PREME: Preference-based Meeting Exploration through an Interactive Questionnaire

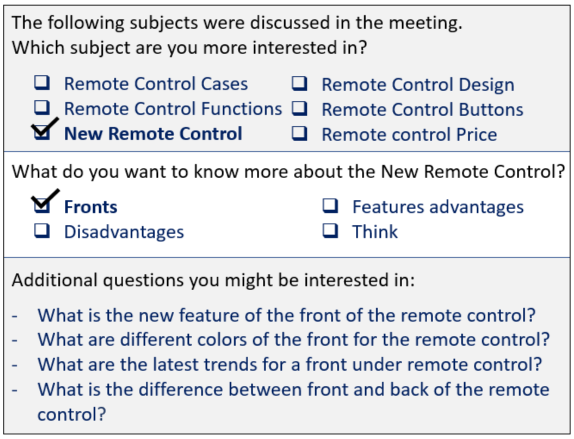
Anonymous ACL submission

Abstract

The recent increase in the volume of online meetings necessitates automated tools for managing and organizing the material, especially when an attendee has missed the discussion and needs assistance in quickly exploring it. In this work, we propose a novel end-to-end framework for generating interactive questionnaires for preference-based meeting exploration. As a result, users are supplied with a list of suggested questions reflecting their preferences. Since the task is new, we introduce an automatic evaluation strategy. Namely, it measures how much the generated questions via questionnaire are answerable to ensure factual correctness and covers the source meeting for the depth of possible exploration.

1 Introduction

In recent years, video conferencing technology has gained substantial improvements, and thus, online meetings have become easily accessible and more prominent. Primarily due to the pandemic and work from home, the need for video calling has grown significantly. For example, the number of meeting minutes held in the Zoom applicatifon has increased by 3300% in 2021 compared to the same quarter of the previous year[1]. Therefore, the high volume of online meetings necessitates automated tools for managing and organizing essential information for the attendees. Especially when an attendee has missed an online meeting, it is critical to quickly access required information since reading through the transcript is quite time-consuming.



The following subjects were discussed in the meeting.
Which subject are you more interested in?

☐ Remote Control Cases
☐ Remote Control Functions
☑ New Remote Control
☐ Remote Control Design
☐ Remote Control Buttons
☐ Remote control Price

What do you want to know more about the New Remote Control?

☑ Fronts
☐ Disadvantages
☐ Features advantages
☐ Think

Additional questions you might be interested in:

- What is the new feature of the front of the remote control?
- What are different colors of the front for the remote control?
- What are the latest trends for a front under remote control?
- What is the difference between front and back of the remote control?

Figure 1: An example of exploring one of the meetings from the collection (Carletta et al., 2005) based on user preferences through an interactive questionnaire. Users may exploit the questionnaire multiple times to explore various parts of the meeting.

Providing meeting summaries is a promising direction (Wang and Cardie, 2013; Jacquenet et al., 2019; Zhao et al., 2019; Singhal et al., 2020). However, recent works (Murray et al., 2010; Mehdad et al., 2013; Li et al., 2019) have demonstrated that approaches designed for document summarization could not effectively apply to meetings transcripts due to the following potential reasons: **(R1) Structure:** standard documents are well structured compared to meeting transcripts; **(R2) Language:** spoken language used in meetings is less regular than documents; and **(R3) Multiple speakers:** the speaker role is essential. Moreover, there is little meeting data publicly available that can be used for experimentation compared to regular documents such as news or articles. In contrast with document summarization, when summarizing a meeting, different users tend different preferences on what content should be included in the summary. Recently, Zhong et al. (2021) attempted to tackle this problem by proposing a query-based multi-domain meeting summary, where a user provides a query in question form, e.g., *'What was the discussion about the jog dial's function when talking about changes in the current design?'* to locate the part of the transcript that related to the query and then summarize. However, when attendees have missed the meeting, they cannot formulate such questions due to no prior knowledge about the meeting. To overcome this, we aim to address the following **research challenge**: *How can attendees effectively explore a meeting content without having prior knowledge about it?*

