Label-Free Topic-Focused Summarization Using Query Augmentation

Wenchuan Mu
Information Systems Technology and Design
Singapore University of Technology and Design
Singapore
wenchuan_mu@sutd.edu.sg

Kwan Hui Lim
Information Systems Technology and Design
Singapore University of Technology and Design
Singapore
kwanhui_lim@sutd.edu.sg

Abstract—In today's data and information-rich world, summarization techniques are essential in harnessing vast text to extract key information and enhance decision-making and efficiency. In particular, topic-focused summarization is important due to its ability to tailor content to specific aspects of an extended text. However, this usually requires extensive labelled datasets and considerable computational power. This study introduces a novel method, Augmented-Query Summarization (AQS), for topic-focused summarization without the need for extensive labelled datasets, leveraging query augmentation and hierarchical clustering. This approach facilitates the transferability of machine learning models to the task of summarization, circumventing the need for topic-specific training. Through real-world tests, our method demonstrates the ability to generate relevant and accurate summaries, showing its potential as a cost-effective solution in data-rich environments. This innovation paves the way for broader application and accessibility in the field of topic-focused summarization technology, offering a scalable, efficient method for personalized content extraction.

Index Terms—Topic-focused summarization, Query augmentation, Transfer learning.

I. INTRODUCTION

In today's world, where we are constantly bombarded with vast amounts of text, the ability to efficiently summarize information has become crucial [1]. Summarization, the process of condensing extensive texts into shorter, digestible formats, is important for both individuals and organizations. It enables quicker understanding and better management of information. However, this process is not without its challenges. One major difficulty is creating summaries that are useful for different purposes and to different people, as everyone has unique needs and interests [2].

Addressing these issues, topic-focused summarization offers a new approach. It focuses on generating summaries based on specific topics, making the information more relevant to the reader's interests [3]. This task involves analyzing the content to identify key themes and then highlighting these in the summary. Yet, implementing topic-focused summarization is challenging. It typically requires a lot of organized data to start with and demands high computational power to process this information effectively [4].

A. Recent Developments and Limitations

Recent advances in automated summarization have primarily focused on abstractive and extractive techniques [5]. Ab-

stractive summarization models [1], such as those employing neural language models, attempt to generate a concise and coherent summary by understanding and paraphrasing the original text. These models, often based on complex architectures like transformers and attention mechanisms, have shown remarkable capability in mimicking human-like summary generation [5], [6]. However, they are typically dependent on extensive supervised training, requiring large labelled datasets that are not always feasible or available for specific topics or domains [7]. Furthermore, the unconstrained nature of abstractive summarization often leads to challenges in controlling the focus of the summaries, particularly in aligning with specific user queries or topics [2].

Extractive summarization, on the other hand, involves selecting relevant sentences or segments directly from the source text [8]. This approach, while simpler and more interpretable, often struggles with generating summaries that are as coherent and fluent as those produced by abstractive methods [9], [10]. Additionally, both abstractive and extractive methods face challenges in topic-focused summarization. They either require substantial customization and retraining for each new topic or fail to adequately prioritize information relevant to specific user queries [3]. This limitation becomes particularly evident in dynamic and diverse real-world applications where the ability to adapt to varying topics and queries rapidly is essential [4]. Therefore, there is a pressing need for a summarization approach that is both flexible and efficient in handling diverse topics without extensive retraining or reliance on large labelled datasets [7].

