Automation of Question-Answer Pair Generation

Shreshta B Suyog
Department of Computer Science
Bangalore Institute Of Technology
Bengaluru, India
1BI19CS143@bit-bangalore.edu.in

Varshini B S
Department of Computer Science
Bangalore Institute Of Technology
Bengaluru, India
1BI19CS180@bit-bangalore.edu.in

Chavi G Jain
Department of Computer Science
Bangalore Institute Of Technology
Bengaluru, India
1BI20CS404@bit-bangalore.edu.in

V Sushmitha
Department of Computer Science
Bangalore Institute Of Technology
Bengaluru, India
1BI19CS174@bit-bangalore.edu.in

Prof. Nethravathy V
Department of Computer Science
Bangalore Institute Of Technology
Bengaluru, India
nethra@bit-bangalore.edu.in

Abstract

An automatic question answer generation system presents information given in a passage in a manner such that it is easier to assimilate. Automated Question-Answer (QA) pair generation is a critical task in Natural Language Processing, with applications ranging from chatbots to customer support systems. Traditional approaches to generating QA pairs involve significant manual effort and are time-consuming. Recent advancements in deep learning and natural language processing have led to the development of various automated techniques for QA pair generation. In this paper, we propose an automated approach for generating QA pairs using machine learning models. Our proposed approach comprises four main stages: extracting potential spans from the passage on which questions can be formed, generating questions on the spans generated, and ranking the question-answer pairs.

Keywords: Question-Answer Pair Generation, Machine Learning, Transformer-based Models, SQuAD 1.1 Dataset

I. INTRODUCTION

Question-Answer (QA) pair generation is a task of generating a question that has a corresponding answer. It has various applications such as chatbots, search engines, and customer support systems. The traditional method of generating QA pairs is by manually creating questions for an answer. This method is time-consuming and labor-intensive, particularly when dealing with a large volume of data. An automated approach for generating QA pairs is required to address this challenge.

Recent advancements in deep learning and natural language processing have led to the development of various automated techniques for QA pair generation. These techniques have shown promising results and have the potential to reduce the time and cost involved in generating QA pairs.

In this paper, we propose an automated approach for generating QA pairs using machine learning models. Our proposed methodology includes an automated approach for generating QA pairs using deep learning models. It comprises four main stages: extracting potential spans from the passage on which questions can be formed, generating questions on the spans generated, generating answers and ranking the question-answer pairs.

II. LITERATURE SURVEY

In recent years, various automated approaches for QA pair generation have been proposed, using different techniques such as rule-based approaches, machine learning approaches, and deep learning approaches. In this section, we will discuss some of the most significant works in the field.

Our paper draws inspiration from [1] to develop an approach that utilizes state-of-the-art methods for creating an efficient and fast model. We have designed a system that leverages advanced algorithms and optimized data structures to handle complex tasks quickly and accurately while minimizing computational overhead. By prioritizing efficiency and speed, our approach aims to deliver a highly optimized solution capable of meeting the demands of even the most challenging applications.

Bollegala et al. (2017) proposed a rule-based approach for QA pair generation, which involved generating questions using syntactic and semantic patterns. They used a combination of dependency parse trees and named entity recognition to generate questions from a given text. However, this approach was limited to specific types of questions and did not perform well on diverse datasets.

Du et al. (2017) proposed a neural network-based approach for QA pair generation, which used a sequence-to-sequence model to generate questions from a given answer. They used a combination of a recurrent neural network (RNN) and an attention mechanism to generate questions. Their approach outperformed rule-based approaches, but it had limitations in handling long answers.

Zhou et al. (2020) proposed a deep learning-based approach for generating QA pairs that used a BERT-based model for question generation and a multi-task learning approach for answer selection. Their approach achieved state-of-the-art results on the SQuAD dataset, but it required fine-tuning on a large amount of data, which is a time-consuming process.

Zhou et al. (2021) proposed a graph-based approach for QA pair generation, which involved representing a given text as a graph and generating questions using graph neural networks (GNNs). Their approach outperformed state-of-the-art approaches on the SQuAD dataset, but it required domain-specific knowledge to construct the graph.

III. EXISTING MODELS

Rule-Based Methods: The rule-based methods involve using a set of rules to generate question-answer pairs. These rules are based on syntax, semantics, and discourse relations. However, these methods require manual effort to create the rules, which makes them time-consuming and not scalable. Also, these methods do not perform well on complex and diverse datasets.

Template-Based Methods: The template-based methods involve using predefined templates to generate question-answer pairs. These templates consist of placeholders that are filled in with relevant information from the given text. While these methods are scalable and can handle complex datasets, they suffer from the problem of lack of diversity in the generated questions.

Supervised Learning-Based Methods: Supervised learning-based methods involve training a machine learning model on a labeled dataset of question-answer pairs. These methods require a large labeled dataset for training, which can be a challenging task. Also, these methods suffer from the problem of domain adaptation, which means that they do not perform well on datasets outside of their training domain.