[1] https://investors.zoom.us/

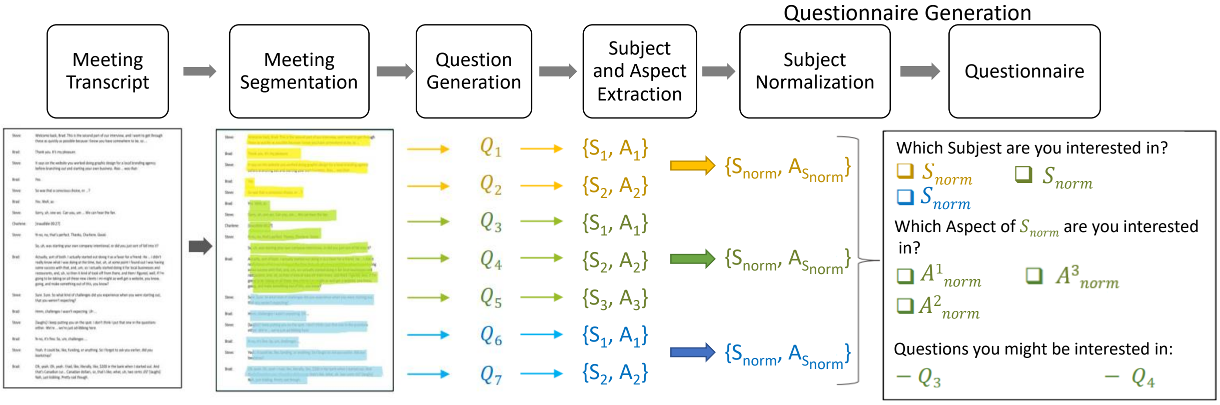


Figure 2: Overview of our framework, Preference-based Meeting Exploration through an Interactive Questionnaire (PREME), where Q is a comprehensive set of questions, and S_i and A_j are extracted pairs of subjects and aspects.

This work is motivated by the fact that asking questions is a more efficient way for humans to acquire information than notes in plain text (Lawson et al., 2007, 2006). Thus, we address preference-based meeting exploration by automatically generating a structured interactive questionnaire for a transcript that covers most of the discussed topics and quickly walks users through the discussed content. An example of the desired questionnaire is shown in Fig. 1[2]. First, the user has the ability to express their preferences regarding *subjects* that have been discussed (Solbiati et al., 2021; Huang et al., 2018; Zhang and Zhou, 2019; Sehikh et al., 2017). Next, the questionnaire interactively suggests narrowing down their exploration if possible by displaying a list of possible related *aspects*. As a result, a ranked list of questions reflecting user preferences is generated. Next, the user can pick a question that demonstrates their seeking needs the most and is redirected to the meeting part containing an answer. Interactively asking for preferences in the questionnaire is beneficial because the user oversees what has been covered during the meeting they have missed. Hence, the goal of proposed questionnaires is two-fold: **(G1)** to compactly represent the discussed content; **(G2)** to guide users to form questions that express their preference regarding the transcript. We require the generated questionnaire to satisfy the following properties:

P1 *Coverage:* coverage is the amount of the information from the source text that a questionnaire points to. The generated questionnaire must cover the meeting as much as possible;

P2 *Answerable:* a given meeting transcript should contain the answers to the questions generated as a result of the questionnaire.

To address the defined challenge, we propose a framework, PREME, which consists of several concrete sequential steps highlighted in Fig. 2. We start by enchaining the method to extract meeting segments (Solbiati et al., 2021). Due to the conversational nature of the meeting, topic detection from the segments is challenging (Huang et al., 2018; Zhang and Zhou, 2019; Sehikh et al., 2017). Thus, we indirectly extract the topics as follows. First, we generate questions from each segments (Brown et al., 2020) since extracting topics from the questions is much more well studied. Further, we employ a trained Conditional Random Field (CRF) model to tag subjects and aspects (Fig. 1) from generated questions originated from each segments (Wallach, 2004). Once we got each segment's topic list, we proposed a strategy to normalize them to reduce the number of options in the questionnaire. Recently, Deutsch et al. (2020) demonstrated that QA-Based evaluation is strongly correlated with human opinion. Thus, to evaluate PREME, we employ a similar QA-based strategy.

To summarize, the main contributions are:

C1 We propose PREME, a novel framework to enable meetings exploration based on user's preferences through an interactive questionnaire;

C2 We propose a new method for subject normalization which returns the most informative subject from a set of phrases and keywords;

C3 We introduce a new automatic evaluation strategy for measuring the effectiveness of the proposed questionnaire to assess the required properties **P1** and **P2**, which according to (Deutsch et al., 2020) has a strong correlation with human judgments; and

C4 We open-source a dataset that includes 1000 questions comprehensively annotated with subject to their subjects and aspects.

[2] As a sanity check we interviewed a number of professionals if they find such application useful for their daily job. The responses were all positive.