B. Our Contributions

To address the challenges of topic-focused summarization in the absence of labelled data, our proposed method, Augmented-Query Summarization (AQS), employs a novel approach that leverages the strengths of existing natural language processing techniques [11]. AQS integrates four key components: paraphrasing for query augmentation, standard question answering, hierarchical clustering, and generic abstractive summarization. This combination of techniques allows us to effectively generate topic-focused summaries by processing and analyzing the context and content of source texts through a sequence of finely tuned steps. Each component
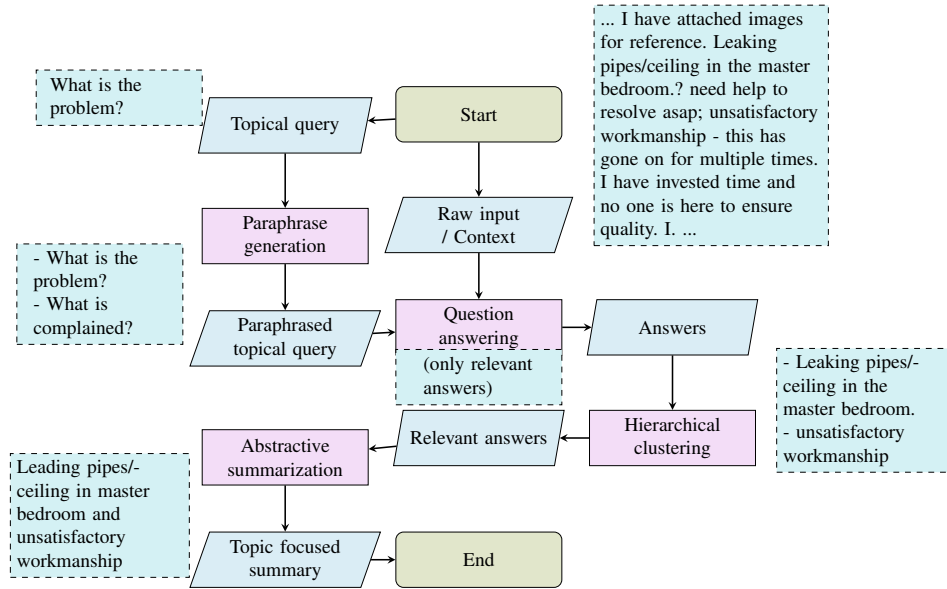
Fig. 1: The Augmented-Query Summarization (AQS) pipeline consists of four pretrained key components: a paraphrasing model, a question-answering model, hierarchical clustering, and an abstractive summarization model. AQS takes two text inputs: the query related to the topic and the typically longer context. It generates a single, topic-focused summary as its output. AQS is an adaptation approach, as all the key components can be derived from generic tasks, such as generic abstractive summarization.

of AQS plays a crucial role in ensuring that the final summary is not only contextually relevant but also concise and coherent, addressing the specific topic of interest without requiring extensive labelled datasets for training.

The practicality and efficiency of AQS are further bolstered by its design, which accommodates the integration of a range of pre-trained models [11], [12]. This flexibility not only makes our method adaptable to various domains and topics, but also significantly reduces the time and resources typically required for model training and fine-tuning. Our approach is especially advantageous in situations where rapid deployment and adaptability are essential. Furthermore, by harnessing the capabilities of pre-existing models, AQS stands at the forefront of summarization technology, offering a robust, scalable solution that promises significant improvements in personalized content extraction and information synthesis in data-rich environments. A summary of our key contributions are:

1) Investigating how query and context variations affect the transferability of question-answering models, as well as the impact of input changes on generic abstractive summarization models. 2) Introducing an algorithm that does not require training on topic-focused summarization or extensive labelled datasets. 3) Illustrating the effectiveness and efficiency of our method on real-world data through qualitative and quantitative analysis and experiments.

II. RELATED WORK

Abstractive summarization models strive to produce brief, precise, and easily understandable text that captures the most essential information from a document. In recent years, notable progress has been achieved in the field of generic abstractive summarization [9], [10], [13]. This success can be credited to the development of advanced neural architectures and the accessibility of extensive datasets [1], [14], [15]. The BART architecture [11], for example, employs transformers and implements sequence-to-sequence functionality, eliminating the need for a separate encoder-decoder structure.

Topic-focused summarization is a more challenging task that involves generating a summary specifically tailored to a given query and its relevant document(s), as illustrated in Example 1. [8] involved treating the extractive text of an extractive-abstractive model as a latent variable. [16] introduced an encode-attend-decode system incorporating query and diversity-based attention mechanisms. This approach aimed to produce a summary that is more relevant to the given query. [4] trains a relevance prediction model directly on data using the original, non-masked query. [17] ranks the sentences in context (specific body of text from which answers to the augmented queries are to be extracted and summarized) according to embedding similarity to create weak-supervising labels before end-to-end training. [18] propose a model that incorporates the explicit answer relevance of the source documents given the query via a question-answering model, to generate coherently and answer-related summaries. [19] and [20] also employed QA models to rank answer evidence at the sentence or paragraph level. However, these methods might not be able to fully harness the potential of QA models in actively retrieving highly relevant text segments from the context.

Our work leverages the power of pretrained QA models, which excel at attending to queries, with the capabilities of pretrained abstractive summarization models, which generate

coherent sentences. In addition, our method is able to combine with any neuron architecture without the training needed.

Example 1. **Context** ▷ After 30 years of being 'stuck' on the spot, the world's biggest iceberg is on the move. Called A23a, the block of ice is around 1,540 sq miles in area – more than twice the size of Greater London ...[1].

