The Ohio State University
CSE 5525 - Speech and Language Processing
News Summarization Project Report

Sunit Singh
singh.1790@osu.edu

Abstract

This paper focuses on abstractive text summarization aided by the question-answering task. Specifically, we propose a neural abstractive summarization model whose goal goes beyond providing plain summaries by tackling a more challenging task—generating coherent summaries with the constraint of incorporating answers from the original text that readers may seek. To achieve this objective, we employ a cross-attention architecture, utilizing a pretrained T5 transformer model. We evaluate our model in terms of both the quality of the generated summary and its ability to address potential questions. Our results demonstrate that our model can produce coherent summaries containing relevant key points from the original text while also being able to provide answers to questions, as evident in our Rouge, entailment, and BERT scores. These findings warrant further investigation into question-aided text summarization, laying the foundation for more coherent and informative summaries.

1 Introduction

The increasing prevalence of human-human and human-machine interactions has given rise to a staggering amount of generated data, which is also growing exponentially each day. While the generated data comprises different modalities, the contribution of text is significant, thanks to websites, blogs, news, social networking sites, and human-to-human communication platforms (S et al., 2023). For example, although Twitter is neither the only social media platform nor the most used one, it accommodates 500 million tweets per day only by itself (Antonakaki et al., 2021). This vast quantity of raw text data makes it harder to organize, store, and even understand the content, as it includes irrelevant content and even noise in the form of inaccurate punctuation, typos, or miswording (Al Sharou et al., 2021). In response to these problems, text summarization has been used as a viable solution to convey the intended meaning rather than provide the exact text message. This process compresses the text into a shorter version, decreasing storage costs while preserving the underlying content (Khan et al., 2023). While the summarization task can be done manually, it can be too time-consuming, which motivates researchers to automate this process by proposing practical algorithms, leading to the concept of *automatic text summarization* (Mridha et al., 2021).

Current studies in *automatic text summarization* mainly revolve around two concepts: *extractive* and *abstractive*. In the former, the primary goal is to identify key sentences that convey the fundamental meaning and then combine them sequentially to construct a summary. In contrast, in the latter, the aim is to first comprehend the message itself and then generate a summary from scratch. Although *abstractive* summarization is a more challenging task, it is more beneficial in terms of improving understandability, readability, and cohesion (Kouris et al., 2019). Recently, the summarization process has started to be aided by the task of question-answering to mitigate factuality-related problems and improve the conveyed information from the original passage. In this context, the questions can be in the form of a factoid or non-factoid. The former refers to simple and fact-based questions asking for a short phrase or sentence, e.g., who?, when? (Mishra and Jain, 2016). In the latter, questions are supposed to be more complex and informative, requiring a comprehensive understanding whose answers are generally in the form of a text span exceeding short phrases (Deng et al., 2020).

In this paper, we concentrate on the abstractive text summarization task, complemented by the non-factoid question-answering (Q&A). Our primary objectives are to deliver a coherent summary with a logical flow and contextual information, as well as to generate a text capable of addressing potential questions that readers might pose regarding the passage. The main contributions of this work can be stated as follows.

- We propose a cross-attention-based architecture that utilizes pre-trained T5 transformer models as our question encoder and text decoder blocks. Our encoder comprises the T5 model trained on a Q&A, while the decoder is structured by the T5 model trained on a summarization task. In this context, we establish a link between these encoder and decoder by using keys and values from the summarizer decoder, while employing queries from the question encoder model. Additionally, we investigate the concept of low-rank adaptation within this framework.
- Our simulations conclusively demonstrate our model's proficiency in generating coherent summaries, as indicated by the Rouge, and Bert scores. Importantly, the model not only excels in summary generation but also exhibits the capability to answer questions accurately which is assessed by the entailment score.

The remainder of the paper is organized as follows. Section 2 presents the problem for which we propose the model. In section 3, we introduce our summarizer model with cross attention. Section 4 covers the evaluation metrics, dataset, and simulation results. We discuss the ethical issues in section 5. Finally, section 6 concludes the paper.