2 Related Work

2.1 Automatic Textual Summarization

Automatic text summarization task has attracted lots of attention across Natural Language Processing (NLP) community recently. Many systems are proposed to summarize documents in different domains, including news (Rush et al., 2015; Nallapati et al., 2017; See et al., 2017; Celikyilmaz et al., 2018; Liu and Lapata, 2019; Zhang et al., 2020), academic papers (Manakul and Gales, 2021; Huang et al., 2021) and books (Kryściński et al., 2021). Meeting summarization has also emerged as a widespread need recently. Due to the unique discourse structure of dialogues, conventional document summarization systems are facing challenges when summarizing meetings (Li et al., 2019; Zhu et al., 2020). Thus, new models are proposed for tackling this task. Wang and Cardie (2013) employ decisions, action items in dialogues to progressively generate the summary. Oya et al. (2014) propose a template-based meeting summarization system by learning the relationship between summaries and their source meeting transcripts. Shang et al. (2018) design an unsupervised meeting summarization model with multi-sentence compression techniques. Li et al. (2019) introduce multi-modal information into meeting summarization with a hierarchical attention mechanism. Zhu et al. (2020) propose a hierarchical meeting summarizer that can process both word-level and turn-level information of dialogues. Furthermore, it comes into sight of the community that, due to the lengthy content and distributed information, a general summary of the meetings does not necessarily satisfy what users are seeking. Thus, Query-based summarization methods become more prevailing in which the summaries are specifically and concisely generated according to user queries (Litvak and Vanetik, 2017; Nema et al., 2017; Baumel et al., 2018; Ishigaki et al., 2020; Kulkarni et al., 2020, 2021; Pasunuru et al., 2021). Recently, Zhong et al. (2021) propose a new framework of query-based summarization for meetings, in which they annotate QMSUM, a query-based multi-domain meeting dataset. Each QMSUM meetings come along with a set of queries with different levels of abstractness, i.e., general queries and specific queries. Human annotators write these queries and the summaries aligned with these queries after reading the meeting transcripts.

While query-based summarization can be a proper path to provide users with meeting information at different specificity levels, we argue that issuing such specific queries still requires a certain degree of background knowledge. In real-life scenarios, users might not be equipped with that knowledge and issue informative queries, especially when they did not attend the meeting. Hence, they can not benefit from query-based summarization techniques to explore the meetings. We address the drawbacks of query-based summarizers by providing users with an interactive questionnaire which provides them with potential queries and allows them to explore the meetings more flexibly.

2.2 Evaluation of Summaries Factuality

The summaries often has called out for hallucination issues (Maynez et al., 2020). Thus, Wang et al. (2020) propose a framework to evaluate factual consistency of summaries with the source text. Their intuition is that the summary and the source should similarly and consistently answer the factual questions about the context. Similarly, Deutsch et al. (2020) propose a Question Answering (QA)-based evaluation approach on summaries' content quality. They measure how much information is contained in a candidate summary by calculating the proportion of questions it can answer. These approaches inspirited our way of thinking about automated end-to-end evaluations of the questionnaires.

2.3 Question Generation and Filtering

Initial works in Question Generation task leveraged crowd-sourcing or rule-based methods to generate pre-defined question templates (Mostow and Chen, 2009; Rus et al., 2010; Lindberg et al., 2013; Fabbri et al., 2020; Mazidi and Nielsen, 2014; Labutov et al., 2015). Heilman and Smith (2010) tackled this problem in a different manner by over-generating candidate questions and then using a learning to rank framework to rank them. Ranking the questions helped filter the low-quality questions as they would rank lower. SQUASH (Krishna and Iyyer, 2019) is one of the recent works in which authors used question generation methods to convert a document into a hierarchy of question-answer pairs with the focus on questions' granularity level. They employed a neural encoder-decoder model trained on three reading comprehension data sets, i.e., SQuAD (Rajpurkar et al., 2016), QuAC (Choi et al., 2018), and CoQ (Reddy et al., 2019) to generate the questions, and further, they filtered out the unanswerable questions using some heuristics and question answering models. While question generation using question answering data sets seems a general approach, this method does not work well on meeting-related questions generated due to many reasons, including: **(1)** Different structure of meetings compared to documents; **(2)** There is not many question-answering datasets available from meetings; **(3)** Sometimes, the answer to questions generated from meetings could be very long, making it hard to fit the context in neural models. In our work, we introduce an automatic method that can generate questions regarding the meeting to overcome the high price of collecting with annotators.

2.4 Questionnaire Organization

Obtaining users preferences has always shown to be a challenging task (Jiang et al., 2008; Rokach and Kisilevich, 2012; Anava et al., 2015; Christakopoulou et al., 2016; Sepliarskaia et al., 2018). The task becomes more challenging when we aim to minimize the number of interactions with users to get to know their preferences. For example, in (Sepliarskaia et al., 2018), the authors reformulate this task as an optimization problem. They propose a static questionnaire by choosing a minimal and diverse set of questions to solve the cold start problem in recommender systems. Similarly, in Liu et al. (2019) proposed a dynamic questionnaire generation method for search of clinical trials. Quiz-style question generation has also been explored recently by Lelkes et al. (2021). The authors have formulated the problem as two sequence to sequence tasks, including the question-answer generation step and incorrect answer generation step. We argue that while the former step seems relevant to our work, it could not be adapted to meeting transcripts since their proposed dataset has been trained on factual question answering data sets and cannot be used for meeting purposes. All in all, we can conclude that creating questionnaires are still under exploration in different domain. Hence, our effort in organizing a questionnaire, especially for meetings, is timely and useful for future research.