Generic summary ▷ *With an area of 1,540 sq miles, iceberg A23a is the current world record holder The largest iceberg in the world was A76 before it fragmented into three pieces*

Topical query ▷*What did Dr Fleming say?*

Topic-focused summary ▷ A structure, likely an ice shelf, stationary since 1986, began moving due to size reduction over time, with consensus suggesting natural progression rather than temperature changes as the cause.

III. PROPOSED METHOD: AUGMENTED-QUERY SUMMARIZATION (AQS)

We present Augmented-Query Summarization (AQS), which integrates four elements: paraphrase generation, question answering, hierarchical clustering, and generic abstractive summarization. The workflow of this pipelined method can be viewed in Figure 1.

In this section, we introduce how the proposed pipelined method is constructed and explain how all elements collectively offer a potential solution for topic-focused summarization without training on dedicated data.

1) Paraphrase generation: Modifications need to be made to the traditional QA pipeline. In other words, the answers should encompass more than a single segment, allowing for the consideration of information from different parts of the context to provide comprehensive responses. To this end, we resort to tweaking the question/query. The intuition is that when the query is subtly changed while retaining the main focus, we may be able to capture varied content segments. This is because each segment related to the topic can be a valid QA model output, but typically only one is produced. By introducing a degree of uncertainty or variability to the input, we anticipate a broader range of outputs. To this end, our options include varying the query, context, or both. In this study, we primarily delve into modifying the query.

We propose to augment the query through paraphrase generation. By generating multiple queries with the same meaning, we naturally generate multiple potential answers, which may originate from different locations within the context. Additionally, by perturbing the query with semantic constraints, we hypothesize that some of the query paraphrases will effectively activate the transferability of the QA model. The effect of inputting semantically similar but syntactically different queries into the QA model can be captured in Example 2 below.

---

[1]From Dailymail article "The world's biggest iceberg is on the move" https://www.dailymail.co.uk/sciencetech/article-12786619/

Example 2. **Context** ▷ The Amazon rainforest includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and minor amounts in Venezuela, Ecuador, Bolivia, Guyana, ...

Query 1 ▷ *Which country is the 3rd largest in the forest?*

Response ▷ Peru (incorrect)

Query 2 ▷ *Which country has the third largest land in the forest?*

Response ▷ Colombia (correct)

We note that a slight syntax change in the query leads to a significant improvement in extracting the correct answer. Although the original and paraphrased queries convey the same meaning, they differ in their phrasing. The improved performance can be attributed to the QA model's sensitivity to query syntax and structure. This phenomenon has also been discussed in literature about the robustness of models [21].

The paraphrased query aligns with the patterns and structures learned by the model during training, enabling more accurate identification of relevant information in the context and producing a more precise answer. Therefore, by using paraphrase generation to augment the original query, we can enhance the chances of extracting a satisfactory answer from the context. This part of the method is shown in lines 2-12 in Algorithm 1. Now that we have augmented the topic-related query, the next step is to extract answers using these queries.

2) Question answering: To address this challenge, we propose to leverage question-answering (QA) models. The intuition of using QA for topic-focused summarization is that we can use the topic to form a question and then extract the topic-related content from the given context. In an adaptation setup, QA is suitable for extracting topic-focused content from a context for the following reasons. First, QA and topic-focused summarization share a common input format, consisting of two texts: one for the query/question and the other for the context. Secondly, they have a comparable output format, both being text strings.

However, there are also challenges in using QA for topic-focused summarization, as there is a key distinction in their output constraints. In QA, the output is typically a substring extracted from the context, whereas topic-focused summarization does not impose such limitations. In other words, the (single-answer) QA task is a task of extracting a single segment from the context, whereas topic-focused summarization may expect content from many discrete places in the context. We suggest a parallel question-answering method that processes augmented queries alongside a singular context. In this approach, every query is matched with the context to generate a response. Theoretically, if a standard QA model has a probability p of producing successful answers, the efficacy of the collective response improves as the quantity of augmented queries rises, converging towards p. Moreover, if p exceeds

1/2, it is highly probable to obtain a successful answer, as depicted in Figure 2(b).

