Generating Multiple Questions from Presentation Transcripts: A Pilot Study on Earnings Conference Calls

Yining Juan,[1] **Chung-Chi Chen,**[2] **Hen-Hsen Huang,**[3] **Hsin-Hsi Chen** [1]
[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[2] AIST, Japan
[3] Institute of Information Science, Academia Sinica, Taiwan
ynjuan@nlg.csie.ntu.edu.tw, c.c.chen@acm.org,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

In various scenarios, such as conference oral presentations, company managers' talks, and politicians' speeches, individuals often contemplate the potential questions that may arise from their presentations. This common practice prompts the research question addressed in this study: to what extent can models generate multiple questions based on a given presentation transcript? To investigate this, we conduct pilot explorations using earnings conference call transcripts, which serve as regular meetings between professional investors and company managers. We experiment with different task settings and methods and evaluate the results from various perspectives. Our findings highlight that incorporating key points retrieval techniques enhances the accuracy and diversity of the generated questions.

1 Introduction

Preparing for an oral presentation is a common task in various domains, particularly in professional settings. For instance, researchers who have had their papers accepted at conferences need to deliver either an oral or poster presentation to share their findings with fellow researchers. Politicians must prepare for debates during election periods, while company managers are required to deliver speeches to update investors on company operations. When crafting their presentation drafts, a fundamental concern arises: what kinds of questions might the audience ask? This paper introduces a novel task, Multi-Question Generation (MQG), to assist presenters in preparing for Q&A sessions.

Diverging from previous studies that predominantly focused on one-to-one question generation tasks (Du et al., 2017; Song et al., 2018), the proposed MQG task is a one-to-many question generation task. In other words, after the presentation, audiences typically pose multiple questions. Table 1 showcases examples of one-to-one question generation and the proposed MQG task. The objective of the one-to-one question generation task is to generate a question based on a given condition (e.g., gross margin). In contrast, the proposed MQG task aims to generate multiple possible questions concurrently. We argue that exploring question generation tasks in a one-to-many setting closely aligns with real-world scenarios. However, the one-to-many setting also presents numerous challenges and research questions, including:

Presentation
Good day, and welcome to the Apple Q4 fiscal year 2022 earnings conference call…
One-to-One Question Generation
Condition: gross margin
Can you talk a bit about gross margin puts and takes?
Proposed MQG
1. Can you talk a bit about gross margin puts and takes?
2. How you think about balancing the consumer price versus your own costs and kind of the associated follow-through?
3. Any preliminary thoughts around capital intensity into fiscal 2023?

Table 1: Examples of one-to-one question generation and the proposed MQG.

1. **Retrieving Keypoints from Long Documents**: Language model limitations prevent the entire speech draft from being inputted into the models. This raises the research question of identifying which parts of the speech are important and likely to prompt questions. Consequently, keypoint retrieval becomes a crucial aspect for question generation. Can these keypoints improve the performance of the MQG task?

2. **Task Setting**: Differing from the one-to-one task setting, which involves generating one question given a passage and a condition, the proposed MQG task requires generating multiple questions. This leads to the following research questions: Can models generate all questions at once? Does generating questions

Proceedings of the 16th International Natural Language Generation Conference, pages 449–454
September 11–15, 2023.

sequentially yield better results?

3. **Evaluation**: As previously generated terms can influence the output of models, evaluating the accuracy and diversity of the generated questions becomes challenging. Specifically, can models generate several questions on the same topic, or can they generate questions from different perspectives?

To investigate these research questions, we collect earnings conference call transcriptions, regular meetings between company managers and professional analysts. Our aim with the proposed MQG task is to generate questions similar to those posed by analysts after listening to managers' presentations. We provide the collected dataset for future research endeavors.

To address these research questions, we propose the MQG-KR approach, combining MQG with Keypoint Retriever (KR). Keypoints are retrieved using BERT, enhancing question generation performance. We explore two task settings: generating all questions simultaneously and generating questions sequentially. Preliminary evaluations show that the MQG-KR approach improves the diversity of generated questions.