3 Proposed Framework: PREME

This section explains PREME, our proposed novel methodology to explore meetings based on users' preferences through an interactive questionnaire. An overview of our methodology is shown in Fig. 2 in which we first apply a topic segmentation method (Solbiati et al., 2021) on meeting transcript to retrieve segments with different topics (Section 3.1). Then, we generate a set of all possible questions from each segment (Section 3.2). Further, we extract the most informative part of the questions, i.e., the subject and aspect of each question (Section 3.3). In the last step, we map the normalized subjects and aspects with generated questions and form the questionnaire (Section 3.4).

3.1 Meeting Segmentation

A meeting transcript can be extremely long and contain discussions of various topics.Therefore, our goal is to divide the meeting text into a sequence of topically coherent chunks. Thus, we adopted an unsupervised topic segmentation method based on the contextualized presentation of meeting (Solbiati et al., 2021). In this topic segmentation method, the authors compute the BERT embeddings for every utterance of the meeting transcript. Further, they curated blocks of utterances and performed a block-wise max-pooling operation to generate contextualized embedding for each block. Then, the semantic similarity between two adjacent blocks is captured, and a change in the topic is detected if two adjacent blocks show similarity below a certain threshold. This approach has several advantages, including: **(1)** It is unsupervised; **(2)** Since we are just converting the meeting into smaller pieces, and we are not losing any part of the meeting.

3.2 Question Generation

For question generation from a segment, we leveraged the powerful GPT-3 model (Brown et al., 2020).[3] An impressive capability of the GPT-3 is to generate very realistic results from few training samples or even no training sample (few-shot and zero-shot learning). The variety of the generated content can be controlled using a temperature hyper-parameter. To expand the size of generated questions' pool as much as possible, in each segment, the API is called in a zero shot learning model with different temperature values between [0-1] with a 0.05 margin, where the value closer to 1 means more diversified questions. We set the maximum output length to 128 tokens and then we repeat the process for 10 trials for each specific temperature. Given that the maximum context window for the API was 2048 tokens, we truncate and slide by half-a-window size of 2048 tokens when-

[3] GPT-3 is a large autoregressive Transformer-based language model developed by OpenAI, with 175 billion parameters. We employed the model through API calls from https://beta.openai.com/

Table 1: Examples of annotated questions with their subjects and aspects for a product meeting from (Carletta et al., 2005). Subjects are highlighted in red and Aspects are highlighted in green.

Q1	What is the arrow symbol on the remote control for?
Q2	What are the main frustrations people have with the remote control?
Q3	How will the logo and color scheme be incorporated into the product?
Q4	What are pros and cons of having a remote with a large number of buttons?
Q5	What is the most difficult part of the project from the industrial engineer's point of view?

ever a segment includes more than 2048 tokens. As a results, A list of questions is extracted based on random initialization in each API call, meaning different results are achieved even with the same hyper-parameters. We extracted five questions on average per segment in each call. Finally, a union across all runs is used to form our question pool.

3.3 Subject and Aspect Extraction

Every of the generated questions has one or more *subject(s)* that is defined as the principal matter that attendees have discussed, i.e., the main concern of the questions. In addition, some questions might point to a specific *aspect(s)* of the subject which is defined as the mentioned details about a given subject. We aim to extract the primary subjects from any question and the detailed aspect if it is mentioned. Table 1 shows examples of annotated *subjects* and *aspects* for a few questions. For instance, in the question *"What is the arrow symbol on the remote control for?"*, "remote control" is annotated as the subject and the "arrow symbol" is the specific aspect of the subject. In order to extract the subjects and aspects from the questions, we use CRF (Wallach, 2004). We examined SOTA keyword extraction and contextualized neural embedding-based topic extraction models; however, the CRF model which uses word identity, word suffix, word shape and word POS tags as features, seems to work the best among them. To train the CRF model, we were required to have annotated questions with subjects and aspects labels. We designed an annotation study using the UHRS[4] crowd-sourcing platform, where we carefully trained annotators with detailed instructions to label 1000 questions with their subject and aspects[5]. Each question has been assigned to two annotators, and we report the agreement rate between annotators in Section 4. Further, we employ the trained CRF model to extract subjects and aspects from the questions.

3.4 Questionnaire Generation

Given a meeting transcript, for each of its segment T which was initially supposed to coherently point out one subject, we generate Q_T, a set of generated questions from T. Further, We create a set S_{Q_T} by extracting the subjects from each question in Q_T. Therefore, for the segment T, we have at least $|Q_T|$ number of subjects. Extracted subjects from a question set with the same origin segment must be normalized so that one comprehensive, general, and informative subject presents a segment. The more the selected subject representative covers other concepts in S_{Q_T}, the better normalization we employed. This subject normalization reduces the number of subjects shown to the user at the first step of the questionnaire and will decrease the user's effort, causing figuring out users' preferences by asking them the minimum number of questions. In other words, our goal is to select a single subject S_{norm} from S_{Q_T} which represents S_{Q_T}in the most informative way. To do so, we define the notion of the subject network as follows.