3) Hierarchical clustering: The challenge that comes along with using an augmented set of queries is effectively managing the increased complexity and diversity of the query space. When we introduce a larger set of augmented queries, it becomes crucial to ensure that the QA system can efficiently process and evaluate each query in a timely manner. The system should be able to handle the expanded query space without sacrificing performance or introducing unnecessary computational overhead. We need to determine the optimal balance between query diversity and redundancy. While generating a diverse set of augmented queries can enhance the chances of capturing different perspectives and nuances, it also increases the likelihood of introducing redundant or overlapping queries. Managing this redundancy becomes important to avoid excessive computational costs and potential confusion in the answer selection process.

Moreover, there is a need to assess the quality and relevance of the augmented queries. Not all generated paraphrases or variations may be equally useful or effective in extracting the desired information from the context. Careful evaluation and filtering techniques are necessary to identify the most promising queries that are likely to yield satisfactory answers.

Addressing these challenges requires the development of efficient query management strategies, intelligent query selection techniques, and robust evaluation mechanisms. By effectively addressing these challenges, we can harness the benefits of using an augmented set of queries to improve the overall performance and effectiveness of QA systems. Thus, we propose clustering answers into relevant and irrelevant ones, eliminating off-topic or poorly formatted answers, to improve coherence and focus in the summarized output.

Hierarchical clustering is employed in this approach, where a symmetric similarity score is calculated between each pair of answers. The two most similar answers are then combined into a longer answer. This process is repeated until the longest answer contains at least half of the original individual answers. Subsequently, the longest answer is used as input for the abstractive summarization model. Refer to lines 17-20 in Algorithm 1 for a detailed description.

4) Generic abstractive summarization: Now that we have obtained effective answers, the next step is to write a fluent overall answer. We hypothesize that a generic text-to-text model can cope with this problem. Generic abstractive summarization condenses text by generating concise summaries that capture main ideas and context. Moreover, there are abundant off-the-shelf models. The following example illustrates the impact of introducing irrelevant information into a text on the performance of abstractive summarization models. Such perturbations can lead to a decrease in the quality of the generated summaries.

Example 3. <u>Input</u> ▷ *Brief of Incident, Subject Vehicle Illegal passing the gantry place1Case 01, Gantry at place1Case 01*
Vehicle Illegal.

Generated Summary ▷ Case 01, Gantry at place1. Vehicle Illegal passing the gantry. (improved)

<u>Slightly Changed Input</u> ▷ *Thank you subject vehicle ID 4e9aM0$3N5Thank you Incident, Subject Vehicle Illegal passing the gantry placeCase 01, Gantry at place1Case 01 Vehicle Illegal, Thank you.*

Generated Summary ▷ Vehicle ID 4e9aM0$3N5Thank you Incident, (hardly readable)

However, current abstractive summarizers, especially neural models, can be sensitive to input syntax, resulting in potential errors or inconsistencies (Example 3). It is unwise to directly input answers from QA tasks into abstractive summarizers for two reasons. Firstly, answers with bad syntax or characters extracted from the original context can cause out-of-distribution issues. Secondly, redundancy in answers can propagate to the summaries if concatenated directly. Therefore, we need to use hierarchical clustering to select appropriate queries among all we obtain from paraphrase generation, which leads us to the topic-focused summary.

IV. Experiment and Result

A. Setup

a) AQS Implementation: We implement AQS with 4 established models: Pegasus [22] for paraphrase generation, BART [11] for QA, RoBERTa [12] for text embedding, and BART for abstractive summarization. These models are respectively pretrained on the QQP dataset [2], the SQuAD [23], and the CNN/Dailymail dataset [24]. While specific models are utilized in our experiments, our framework's design allows for the substitution of comparable models, such as T5 [25], to affirm its versatility. It is critical to underscore that all models are trained on generic datasets recognized for each task's benchmarking, such as SQuAD for QA and CNN/Dailymail for summarization, without fine-tuning on the same or similar datasets for the task at hand. This ensures the validity of adaptation in the proposed method. The paragraph used is shown in Figure 2(a) and they are generated using the off-the-shelf model.

b) Validation Datasets: We assess AQS on two benchmark datasets: Debatepedia [16] and QMSum [26]. Debatepedia comprises 13,719 query-context-summary triplets, specifically designed for debate summarization tasks. QMSum offers a collection of 1,808 triplets from 232 English-language meetings across varied fields, including product design, academia, and politics. AQS is further evaluated on the real-world Estate Complaints/Feedback (ECF) dataset to avoid adaptive overfitting, which contains feedback records for an estate department from 2016 to 2021. The structure of the ECF dataset samples is outlined in Table I. The evaluation utilizes the most recent data split from 2021, with 161,101 samples. For ECF Feedback, we use two fixed queries: "What is complained?" to identify