2 Related Work

Generating good questions is a challenging task for both humans and machines. Previous studies have primarily focused on one-to-one question generation, often centered around generating questions for reading comprehension tests. Heilman and Smith (Heilman and Smith, 2010) introduced syntactic transformations to convert given statement sentences into questions. Jia et al. (Jia et al., 2020) improved performance by incorporating a paraphrase module into their model. Wang et al. (Wang et al., 2020) generated questions based on the knowledge graph path of the input sentence. Song et al. (Song et al., 2018) matched given answers and paragraphs to augment context information for question generation.

In contrast to previous studies focusing on generating questions for machine reading comprehension tests, our work proposes generating questions in live presentations. There are distinct differences between these tasks. Questions for machine reading comprehension tests inquire about content, with most answers explicitly provided within the text. Complex questions may require some common sense inference. However, professional analysts posing questions during presentations will not ask for information already provided; instead, they seek clarification or further explanation (Palmieri et al., 2015). Consequently, models generating such questions must first identify unclear or insufficiently detailed portions of the presentation. Although earnings conference calls have been widely studied for various tasks such as stock movement prediction (Medya et al., 2022), volatility forecasting (Qin and Yang, 2019; Sawhney et al., 2020), and summarization (Mukherjee et al., 2022), the question generation task has received little attention. Thus, our paper aims to fill this research gap by providing an initial exploration of question generation from earnings conference calls.

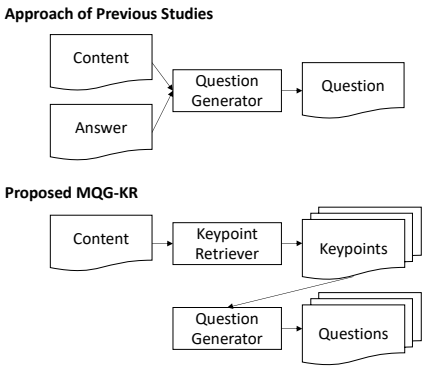

Figure 1: Overview of the proposed MQG-KR.

3 Method

Approach comparison is illustrated in Figure 1, highlighting two key distinctions between the proposed MQG-KR and previous approaches. Firstly, instead of providing the answer (condition) as input to the models, the proposed approach identifies keypoints that are likely to prompt questions from the audience and generates questions based on these keypoints. Secondly, while previous studies primarily focused on generating a single question, the proposed MQG-KR aims to generate multiple questions. In this section, we present two modules within the proposed MQG-KR framework: the Keypoint Retriever and the Question Generator.

3.1 Keypoint Retriever

As highlighted in Section 1, the length of the entire presentation exceeds the input capacity of most models. Additionally, not all sentences within the presentation hold investment-related significance, and professional analysts may not base their ques-

tions on these unimportant sentences. For instance, the greeting sentences in Table 1 do not provide valuable information to investors and may introduce noise to the models. To address this concern, we propose a keypoint retriever to refine the input.

Ideally, manual annotations for keypoints would be beneficial. However, the process of manual annotation is both expensive and time-consuming. Moreover, annotators without financial backgrounds may find it challenging to identify relevant keypoints. Therefore, we leverage the managers' answers to pinpoint the related section in their presentation. Specifically, we employ BM25 (Robertson et al., 2009) to calculate the similarity between the answer and each sentence in the presentation. Subsequently, we select the top-5 most similar sentences as the keypoints. This approach allows us to obtain fuzzy annotations for keypoints. Importantly, this process is applied solely to the training data, ensuring there are no issues related to data leakage. Once the keypoint labels are established, we train a classifier to identify the keypoints within the presentation, enabling the generation of questions based on these keypoints. For the keypoint retriever, we employ BERT (Devlin et al., 2019).

3.2 Question Generator

Our objective is to assess the effectiveness of the proposed approach, MQG-KR, in the novel task of Multi-Question Generation (MQG). We employ the well-performing generative model, FROST (Narayan et al., 2021), as our question generator. During the training process of FROST question generator, the entity chain of the presentation and questions is provided. In the inference (test) process, the models are required to generate both the entity chains and the questions. This approach has shown promise in abstractive summarization tasks (Narayan et al., 2021). In this paper, we present an initial exploration of FROST in the context of the one-to-many question generation task.

