR: Too short to represent, summarize with details! intro-guided extended summary generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–335, Seattle, United States. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.

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

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. 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 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Chrysoula Zerva, Minh-Quoc Nghiem, Nhung T. H. Nguyen, and Sophia Ananiadou. 2020. Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics*, 125:3109–3137.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *AAAI*.