---
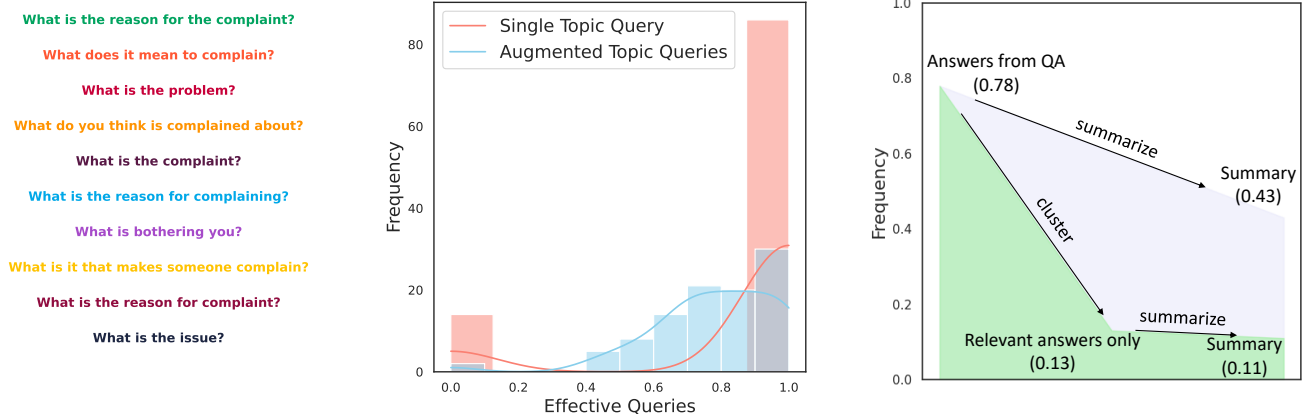
(a) Examples of paraphrased queries. These queries are paraphrased from an original query, "What is complained?"

(b) Distribution of the effective queries: Single queries lead to binary outcomes ('correct'/'incorrect'). Multiple queries produce probabilistic outcomes. Horizontal axis: the proportion of queries with correct answers (effective queries). Vertical axis: the number of documents input (%). Line curves: kernel density estimation.

(c) Text redundancy, a measure of irrelevant content in answers or summarised answers. Without clustering, answer redundancy is 0.78, leading to a summary redundancy of 0.43. With clustering, answer redundancy decreases to 0.13, resulting in a summary redundancy of 0.11.

Fig. 2: Illustration of the functionality of each part in the proposed method. Paraphrasing generates multiple queries from a single one. Using multiple queries may stabilise the QA model performance. On some inputs, a single query fails to yield a correct answer, while using multiple queries may yield, say 70%, correct answers. When the majority part of answers are correct, then the answer can be considered as correct. Note that a correct answer from QA might also contain irrelevant content, *e.g.*, function words. Clustering significantly reduces redundancy in QA. This is likely because some incorrect answers are pretty long, and thus account for a large proportion of content. Therefore, removing these incorrect answers before summarization is needed. Besides, summarization helps reform a fluent sentence from the clustering-selected answers. Intuitively, the text redundancy in the summaries is even lower.

TABLE I: Estate Complaints/Feedback (ECF) Dataset Demo with five fields: Subject, Description, Category, Sub Category, and Record Key. It includes 13 categories and 64 subcategories. The Subject field often lacks or is poorly written, necessitating topic-focused summarization.

Unique Case Record Key	Subject	Reporting Category	Reporting Sub Category	Description
1***81	FW: Vehicle Illegal Bypassing the Gantry at *place*-Case 01	Illegal Parking	Illegal Parking - Public *org* Car Parks/Service Roads	*date* *time* PM To ... Subject Vehicle Illegal Bypassing the Gantry at *place* 01 á Dear *name* á Brief of Incident á On the ... á á Please see the document as attached á á Thank You á Best Regards ... *org* Pte Ltd *org*

the main issues, and "What is the emotion?" to determine sentiment tones.

c) Metrics: ROUGE [27] is the standard for summarization model assessment. This work utilizes (F1) ROUGE-1, 2, and L for automatic model evaluation on QMSum and Debatepedia. For the ECF dataset, we measure ROUGE and Bert-Score [28] with reference to the Reporting Sub Category. A sentiment analysis model[3] compares original and summary sentiment, using Pearson's correlation. Two postHuman Annotators also conduct manual evaluations on a random 50-sample subset from ECF's 2021 data.