4 Experiment

4.1 Dataset

We compile a dataset of 995 transcriptions of earnings conference calls obtained from Seeking Alpha[1]. This dataset encompasses presentations from 18 different companies. During these 995 earnings conference calls, a total of 32,115 questions were asked. On average, each presentation received approximately 32 questions from the audience. In our task setting, models are required to generate multiple questions based on the provided presentation. We split the dataset into an 80% training set and a 20% test set for evaluation purposes, respectively.

4.2 Baselines

In addition to the vanilla FROST model, we consider two other baselines: Longformer (Beltagy et al., 2020) and LongT5 (Guo et al., 2022). These models are specifically designed to handle longer documents. Longformer employs sparsity in the attention matrix and utilizes a global and sliding window approach for encoding longer sequences. On the other hand, LongT5 is an extension of the T5 model (Raffel et al., 2020) and adopts a similar approach to Longformer. Notably, LongT5 has demonstrated superior performance compared to Longformer across six summarization datasets (Guo et al., 2022).

4.3 Evaluation

We employ the ROUGE-L score for evaluating the generated results (Lin, 2004). Additionally, as the proposed MQG task involves generating multiple questions, we propose two additional evaluation metrics, namely ROUGE-AMG and ROUGE-AMR, to assess the results from different perspectives. Each generated question (GQ_i) is assigned a list of ROUGE-L scores (GL_i) with each reference question (ground truth).

ROUGE-AMG is calculated using the following equation:

$$ROUGE - AMG = \frac{\sum_{i=1}^{N} \max(GL_i)}{N}, \quad (1)$$

where N is the number of generated questions. ROUGE-AMG measures the extent to which the generated question is similar to the reference question.

On the other hand, we also evaluate the results from the perspective of reference questions using the ROUGE-AMR metric. Each reference question (RQ_j) receives a list of ROUGE-L scores (RL_j) with each generated question. ROUGE-AMR is calculated as follows:

$$ROUGE - AMR = \frac{\sum_{j=1}^{M} \max(RL_j)}{M}, \quad (2)$$

where M is the number of reference questions.

As mentioned in Section 1, models may generate questions on the same topic by merely rephrasing

[1] https://seekingalpha.com/earnings/earnings-call-transcripts

	Question Generator	Max Input Length	ROUGE-L (↑)	ROUGE-AMG (↑)	ROUGE-AMR (↑)	Diversity (↓)
Baseline	Longformer	4,096	19.37	18.21	15.54	100.00%
	LongT5	4,096	20.48	19.23	15.37	100.00%
	FROST	1,024	23.08	22.20	17.95	100.00%
MQG-KR	LongFormer	4,096	24.26	21.82	18.29	96.48%
	LongT5	4,096	24.43	22.65	18.66	96.48%
	FROST	1,024	**26.93**	**25.79**	**21.33**	**95.47%**

Table 2: Experimental results. ↑ and ↓ denote the higher the better and the lower the better, respectively.

	ROUGE-L
DialogueVED	22.08
PLATO	22.13
MQG-KR (FROST)	**26.93**

Table 3: Results of sequential generation.

the question. However, since our goal is to generate diverse questions that could potentially be asked, we further evaluate the diversity by examining whether the most similar reference question for each generated question is the same or not. We calculate the ratio based on the maximum number of questions similar to the same reference question. Therefore, from a diversity perspective, a lower ratio is considered a better evaluation metric.

4.4 Experimental Results

The experimental results are presented in Table 2. Firstly, we observe that the proposed MQG-KR consistently yields improved performance across different question generators. These findings emphasize the significance of the keypoint retriever in the context of the proposed MQG task. Secondly, despite FROST having a shorter maximum input length compared to the other two models, it achieves the best performance among the baselines. This outcome may be attributed to the design of the entity chain prediction task within the decoder component, indicating the importance of entity chains in the proposed MQG task. Lastly, with respect to diversity, we find that all questions generated by the vanilla question generators pertain to the same topic. Conversely, the proposed MQG-KR models exhibit the ability to generate a wider range of diverse questions. Notably, MQG-KR combined with FROST demonstrates the highest diversity performance based on the evaluation conducted.