Definition 3.1. Given a segment T, a set of generated questions Q_T, and extracted subjects S_{Q_T}, a subject-network for $G(S_{Q_T})$ is denoted as $G(S_{Q_T}) = (\mathbb{V}, \mathbb{E}, w)$. It is a weighted undirected graph, where $\mathbb{V} = \{s_i \in S_{Q_T}\}$, and $\mathbb{E} = \{e_{s_i}, e_{s_j} : \forall s_i, s_j \in \mathbb{V}\}$. The function $w : \mathbb{E} \rightarrow [0, 1]$ is the cosine similarity between the semantic relatedness of the contextualized embedding vectors of two incident subjects of an edge e_{s_i,s_j}, i.e., v_{s_i} and v_{s_j}.

In Def. 3.1, we propose a subject-network where subjects are connected, and edge weights represent the semantic similarity between the two subjects.We hypothesize that the node with highest similarity and connection to others is the most central one. In other words, since it has great similarity to other subjects, there is a high probability that it points to a more generic concept and that covers the other subjects. Hence, the node S_{norm} should have high centrality attribute to represent the main subject of segment S. We employed PageRank (Haveliwala, 2003) value to find the most im-

[4] https://prod.uhrs.playmsn.com/uhrs/

[5] We invested in having a few well-trained annotators rather than having a high number of annotators who have not been trained well. Thus, annotators were paid hourly and by the quality of their work and they had no intentions for cheating.

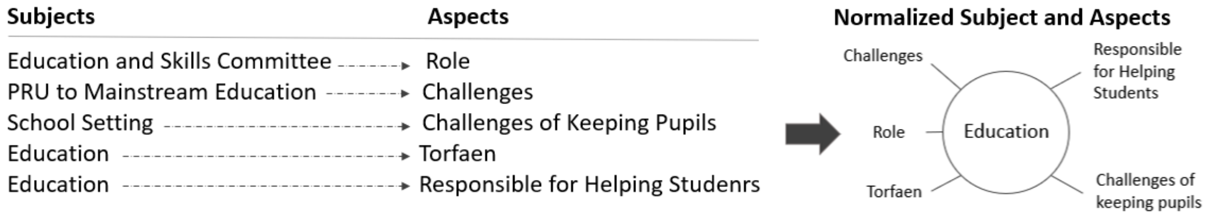

Subjects
Aspects
Education and Skills Committee
Role
PRU to Mainstream Education
Challenges
School Setting
Challenges of Keeping Pupils
Education
Torfaen
Education
Responsible for Helping Studenrs
Normalized Subject and Aspects
Challenges
Responsible for Helping Students
Role
Education
Torfaen
Challenges of keeping pupils

Figure 3: An example of how extracted subjects and aspects from a given segment are normalized.

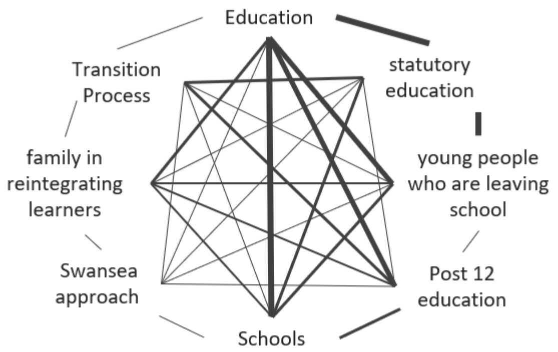


Figure 4: An example of subject-network built for one extracted segments from (Janin et al., 2003). Here, the edge weights is related to the semantic similarity between each nodes and edges with higher weights are shown with higher width.

portant and informative node in this network. Similarly, PageRank has shown to have a high correlation with the most important nodes and has been used in tackling different tasks such as quantifying term's specificity or ranking problems in different information retrieval tasks (Arabzadeh et al., 2020, 2019; Kurland and Lee, 2010). We measure the PageRank score of each node and select the node with the highest PageRank value as the representative subject S_{norm} of the subject set S_{Q_T} for segment T. In other words, we represent each segment T by subject S_{norm} where $PageRank(S_{norm}) > PageRank\ (s_i)$ for every $s_i \in \mathbb{V}$.

Fig. 4 displays a subject-network generated from extracted subjects from one of the meetings' segments in the QMSUM dataset. subjects such as "Education", "Schools," "Young people who are leaving school" are included in this subject set and represented by nodes in this subject-network. Further, we connect every pair of nodes in this graph, and the edge weight is directly related to their semantic similarity. As presented in Fig. 4, some nodes have higher edge weights which their connected lines are shown with greater width. We measure page rank in this weighted network. Here "Education" got the highest PageRank value in this subject-network. Hence, we present these subjects by one subject, i.e., "Education". "Education" can be a promising representative for these subjects as it covers more specific concepts such as "schools", "statutory education," and "post 12 education."