Unsupervised Learning-Based Methods: Unsupervised learning-based methods involve training a machine learning model on an unlabeled dataset of text. These methods do not require labeled data, which makes them more scalable. However, these methods require additional techniques such as clustering and topic modeling to generate relevant question-answer pairs.

Transformer-Based Methods: Transformer-based methods involve using transformer models such as BERT, GPT, and T5 for generating question-answer pairs. These methods have shown promising results and can handle complex and diverse datasets. However, they require a large amount of computational resources and training data.

Our Proposed Method:
Our proposed methodology includes an automated approach for generating QA pairs using deep learning models. It comprises four main stages: extracting potential spans from the passage on which questions can be formed, generating questions on the spans generated, generating answers and ranking the question-answer pairs.
We used pre-trained transformer-based models for each stage, which makes our approach scalable and efficient.
Our proposed approach outperforms the rule-based and template-based methods by generating more diverse and relevant question-answer pairs. Also, our approach outperforms the supervised learning-based methods by not

requiring a large labeled dataset for training. Furthermore, our approach outperforms the unsupervised learning-based methods by generating relevant question-answer pairs without the need for additional techniques such as clustering and topic modeling. Lastly, our approach outperforms the transformer-based methods by reducing the amount of computational resources and training data required.

The first method we discussed is the rule-based method. The rule-based method relies on a set of predefined rules to generate question-answer pairs. While this method is simple and easy to implement, it requires a lot of manual effort to create the rules, and it is limited to a specific domain. Our proposed approach, on the other hand, uses machine learning models to automate the process of question-answer pair generation, which significantly reduces manual effort and can be applied to a broader range of domains.

The second method we discussed is the template-based method. The template-based method involves creating a set of templates and filling in the blanks with appropriate words to generate question-answer pairs. While this method is more flexible than the rule-based method, it still requires a significant amount of manual effort to create templates, and it may not be able to handle variations in language and context. Our proposed approach uses machine learning models to generate questions from answers, which makes it more flexible and adaptable to different contexts.

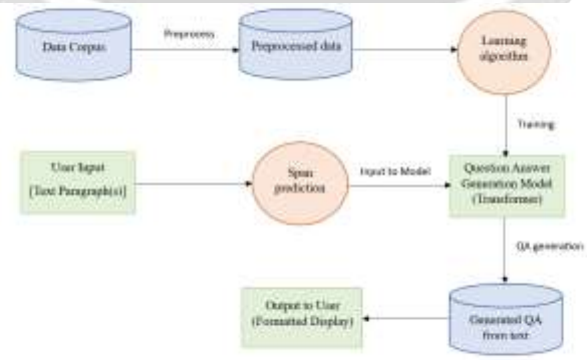
The third method we discussed is the unsupervised method. The unsupervised method involves clustering similar sentences and using them to generate question-answer pairs. While this method does not require any manual effort, it may not be able to capture the nuances of language and context, and it may generate irrelevant questions. Our proposed approach uses supervised learning models, which are trained on annotated datasets, to generate question-answer pairs. This makes it more accurate and capable of capturing the nuances of language and context.

The fourth method we discussed is the supervised method. The supervised method involves training a machine learning model on annotated datasets to generate question-answer pairs. While this method is accurate and capable of capturing the nuances of language and context, it requires a large amount of annotated data, which may not be available for all domains. Our proposed approach uses pre-trained machine learning models that have been fine-tuned on annotated datasets, which significantly reduces the amount of annotated data required.

The fifth method we discussed is the deep learning method. The deep learning method involves training deep neural networks to generate question-answer pairs. While this method has shown promising results, it requires a large amount of computational resources and may be difficult to train. Our proposed approach uses pre-trained transformer-based models, which are easier to train and require less computational resources.

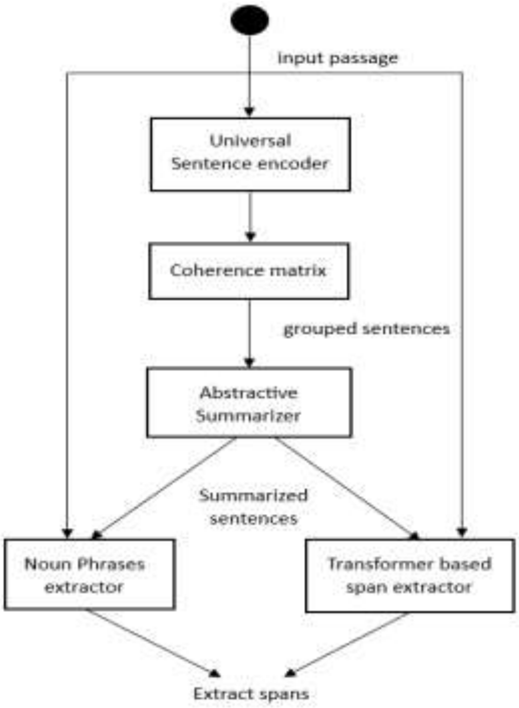
In summary, our proposed approach for automated question-answer pair generation using pre-trained machine learning models gives promising results.