2 Problem Formulation

We focus on the summarization task aided by question-answering, with our objective being to maximize the *coherence* and *answerability* of the generated summary. We posit that by incorporating question-answering, we can enhance the factual content of the generated summary. Specifically, concerning *coherence*, our aim is to preserve contextual information and maintain the logical flow from the original passage, akin to the plain summarization task. Conversely, with regard to *answerability*, we monitor the ability of the generated summary to provide answers to questions posed to the original passage.

To achieve this mixed objective of *coherence* and *answerability*, we formulate a supervised learning task. Our model tries to optimize this objective by having the following inputs:

1. Passage to be summarized
2. Ground truth for the summary
3. Questions associated with the passage
4. Answers corresponding to the given questions
5. Binary label indicating whether the summary contains the answer

For this setup, we propose a cross-attention-based architecture to jointly optimize the objectives of *coherence* and *answerability*, as detailed in section 3. Building upon our prior research on query-prompted passage summarization (Singh, 2022), we conduct hyperparameter tuning, as outlined in section 4.3. We also incorporate BertScore to assess the conveyance of semantic understanding through summarization, as detailed in section 4.1. Additionally, we introduce the concept of low-rank adaptation, as discussed in section 3.4.

3 Proposed System Model

In this section, we will explain our proposed system model, which consists of two blocks, specifically, a text decoder and a question encoder, as depicted in figure 1. We will state the training algorithm which consists of three stages. We also elaborate on the concept of low-rank adaptation within this framework.

In our model, we employed the T5 transformer structure in both blocks, which was retrained to tailor it specifically for our objective. The T5 transformer model is a text-to-text model trained for multiple tasks, encompassing both supervised and unsupervised learning. This training approach results in a powerful model capable of setting a new benchmark in language understanding tasks (Raffel et al., 2020).

3.1 Summarizer Model

Our summarization model utilizes the pre-trained T5 transformer. We perform re-training on this model using a summarization dataset, specifically the CNN DailyMail, as detailed in section 4.2. This process enhances the model's specialization for the context of news summarization, aligning with the primary objective of our project. The training of this block constitutes the initial stage of our training algorithm. We employ the *Cross Entropy* loss during the training phase.

3.2 Question-Answering Model

Our question-answering (Q&A) model, structured once again as a T5 transformer, undergoes retraining on the SQuAD2.0 Q&A dataset, as detailed in section 4.2. The objective of this model is to encode given questions to generate queries for the cross-attention model. Hence, the process of this block enables the incorporation of Q&A in our summarization framework, as discussed in section 2. The training of this block constitutes the second stage of our training algorithm. Similar to the previous section, i.e., section 3.1, we employ the *Cross Entropy* loss function to train our Q&A block.

3.3 Cross Attention Structure and Training Algorithm

We execute the final stage of our training algorithm after obtaining the trained summarizer and Q&A blocks. During this phase, our model takes the input as outlined in section 2. However, none of the utilized datasets, namely CNN DailyMail or SQuAD2.0, can provide these inputs simultaneously. Consequently, we opt for the SQuAD2.0 dataset and leverage the 'facebook-bart-cnn' model to generate summaries from its passages. We chose this model because 'facebook-bart-cnn' is trained on the CNN DailyMail dataset, aligning with our summarization block (Lewis et al., 2020). Additionally, to obtain the binary *answerability* label, we apply the entailment concept, as detailed in section 4.1. We employ the 'roberta-large-mnli' pre-trained model to compare the generated summary and the actual answer to the posed question. This model provides entailment scores for each summary-answer pair (Liu et al., 2019), where we consider scores greater than 0.5 as a positive label.

To integrate the concepts of summarization and Q&A, we introduce cross-attention between these blocks, as depicted in Figure 1. Specifically, we extract queries from the question encoder while utilizing keys and values from the summarization block, as illustrated. This structure enables our summarization model to generate a question-aided summary with inputs as described in section 2. Our model attempts to generate the summary while simultaneously performing binary classification on the *answerability* label. The loss function is a composition of *Cross Entropy* (CE) losses, one for the summarization task and another for the binary classification task. We introduce the parameter α as a weight for the binary classification loss to adjust the contribution of each loss during training. Specifically, our loss function takes the form $CE_{Summarization} + \alpha CE_{Classification}$. This parameter also determines our model's behavior, indicating whether it should focus more on the factual content or the general logic flow of the original passage. Further details about our hyperparameter choices are provided in section 4.3.