Next, the extracted aspects from each question set should be mapped to their representative subject. We remove the redundant and repetitive aspects and subjects by removing those who have highly similar n-grams. Plus, There might be several subjects existing in S_{Q_T} which all point out to S_{norm}, and they might be semantically very similar. In this step, we must be concerned not to lose any aspect because of subject normalization. We aim to map every aspect from S_{norm} and every s_i in S_{Q_T} which is highly similar to S_{norm} to maximize the potential of questions we might want to show at the end of the questionnaire. For instance, in Fig 3 we display a few extracted subjects and aspects from one segment. If we only consider "education" and its related aspect, we will lose many aspects that users might be interested in, and as a result, the questionnaire coverage will drop. On the other hand, if we merge the highly similar representative subjects with, e.g., "school setting" and "Education and Skills Committee," we will have a broader host of questions to suggest to users. Therefore, we will filter out dissimilar subjects from S_{Q_T} to S_{norm} and map extracted aspects from filtered S_{Q_T} to S_{norm} as it is shown in Fig. 3. As a result, if "education" is the subject of interest for a user, they have the opportunity to select which aspects of education they are more interested in, such as "Role" of education or "challenges" of education. Finally, we will show users the questions in which the selected aspects and normalized subjects have appeared.

4 Evaluation Methodology

For experiments, we use the QMSUM dataset (Zhong et al., 2021), which includes 232 product, academic, and committee meetings (Janin et al., 2003; Carletta et al., 2005). The dataset consists of 162, 35 and 35 meetings for training, validation and testing purposes respectively. Each meeting comes with a set of general and specific questions; the general ones are out of the scope of this work since they refer to very broad concepts, e.g., *"summarize the whole meeting."*. Further evaluations are conducted on the QMSUM test set.

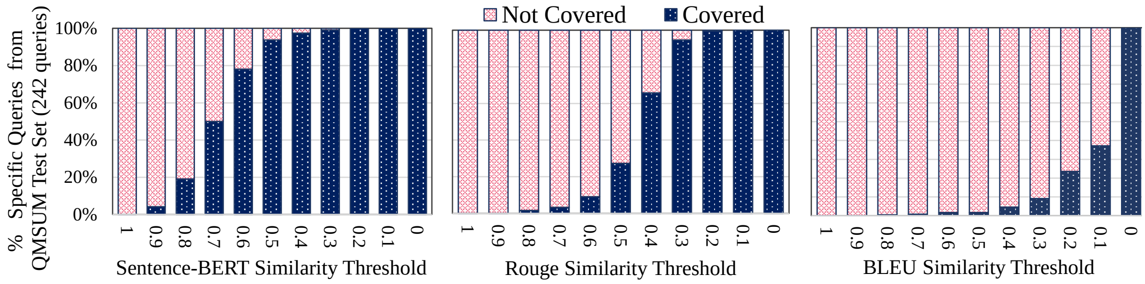


Figure 5: Coverage of specific queries in QMSUM test set among our generated questions considering different similarity metrics and threshold as coverage definition.

Table 2: Annotators agreement on annotated questions with respect to subjects and aspects using Kripendorff's score (Krippendorff, 2011)

	Subject	Aspect
Hard [Exact Match]	0.459	0.415
Soft [At least one term matched]	0.490	0.485

Table 3: CRF performance on extracting subjects and aspects of questions using 10-fold cross validation

	Precision	Recall	F1-Score
Subject	0.64	0.69	0.67
Aspect	0.89	0.80	0.84
N/A	0.63	0.73	0.68

4.1 Evaluating Framework Components

The proposed framework consists of several steps (Fig. 2). The used *meeting segmentation* (Solbiati et al., 2021) method has shown to outperform baselines (Hearst, 1997; Beeferman et al., 1999; Badjatiya et al., 2018). Hence, we refer to original paper for evaluation results.

Evaluating Question Generation We evaluate the quality of our generated questions by measuring the fraction of generated questions by human annotators in QMSUM that we covered in PREME. We assume the specific queries in the QMSUM dataset enjoy relatively high quality because annotators issued them after comprehensively reading the transcript (gold standard questions). Hence, Fig. 5 reports the similarity between most similar questions generated by PREME and the gold questions by three different similarity metrics i.e., Sentence-BERT similarity (Reimers and Gurevych, 2019), Rouge F-1 score (Lin, 2004), and BLEU score (Papineni et al., 2002). We assume a questions from QMSUM is covered if there is at least a question generated by PREME that has similarity is higher than a certain threshold $t \in [1, 0.9, ..., 0.1, 0]$. We report the percentage of 'Covered/Not Covered' questions based on different similarity matching thresholds. Based on Fig. 5 we conclude while we cover a relatively fair number of specific questions, there is still room for improvement. However, we should note that the questions in QMSUM are very limited, and initially, they were not supposed to cover all possible questions that one could raise from the meeting. Additionally, we observe that questions in QMSUM, which are issued by humans, include more abstractive questions while our generated questions inclined toward more factual ones.