[3]Sentiment analysis pipeline from Hugging Face https://huggingface.co/blog/sentiment-analysis-python

d) Baselines: We benchmark our AQS against 1) Pre-trained transformers like T5 [25] which are conventionally adapted for topic-focused summarization at inference; 2) Transformers fine-tuned on combined queries and contexts, suitable for supervised or adaptation modes; 3) Other label-free models, such as CAiRE-COVID, originally design for shortlisting COVID-19 information [20]; 4) Advanced end-to-end methods such as QFS-BART [29] or SEGENC [4]. This method can also be employed in either supervised or adaptation mode. For instance, SEGENC can be trained on QMSum and then validated on Debatepedia or ECF dataset.

B. Result

a) Effectiveness: Our primary goal is to assess how well a method performs on a new dataset it has not seen

Algorithm 1 Topic-Focused Summarization $(\mathbf{x}, \mathbf{z}, n \in \mathbb{N}_+), q \in [0, 1)$

Require: $\mathbf{x}$: Topic inquiry, $\mathbf{z}$: Input text, n: Beam size, q: patience factor

1: $\mathcal{H}_{cur} \leftarrow \{(\epsilon, 0.0)\}$ $\triangleright$ Initialize with empty generation prefix and zero score
2: **repeat**
3:     $\mathcal{H}_{next} \leftarrow \emptyset$
4:     **for** $(\mathbf{y}, p) \in \mathcal{H}_{cur}$ **do**
5:        **if** $y_{|\mathbf{y}|} = </\mathbf{s}>$ **then**
6:           $\mathcal{H}_{next} \leftarrow \mathcal{H}_{next} \cup \{(\mathbf{y}, p)\}$ $\triangleright$ Hypotheses ending with $</\mathbf{s}>$ are not expanded
7:        **else**
8:           $\mathcal{H}_{next} \leftarrow \mathcal{H}_{next} \cup \bigcup_{w \in \mathcal{T}}(\mathbf{y} \cdot w, p + \log P(w|\mathbf{x}, \mathbf{y}))$ $\triangleright$ Add all possible continuations
9:        **end if**
10:    **end for**
11:    $\mathcal{H}_{cur} \leftarrow \{(\mathbf{y}, p) \in \mathcal{H}_{next} : |\{(\mathbf{y}', p') \in \mathcal{H}_{next} : p' > p\}| < n\}$ $\triangleright$ Select n-best
12: **until** $\forall(\mathbf{y}, p) \in \mathcal{H}_{cur}. \ y = </\mathbf{s}>$
13: $\mathcal{H}_{seg} \leftarrow \emptyset$
14: **for** $(\mathbf{y}, p) \in \mathcal{H}_{cur}$ **do**
15:    $\mathcal{H}_{seg} \leftarrow \mathcal{H}_{seg} \cup \{ Q(\mathbf{y} \cdot \mathbf{z}) \}$ $\triangleright$ Each answer is a group
16: **end for**
17: **repeat**
18:    $(\mathbf{u}, \mathbf{v}) \leftarrow \arg\min_{i,j \in \mathcal{H}_{seg}} d(i, j)$
19:    $\mathcal{H}_{seg} \leftarrow \mathcal{H}_{seg} \cup \{ \mathbf{u} \cdot \mathbf{v} \} \setminus \{ \mathbf{u}, \mathbf{v} \}$ $\triangleright$ Merge the two most similar groups
20: **until** $\max_{\mathbf{u} \in \mathcal{H}_{seg}} |\mathbf{u}| > q \sum_{\mathbf{u} \in \mathcal{H}_{seg}} |\mathbf{u}|$ **return** $S(\arg\max_{\mathbf{u} \in \mathcal{H}_{seg}} |\mathbf{u}|)$ $\triangleright$ Summarize the largest group

---

TABLE II: ROUGE-F1 scores are presented for the QMSum and Debatepedia datasets. The previous work can be classified into three categories: 1) pre-trained transformers on a generic summarization dataset, 2) fine-tuning of these transformers on a query-focused summarization dataset, and 3) models specifically designed for supervised training in query-focused summarization.