3.4 Low Rank Adaptation in Summarization

Owing to the fusion of 2 complex architectures, we fine-tune a variant of the cross-attention architecture using low-rank adaptation(Hu, 2021) on the XSUM dataset. We propose that the cross-attention paradigm can be viably applied to develop models wherein we need a cross-task reference, in our case summarization and question-answering tasks. Since low-rank matrices generated during fine-tuning make the model portable by allowing plug-and-play capability, we propose that it opens the possibility for automated domain adaptation, wherein a domain classifier model identifies the appropriate LoRA weights and the main task executor (in our case, the cross attention network), is used for inference.

We use the open-sourced parameter efficient fine tuning(Mangrulkar et al., 2022) library from Huggingface to carry out low-rank adaptation during model fine-tuning. For pre-trained language models like FLAN-T5, PEFT methods fine-tune a small number of the model's trainable parameters. This helps in great training cost reduction and also significantly reduces the size of checkpoint files. In addition to LoRA fine-tuning we also use int8 quantization from the open-sourced bitsanbytes library thereby further reducing the memory footprint by a factor of 4.

We set the LoRA rank to 16 and the scaling factor to 32. We also use a dropout of 0.05 and feed the task type as a sequence-to-sequence task. We chiefly target *query* and *value* matrices to apply low-rank adaptation to. In another experiment, we freeze the attention weight matrices for the initial layers of the model and only make the last few layers of the summarizer model decoder and the cross-attention network layers trainable. We observe that this approach lowers the memory requirement of the model and its training time, but has a significant impact on the model's performance. Low-rank adaptation on the other hand fares much better, exhibiting on par performance with the orig-