Evaluating Subject and Aspect Extraction To assess the quality of the collected dataset, we measure Krippendorff's alpha agreement between annotators (Krippendorff, 2011) for extracted *subject* and *aspect* of the 1000 questions generated from the training set. Tab. 2 shows annotators have agreement ~ 0.4, which is interpreted as "Moderate" agreement for such a challenging task. Since different annotators might selected different section of the text, Tab. 2 reports both *hard* and *soft* agreements. we trained the CRF model using *crfsuite* library and evaluated it by 10-fold cross-validation. Given each term in the questions, the model predicts whether the term is considered the subject, aspect, or not applicable for labeling (N/A). Tab. 3 shows the result of the CRF model evaluation in terms of precision, recall, and F1 scores. We notice that the model shows better performance on detecting aspects compared to the subject.

4.2 Evaluating Questionnaires

To the best of our knowledge, we are first to propose a preference-based questionnaire as a way for meeting exploration; thus, no particular gold standard benchmark or evaluation metrics. We introduce a new evaluation strategy that satisfies the desired properties on coverage (**P1**) and the existence of answers in the transcript (**P2**). Since we require users to express their preference, it makes it challenging to simulate *'enough imaginative context'* among annotators. The proposed automatic metrics give good insights if our framework is ready to be tested through a user study in the future.

For our experimentation, we utilize the model SOTA called Locator in (Zhong et al., 2021) in

Table 4: Test set statistics and PREME Performance: Average number of generated questions and Coverage.

	#Meetings	Average # Turns	Average # Questions	Coverage (%)
Academic	9	893	1257	83.07%
Committee	6	214	1105	64.04%
Product	20	569	724	86.25%
All	35	591	927	81.62%

which, given the query, it can extract the relevant spans from the meeting. The Locator employs a hierarchical ranking-based model structure based on CNN (Kim, 2014) and Transformers (Vaswani et al., 2017) architecture. The Locator embeds each utterance of the meeting and feeds it to a CNN network by capturing the local features, and utilize Transformer layers to obtain contextualized turn-level representations. In addition, the speaker's embedding is also concatenated to the features list. Finally, the model uses MLP to score each turn, and the turns with the highest scores are considered the relevant spans for each question.

To measure the coverage (to satisfy **P1**), we adopt the newly proposed QA-style of evaluation (Deutsch et al., 2020; Wang et al., 2020) which has shown to have substantial correlation with human judgments in terms of questions quality assessments. *Coverage* is defined as the fraction of a meeting that a questionnaire encompasses. To measure the coverage, first, the relevant answer spans for the existing questions in a questionnaire are located. Further, the proportion of utterances that were already located as relevance answer spans w.r.t. the whole meeting transcripts, is measured as the coverage. We believe that that is a promising indicator of questionnaire informativeness. We run our experiments on the QMSUM test set. Tab. 4 shows the details of this test set. We over generate the questions and after removing the duplicates, on average, the questionnaire has 1257 unique questions from Academic meetings, 1105 questions from Committee meetings, and 724 questions from Product meetings. Further, Tab. 4 reports the percentage of utterances covered in each meeting. On average, our proposed questionnaire can cover 81% of the meeting. We also compared the coverage on different types of meetings. While our generated questionnaire covered Committee meetings the least (64%), the Product and Academic meetings show higher coverage (over 80%). Further, we evaluate how much the generated questions in PREME are answerable (to satisfy **P2**). Inspired by (Krishna and Iyyer, 2019), we run a pretrained QA model (Sanh et al., 2019) over generated questions and report the confidence score for each QA pair in Fig. 6. We use DistilBERT fine-tuned on SQUAD (Rajpurkar et al., 2016) dataset[6]. We observe that more than 73% of generated questions from PREME on meetings in test set of QMSUM shows confidence score higher than 0.5 and more than 42% of questions shows confidence score greater than 0.7. The results confirm that a promising portion of generated questions are answerable.

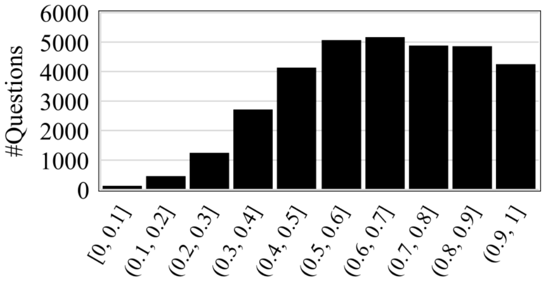


Figure 6: Histogram of Confidence Scores of Question-Answering (Sanh et al., 2019) model on generated questions from PREME.