Approach (Train set)	Validate on QMSum			Validate on Debatepedia		
	R-1	R-2	R-L	R-1	R-2	R-L
BART (CNN/Dailymail)	31.87	9.08	27.50	37.03	27.63	36.77
T5 (CNN/Dailymail)	32.45	**9.80**	28.48	39.72	29.33	37.37
CAiRE-COVID	30.98	4.70	25.80	13.32	2.47	12.18
BART (QMSum)	32.42	9.62	28.37	*42.11*	**38.04**	*41.42*
T5 (QMSum)	34.24	9.30	28.91	40.96	37.61	38.71
SEGENC (QMSum)	37.80	13.43	33.38	*41.50*	*38.02*	*39.83*
QFS-BART (QMSum)	36.07	10.89	34.65	38.94	36.68	39.63
BART (Debatepedia)	31.93	8.33	27.27	57.96	44.09	57.40
T5 (Debatepedia)	31.38	8.39	27.85	53.45	40.39	52.42
SEGENC (Debatepedia)	*33.77*	*9.51*	*31.86*	54.41	38.49	51.18
QFS-BART (Debatepedia)	*33.30*	8.94	30.76	59.02	44.59	57.44
AQS (**Label-free**)	**33.84**	*9.79*	**32.38**	**46.07**	*37.98*	**44.43**

The best adaptation method is indicated in **bold**, second and third *italics*, and shaded cells represent different splits of the same dataset, not demonstrating adaptation behaviour.

during training. This 'different dataset' is distinct from merely splitting the training and validation data. As shown in Table II, AQS outperforms other methods on both the QMSum and Debatepedia datasets, in terms of summary effectiveness. The summary effectiveness is measured by the closeness [27] between the generated summary and the reference in the test dataset. In contrast, when generic summarization models are used for query-focused summarization, their effectiveness is notably limited. Similarly, dedicated query-focused summarization methods find it challenging to adapt to shifts in data distribution between training and targeted test data. In short, AQS produce well-formed summaries compared to baseline methods on unseen topic-focused summarization tasks.

b) Efficiency: AQS, featuring four transformer models with approximately 380 ± 100 M parameters each, emphasizes efficiency. It can execute sequentially (or parallel), sidestepping processing bottlenecks, and achieves rapid performance, processing 183 samples per minute on a single NVIDIA Tesla T4 GPU. In contrast, QFS-BART requires double the GPU memory due to larger model dimensions, while SEGENC, over ten times larger than AQS, demands a high-end A100 GPU. AQS represents a significant step forward in optimizing the balance between model complexity and operational efficiency.

c) Use Case: Since the ECF dataset lacks reference summaries, the Reporting Sub Category serves as a reference. While exact matches are not always necessary between

TABLE III: ROUGE-F1 scores and Bert-Score are reported for the real-world ECF dataset. Moreover, we present the sentiment consistency (CNS) and human satisfactory score, as described in Section IV-A. All the methods listed are applied in adaptation mode, *i.e.*, with no training or fine-tuning on the tested distribution.

Approach	R-1	R-2	R-L	Bert-Score	Sentiment CNS.	Human Eval
BART (CNN/Dailymail)	4.23	2.02	3.84	82.17	0.211	1.95
T5 (CNN/Dailymail)	3.51	1.47	2.33	84.62	0.219	0.63
CAiRE-COVID	11.97	2.23	11.78	86.56	0.672	4.03
BART (QMSum)	10.91	3.45	8.74	82.48	0.607	2.48
T5 (QMSum)	12.78	3.92	9.58	84.93	0.699	3.87
SEGENC (QMSum)	15.87	4.25	*12.64*	*88.56*	0.493	*4.39*
QFS-BART (QMSum)	13.03	4.29	12.61	87.88	0.515	2.01
BART (Debatepedia)	2.71	0.63	1.17	86.47	0.528	0.42
T5 (Debatepedia)	4.35	1.04	2.98	81.22	0.716	3.57
SEGENC (Debatepedia)	*16.23*	*5.92*	11.09	86.79	*0.745*	4.15
QFS-BART (Debatepedia)	10.56	4.78	9.26	82.97	0.453	1.33
AQS (**Label-free**)	**21.94**	**6.37**	**15.68**	**88.75**	**0.807**	**4.82**

The best adaptation method is indicated in **bold**, the runner-up in *italics*

TABLE IV: Ablation study on the components of the proposed AQS. In rows 2 to 4, we made the following modifications: a) replacing the pretrained transformer from BART to T5, b) removing the clustering step, and c) eliminating the query paraphrasing step. This table shows what are the factors influential/non-influential to our method.