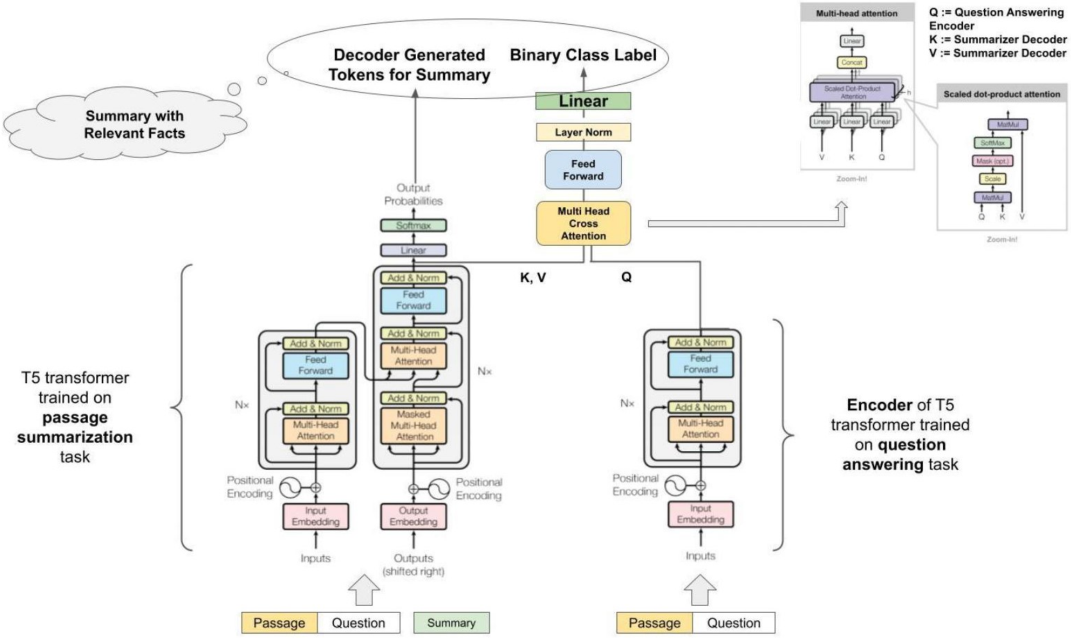

Figure 1: T5 Transformer-based Cross Attention Summarizer architecture

inal model. We argue this is because, unlike the former experiment, we are lowering the model's capacity across all feature matrices rather than empirically choosing whole encoder-decoder layers and freezing them during fine-tuning. After applying low-rank adaptation we observe the number of trainable parameters reduces to roughly 2M parameters as compared to 743M parameters in our original model, thereby making using a mere 0.18% of the parameters!

4 Evaluation and Numerical Results

In this section, we will briefly explain the performance metrics and dataset employed in the evaluation process. Subsequently, we will present the simulation settings and numerical results.

4.1 Performance Metrics

Our objective is to design a model to maximize *coherence* and *answerability*, as mentioned in section 2. To achieve this goal, we employ Rouge and Bert scores to evaluate the model's summarization quality i.e., *coherence*, while utilizing entailment scores to assess the factual content, i.e., *answerability* (Lin, 2004; Zhang et al., 2020; Paramasivam and Nirmala, 2022). We explain these metrics as follows.

1. **Rouge Score**: Rouge is a metric commonly used in sequence-to-sequence tasks, such as machine translation or automatic translation, as is the case here. It is a syntactic metric that compares the generated text with a reference, producing an output in the range of $[0, 1]$, where 1 signifies a perfect output that exactly matches the reference (Lin, 2004). Specifically, we utilize Rouge-1 and Rouge-2, investigating the match of 1-grams and 2-grams, respectively. This metric enables us to evaluate coherence syntactically.

2. **Bert Score**: To complement the syntactic Rouge metric, we incorporate Bert Score to gain insights into the semantic aspect of the generated summary. This scoring method leverages token embeddings to assess the output, indicating the degree of alignment with the reference in terms of meaning (Zhang et al., 2020). This metric enables us to evaluate coherence semantically.

3. **Entailment Score**: We employ entailment scores to assess the *answerability* of our model. Specifically, we calculate the entailment between the generated summary and the answers to gauge how well they are aligned (Paramasivam and Nirmala, 2022).

4.2 Dataset

We utilized the SQuAD2.0 Q&A dataset to train our question encoding block, as explained in section 3.2 (Rajpurkar et al., 2016). We preprocess the dataset by lowercasing all text, removing non-ASCII characters, eliminating protocols from URLs, and simplifying common contraction expressions. For instance, the contraction 'I am' was converted. Metadata was collected after this preprocessing to determine the highest token count and distribution for articles, summaries, and information from the SQuAD dataset. Following a thorough analysis, we set the token limit to 1024 to accommodate both the passage and questions for further examination of the cross-attention model. Our classification model was trained on over 90000 samples, with an average token count of 256.

We employed the Cnn DailyMail dataset to train our summarization model, as elaborated in section 3.1 (Nallapati et al., 2016). This dataset comprises news articles and their summaries, spanning over 300000 articles written between 2007 and 2015 by CNN and the Daily Mail. The average token count for news articles is 781, while summaries contain an average of 56 tokens. All samples are in English and were curated by affiliates of the University of Oxford. We chose this dataset over others, such as the multi-news and BBC News Summary datasets, due to its comprehensiveness, featuring a larger number of samples and a noticeable ideological contrast between news outlets. Specifically, the Daily Mail exhibits a right-leaning perspective, while CNN leans left. Additionally, we prioritized the dataset's recent validation, conducted in 2019, to identify and address potential biases, which were found to be minimal.

The XSum dataset was used in the fine-tuning of the cross-attention structure. The tuning was conducted on 226,711 archived BBC articles between 2010 and 2017 (Narayan et al., 2018). The data set is split 90% for training and 5% each for training and validation. The dataset consists of the text of the article and its summary.