5 Conclusions and Future Work

Due to the increasing amount of meeting transcripts, there is a need for automatic tools for interactive preference-driven exploration that allows to quickly examine a meeting . We have proposed an end-to-end framework, called PREME, that allows automatically building a questionnaire that will enable users to explore the most of discussed subjects and their aspects if desired. As a result, users are supplied with questions about the meetings that express their information needs, and answers can be found in the transcript. Since simulating actual users' preferences is challenging and requires hired annotators, we have proposed an automatic end-to-end evaluation strategy to demonstrate the desired properties (**P1** and **P2**) of the generated questionnaires. The future works should include an extensive survey that will reveal additional requirements for the PREME to satisfy, which will suggest additional evaluation metrics. Plus, proposing a new method for questionnaire generation will allow us to run a user study for pair-wise comparison of the methods and make correlation analysis to reveal the correlation between human and automatic evaluation metrics for the suggested task. We publicly release the collected dataset of annotated questions concerning its subjects and aspects, the code for questionnaires generation, and our evaluation procedure to carry forward the proposed state-of-the-art for the newly formulated problem.

[6] `https://huggingface.co/distilbert-base-cased-distilled-squad`

References

Oren Anava, Shahar Golan, Nadav Golbandi, Zohar Karnin, Ronny Lempel, Oleg Rokhlenko, and Oren Somekh. 2015. Budget-constrained item cold-start handling in collaborative filtering recommenders via optimal design. In *Proceedings of the 24th international conference on world wide web*, pages 45–54.

Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, Feras Al-Obeidat, and Ebrahim Bagheri. 2020. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing & Management*, 57(4):102248.

Negar Arabzadeh, Fattaneh Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. 2019. Geometric estimation of specificity within embedding spaces. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2109–2112.

Pinkesh Badjatiya, Litton J Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based neural text segmentation. In *European Conference on Information Retrieval*, pages 180–193. Springer.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. *arXiv preprint arXiv:1803.10357*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2020. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *CoRR*, abs/2010.00490.

Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. *CoRR*, abs/2004.11892.

Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Tai-Chia Huang, Chia-Hsuan Hsieh, and Hei-Chia Wang. 2018. Automatic meeting summarization and topic detection system. *Data Technologies and Applications*.

Tatsuya Ishigaki, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, and Manabu Okumura. 2020. Neural query-biased abstractive summarization using copying mechanism. In *European Conference on Information Retrieval*, pages 174–181. Springer.

François Jacquenet, Marc Bernard, and Christine Largeron. 2019. Meeting summarization, a challenge for deep learning. In *International Work-Conference on Artificial Neural Networks*, pages 644–655. Springer.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Bin Jiang, Jian Pei, Xuemin Lin, David W Cheung, and Jiawei Han. 2008. Mining preferences from superior and inferior examples. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 390–398.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kalpesh Krishna and Mohit Iyyer. 2019. Generating question-answer hierarchies. *arXiv preprint arXiv:1906.02622*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2021. Comsum and sibert: A dataset and neural model for query-based multi-document summarization. In *International Conference on Document Analysis and Recognition*, pages 84–98. Springer.

Oren Kurland and Lillian Lee. 2010. Pagerank without hyperlinks: Structural reranking using links induced by language models. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898.

Timothy J. Lawson, James H. Bodle, Melissa A. Houlette, and Richard R. Haubner. 2006. Guiding questions enhance student learning from educational videos. *Teaching of Psychology*, 33(1):31–33.

Timothy J Lawson, James H Bodle, and Tracy A McDonough. 2007. Techniques for increasing student learning from educational videos: Notes versus guiding questions. *Teaching of Psychology*, 34(2):90–93.

Adam D Lelkes, Vinh Q Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021*, pages 2501–2511.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. 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.

David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 105–114.

Marina Litvak and Natalia Vanetik. 2017. Query-based summarization using MDL principle. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 22–31, Valencia, Spain. Association for Computational Linguistics.

Cong Liu, Chi Yuan, Alex M Butler, Richard D Carvajal, Ziran Ryan Li, Casey N Ta, and Chunhua Weng. 2019. Dquest: dynamic questionnaire for search of clinical trials. *Journal of the American Medical Informatics Association*, 26(11):1333–1343.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Potsawee Manakul and Mark Gales. 2021. Long-span summarization via local attention and content selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6026–6041.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Karen Mazidi and Rodney D. Nielsen. 2014. Linguistic considerations in automatic question generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 321–326, Baltimore, Maryland. Association for Computational Linguistics.

Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146.

Jack Mostow and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *AIED*, pages 465–472.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. 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)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13666–13674.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Lior Rokach and Slava Kisilevich. 2012. Initial profile generation in recommender systems using pairwise comparison. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1854–1859.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. The first question generation shared task evaluation challenge.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Imran Sehikh, Dominique Fohr, and Irina Illina. 2017. Topic segmentation in asr transcripts using bidirectional rnns for change detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 512–518. IEEE.

Anna Sepliarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. 2018. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 172–180.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Daksha Singhal, Kavya Khatter, A Tejaswini, and R Jayashree. 2020. Abstractive summarization of meeting conversations. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–4. IEEE.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings. *arXiv preprint arXiv:2106.12978*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Hanna M Wallach. 2004. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

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

Leilan Zhang and Qiang Zhou. 2019. Topic segmentation for dialogue stream. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1036–1043. IEEE.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. 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*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.