Approach	R-1	R-2	R-L	Bert-Score	Sentiment CNS.	Human Eval
AQS	**21.94**	*6.37*	**15.68**	**88.75**	*0.807*	**4.82**
AQS - BART + T5	*19.40*	**6.38**	*13.29*	*88.60*	**0.833**	*4.40*
AQS - Clus.	6.47	1.22	2.39	79.45	0.203	1.57
AQS - Para.	1.51	0.17	1.23	81.43	0.208	0.47

The best adaptation method is indicated in **bold**, the runner-up in *italics*

sub-category names and predicted summaries, assessing semantic similarity can gauge model performance. When a summary closely aligns with its sub-category name, it indicates better performance. Similarly, sentiment preservation evaluation matches the predicted summary with the original sentiment. As shown in Table III, AQS generates highly favourable summaries for Descriptions, benefiting from its use of well-pretrained transformers with strong transferability from generic task training.

d) Ablation: We explore how AQS effectively combines pre-trained transformers for topic-focused summarization. The ablation study presented in Table IV examines the performance when gradually removing specific components of AQS. The results reveal that a significant drop in performance occurs if either the clustering or paraphrasing part is excluded. Without clustering, the presence of redundant text from the answers can dilute the relevant information, leading to off-topic summaries. Similarly, without paraphrase generation, relying solely on a QA model results in a single, often short answer, rendering the presence or absence of downstream models inconsequential. Figure 2 visually demonstrates the impact of AQS's paraphrase module and clustering module on the relevance of information at each stage of text processing.

V. CONCLUSION

This study investigates how pretrained neural models can be adapted to create summaries focused on specific topics, especially when there are no labelled examples of such summaries available. Our approach involves using paraphase-based query augmentation, question answering, and hierarchical clustering to shift from generic summary generation to topic-specific summarization. This method does not need data specifically labelled for topic-focused summarization. Instead, it uses four distinct elements that can be developed separately and thus benefit from a broader range of data, such as that used for generic abstractive summarization. The effectiveness of this method is validated on an existing test set for topic-focused summarization. Besides, this strategy presents a viable option for real-world applications without demanding substantial computational resources.

REFERENCES

[1] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 379–389.

[2] S. A. Bahrainian, S. Feucht, and C. Eickhoff, "NEWTS: A corpus for news topic-focused summarization," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 493–503.

[3] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary, "Topic-focused multi-document summarization using an approximate oracle score," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia: Association for Computational Linguistics, Jul. 2006, pp. 152–159.

[4] J. Vig, A. Fabbri, W. Kryscinski, C.-S. Wu, and W. Liu, "Exploring neural models for query-focused summarization," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1455–1468.

[5] E. Chu and P. Liu, "MeanSum: A neural model for unsupervised multi-document abstractive summarization," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1223–1232.

[6] P. Li, W. Lam, L. Bing, W. Guo, and H. Li, "Cascaded attention based unsupervised information distillation for compressive summarization," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2081–2090.

[7] H. Yu, "Survey of query-based text summarization," 2022.

[8] Z. Mao, C. H. Wu, A. Ni, Y. Zhang, R. Zhang, T. Yu, B. Deb, C. Zhu, A. Awadallah, and D. Radev, "DYLE: Dynamic latent extraction for abstractive long-input summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1687–1698.

[9] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083.

[10] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, "Unified language model pre-training for natural language understanding and generation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 13 042–13 054.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[13] S. Gehrmann, Y. Deng, and A. Rush, "Bottom-up abstractive summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4098–4109.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[16] P. Nema, M. M. Khapra, A. Laha, and B. Ravindran, "Diversity driven attention model for query-based abstractive summarization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1063–1072.

[17] Y. Xu and M. Lapata, "Generating query focused summaries from query-free resources," 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 6096–6109.

[18] Z. Su, Z. Tang, X. Guan, L. Wu, M. Zhang, and J. Li, "Improving temporal generalization of pre-trained language models with lexical semantic change," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6380–6393.

[19] Y. Xu and M. Lapata, "Coarse-to-fine query focused multi-document summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3632–3645.

[20] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. Barezi, and P. Fung, "CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, Dec. 2020.

[21] S. Garg and G. Ramakrishnan, "BAE: BERT-based adversarial examples for text classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6174–6181.

[22] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 11 328–11 339.

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392.

[24] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019.

[26] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. Radev, "QMSum: A new benchmark for query-based multi-domain meeting summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5905–5921.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," *CoRR*, vol. abs/1904.09675, 2019.

[29] D. Su, T. Yu, and P. Fung, "Improve query focused abstractive summarization by incorporating answer relevance," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3124–3131.