4.3 Simulation Settings

We conducted hyperparameter tuning through grid searching, where values for each training stage, as detailed in section 3.3, are specified in table 1. Additionally, we selected the value for α to be 1 in loss function, as stated in section 3.3.

Table 1: Training Parameters

Param.	Summ.	Q&A	Cross-Att.
Num. Epochs	5	10	10
Learning Rate	0.0001		
Batch Size	8		
Optimizer	$Adam(\beta_{1,\,2} = 0.9, 0.999)$		

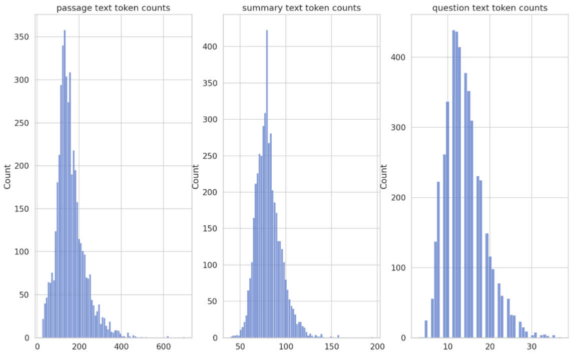

Figure 2: Token counts for the SQuAD2.0 dataset and the generated summaries for its passages by 'facebook-bart-cnn'.

4.4 Numerical Results

In this section, we discuss the numerical results obtained from our simulations. For our analysis, we established two different baselines: the 'facebook-bart-cnn' model and our summarization model alone. Since all these models are trained on the same summarization dataset, namely CNN DailyNews, we anticipate that the effect of our cross-attention architecture will be readily observable.

To begin, we present the token counts for the Squad 2.0 dataset, along with its summaries generated by the 'facebook-bart-cnn' model, as explained in Section 3.3. These plots in figure 2 offers insight into how much the baseline model compresses the text in terms of token count and provides information about the training dataset for our cross-attention model. Additionally, we include the learning curves of our models in figure 3. It can be seen that the loss is consistently decreasing in our training phase.

We further investigate the Rouge scores of the aforementioned models, as detailed in Table 2. The observation of these scores clearly emphasizes that our cross-attention-based model achieves superior performance in terms of Rouge scores, indicating that it provides summaries that are better aligned

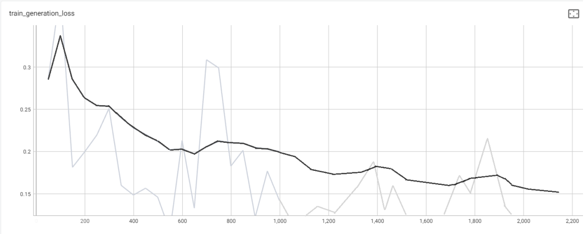

Figure 3: Learning curve of cross-attention model.

Table 2: Evaluate the *coherence* of the models through Rouge scores.

Model	Rouge-1	Rouge-2	Rouge-l
fb.-bart-cnn	0.713	0.603	0.687
Summ. Block	0.557	0.417	0.509
Cross-Atten.	**0.785**	**0.707**	**0.737**

with the actual summary syntactically, i.e., in terms of wording. Particularly, comparing the scores of our sole summarization block and the cross-attention model underscores the advantage of introducing Q&A guidance to the plain summarization task.

To enhance the insights obtained from Rouge through a semantic perspective, we examine the BertScores of the models. This observation demonstrates that our summaries not only align with the actual summary in terms of wording but also semantically. This indicates that our model offers superior semantic conveyance compared to other models.

We observe the *answerability* of our generated summaries through entailment, as explained in section 4.1. The entailment scores for the models are presented in table 4. These scores demonstrate the superior performance of our cross-attention model. Similar to previous findings, Q&A guidance increased the performance of our raw summarization block, underscoring the potential of question-aided summarization. It is evident that our cross-attention model generates summaries containing much more factual content as can be deduced from the entailment score.

Table 3: Evaluate the *coherence* of the models through Bert scores.

Model	BertScore
fb.-bart-cnn	0.793
Summ. Block	0.665
Cross-Atten.	**0.862**

Table 4: Evaluate the *answerability* of the models through entailment scores.