4.5 Sequential Generation

In earnings conference calls, analysts typically ask questions one by one, with later questions often following up on the previous ones. To simulate this process, we also experiment with a sequential question generation setting. After generating a question, it is then used as input to the question generator to generate the subsequent question. For the sequential setting, we employ two well-performing dialogue generation models: PLATO (Bao et al., 2020) and DialogueVED (Chen et al., 2022). PLATO utilizes discrete latent variables to address one-to-many problems, while DialogueVED incorporates a multi-task pre-finetune process to enhance generated results and employs continuous latent variables for one-to-many generation tasks.

The experimental results are presented in Table 3. First, PLATO outperforms DialogueVED in the MQG task. Second, the performance of the models in the sequential generation setting falls short of FROST and the proposed MQG-KR, which operate under the simultaneous generation setting. However, since the models are designed for different purposes, we refrain from determining the best-performing model. Our aim in this paper is to explore the proposed MQG task from various perspectives, sharing our findings and establishing baselines for future studies.

5 Conclusion

This paper introduces the novel task of MQG and explores its potential applications in generating multiple questions based on a given presentation transcript. Our experiments demonstrate that incorporating a keypoint retriever improves the accuracy and diversity of the generated questions. This research contributes to the fields of NLP and Information Retrieval (IR) by offering insights into the MQG task and its relevance in various scenarios.

Our findings provide a valuable starting point for future research in MQG. By better anticipating and preparing for question and answer sessions, presenters can benefit from the generated questions. We believe that this work stimulates further discussions, advancements, and collaborations in the exciting field of Multi-Question Generation, driving the development of more effective and efficient question generation models.

Acknowledgments

This research is supported by National Science and Technology Council, Taiwan, under grants 110-2221-E-002-128-MY3, 110-2634-F-002-050-, and 111-2634-F-002-023-. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Limitations

One limitation of this work is the focus on a specific application scenario, namely earnings conference calls, while only mentioning other real-world presentation and QA scenarios. Acquiring transcriptions for conference oral presentations or numerous meetings can be challenging and resource-intensive. Although video recordings may be available, the transcription process is time-consuming and costly. Moreover, data from politicians' debates are relatively scarce compared to earnings conference call data, as elections are not held every year. In contrast, quarterly earnings conference calls yield four transcriptions per company annually. Future studies can either develop new methods utilizing the proposed dataset or explore MQG tasks using alternative resources. This paper aims to serve as a starting point for the IR and NLP communities to delve into MQG tasks by improving keypoint retrieval performance and enhancing the ability to generate insightful questions.

Furthermore, we propose a more granular future direction for MQG. Subsequent work can involve annotating questions based on the question taxonomy proposed by Palmieri et al. (Palmieri et al., 2015). This taxonomy classifies analysts' questions into three levels, each comprising two to six labels. We believe that incorporating these labels can aid in automatically understanding analysts' questions and generating high-quality questions. Additionally, future research can explore assisting presenters in preparing answers for the generated questions, thereby progressing towards the development of a Q&A session tutor or assistant. This work highlights the significance of not only NLP techniques but also the relevance of IR considerations in this application domain.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

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

Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. 2022. DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

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, Los Angeles, California. Association for Computational Linguistics.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2020. How to ask good questions? try to leverage paraphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6130–6140, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sourav Medya, Mohammad Rasoolinejad, Yang Yang, and Brian Uzzi. 2022. An exploratory study of stock price movements from earnings calls. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 20–31, New York, NY, USA. Association for Computing Machinery.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Rudi Palmieri, Andrea Rocci, and Nadzeya Kudrautsava. 2015. Argumentation in earnings conference calls. corporate standpoints and analysts' challenges. *Studies in communication sciences*, 15(1):120–132.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020. VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013, Online. Association for Computational Linguistics.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 569–574, New Orleans, Louisiana. Association for Computational Linguistics.

Siyuan Wang, Zhongyu Wei, Zhihao Fan, Zengfeng Huang, Weijian Sun, Qi Zhang, and Xuanjing Huang. 2020. PathQG: Neural question generation from facts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9066–9075, Online. Association for Computational Linguistics.