Model	Entailment Score
fb.-bart-cnn	0.652
Summ. Block	0.565
Cross-Atten.	**0.766**

5 Ethical Issues

In the task of passage summarization, it is crucial to consider maintaining fair information and avoiding the introduction of hallucinations to prevent ethical issues. Regarding hallucinations, a recent study has indicated that summarization models may reflect the dataset on which they were trained during the inference phase. They can transfer knowledge acquired from the training phase to the summary, which may not actually exist in the original passage to be summarized. This knowledge poses a danger, as it can manifest in the form of toxic language and hate speech, creating an ethical problem (Marinova et al., 2021). Addressing this issue is not only important for ethical considerations but also for enhancing summarization performance, as biases introduced by the training dataset can influence the model's generalization ability (Zhong et al., 2019).

While mitigating biases originating from the training dataset, it is equally important to ensure the fair presentation of information. Consequently, all properties should be accurately reflected regardless of the political leaning of the text or the gender/race/religion of a person mentioned in the passage. If our model diminishes the visibility of any part of the text based on the aforementioned characteristics, a significant problem arises in terms of ethical considerations (Shandilya et al., 2018). Studies attribute this problem to the lack of representational diversity in the training dataset. Researchers argue that summarization models tend to learn negative stereotypes for underrepresented samples, resulting in unintended bias (Keswani and Celis, 2021). As a solution to this issue, some researchers have proposed the utilization of the Rooney Rule, which imposes constraints on the summarization process by compelling the model to maintain a specific level of representation of diversity while generating the output (Dash et al., 2019). In our setup, our future work includes the utilization of this concept to mitigate any unintended bias.

6 Conclusion

In this paper, we investigated the summarization problem of passages by incorporating assistance from non-factoid question answering. Our focus extended beyond ensuring the coherence of the generated summary; we aimed to generate it in a manner that includes answers to potential key questions that may arise from the original text. We introduced a cross-attention-based architecture, leveraging a pre-trained T5 transformer model. Our results illustrate that our model can achieve superior performance compared to baselines in terms of Rouge, Bert, and entailment scores, emphasizing the advantages introduced by the integration of question-answering assistance.

References

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Despoina Antonakaki, Paraskevi Fragopoulou, and Sotiris Ioannidis. 2021. A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164:114006.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Multi-hop inference for question-driven summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6734–6744, Online. Association for Computational Linguistics.

et al. Hu, Edward J. 2021. Lora: Low-rank adaptation of large language models. *MLA*.

Vijay Keswani and L. Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 3802–3814, New York, NY, USA. Association for Computing Machinery.

Bilal Khan, Zohaib Ali Shah, Muhammad Usman, Inayat Khan, and Badam Niazi. 2023. Exploring the landscape of automatic text summarization: A comprehensive survey. *IEEE Access*, 11:109819–109840.

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2019. Abstractive text summarization based on deep learning and semantic content generalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5092, Florence, Italy. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Iva Marinova, Yolina Petrova, Milena Slavcheva, Petya Osenova, Ivaylo Radev, and Kiril Simov. 2021. Monitoring fact preservation, grammatical consistency and ethical behavior of abstractive summarization neural models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 901–909, Held Online. INCOMA Ltd.

Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, 28(3):345–361.

M. F. Mridha, Aklima Akter Lima, Kamruddin Nur, Sujoy Chandra Das, Mahmud Hasan, and Muhammad Mohsin Kabir. 2021. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, 9:156043–156070.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th*

SIGNLL Conference on Computational Natural Language Learning, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Aarthi Paramasivam and S. Jaya Nirmala. 2022. A survey on textual entailment based question answering. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part B):9644–9653.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Kavyashree S, Sumukha R, Soujanya R, and Tejaswini S V. 2023. Survey on automatic text summarization using nlp and deep learning. In *2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS)*, pages 523–527.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 97–98, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Sunit Singh. 2022. Question prompted passage summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2019. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 80–89, Hong Kong, China. Association for Computational Linguistics.