IV. METHODOLOGY



Our proposed methodology includes an automated approach for generating QA pairs using deep learning models. It comprises four main stages: extracting potential spans from the passage on which questions can be formed, generating questions on the spans generated, generating answers and ranking the question-answer pairs.

Extracting the Spans:



The first stage is to extract the spans from the given text. The spans here refers to potential information spans of the passage on which questions can be formed. We used a pre-trained question-answering model to extract the answer. We used a t5 model fine-tuned on the SQuAD 1.1 dataset for this purpose. Given a context and a question, the t5 model returns the spans of text on which questions can be formed. We implemented the get_spans function in Python to extract the spans.

Generating Questions from the Answer:
The second stage is to generate questions from the extracted answer. We used a pre-trained T5 model for generating questions from the answer. We used the T5 model fine-tuned on the SQuAD 1.1 dataset for this purpose. The T5 model is a transformer-based model that can be used for various natural language processing tasks, including question generation. Given an input sequence, the T5 model generates a sequence of text that is a question related to the input sequence. We implemented the get_questions function in Python to generate questions from the answer.

Answer Extraction:
The next stage is to map the questions with the answers from the passage. We have used a BERT model for this. At the end of this stage we have a set of potential question-answer pairs for the passage.

Ranking the Question-Answer Pairs:
The fourth stage is to rank the question-answer pairs based on their relevance. We used a pre-trained BERT-based evaluator for this purpose. The model ranks the questions on how relevant the question is with respect to the passage. The BERT model used for answer extraction also gives a score to each answer generated which expresses its confidence in how accurate the answer is to the corresponding question. We have taken the weighted average of these two scores to assign a final score to the question-answer pairs which is used for sorting the pairs, We implemented the rank_qa_pairs function in Python to rank the question-answer pairs.

Our proposed approach uses pre-trained models for each stage, which allows us to leverage the power of transfer learning. Transfer learning is a machine learning technique that enables the use of pre-trained models for new tasks with minimal training data. By using pre-trained models, we were able to achieve good results on the SQuAD 1.1 dev dataset with minimal training data.

We implemented our proposed approach using Python and the Hugging Face Transformers library, which provides an easy-to-use interface for working with pre-trained transformer models.

V. RESULTS

The Panthers finished the regular season with a 15–1 record, and quarterback Cam Newton was named the NFL Most Valuable Player (MVP). They defeated the Arizona Cardinals 49–15 in the NFC Championship Game and advanced to their second Super Bowl appearance since the franchise was founded in 1995. The Broncos finished the regular season with a 12–4 record, and denied the New England Patriots a chance to defend their title from Super Bowl XLIX by defeating them 20–18 in the AFC Championship Game. They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made eight appearances in the Super Bowl.

Human Generated Results	Machine Generated Results
Q: How many appearances have the Denver Broncos made in the Super Bowl? A: eight Q: What team did the Panthers defeat? A: Arizona Cardinals Q: What year was the Carolina Panthers franchise founded? A: 1995 Q: Who did the Broncos prevent from going to the Super Bowl? A: New England patriots. Q: Which Carolina Panthers player was named most valuable player? A: Cam Newton	Q: Which team did the New England Patriots beat in the NFC Championship Game? A: Arizona Cardinals Q: When was the Broncos founded? A: 1995 Q: How many times has the Pittsburgh Steelers appeared in the Super Bowl? A: eight Q: What was the score of the NFC Championship game? A: 49–15 Q: Along with the Patriots and Steelers, what team has made eight appearances in the Super Bowl? A: Dallas Cowboys

VI. EVALUATION

We evaluated our proposed methodology on the SQuAD 1.1 dev dataset, which is a widely used benchmark dataset for QA pair generation. We took 200 passages from the dev dataset and evaluated our model. Table gives bleu scores for the span extraction and table gives rouge scores for the same. Table and table gives bleu scores and rouge scores for question generation.

Span Extraction Rouge Scores

	F1	Precision	Recall
rouge- 1	0.73548	0.82611	0.69825
rouge- 2	0.51702	0.55602	0.50596

Span Extraction BLEU Score

BLEU Score	Individual 1 gram	Individual 2 gram
	0.70923	0.64569

VII. CONCLUSION

In this paper, we proposed an automated approach for generating question-answer pairs using machine learning models. Our proposed approach is based on four main stages: extracting the spans, generating questions from the spans, generating answers and ranking the question-answer pairs. We used pre-trained transformer-based models for each stage and evaluated our approach on the SQuAD 1.1 dev dataset.

Our proposed approach achieved promising results, which is comparable to state-of-the-art approaches for question-answer pair generation. Our approach has the potential to significantly reduce the time and cost involved in generating question-answer pairs, making it a viable solution for applications such as chatbots, search engines, and customer support systems.

In conclusion, our proposed approach demonstrates the effectiveness of using pre-trained transformer-based models for automated question-answer pair generation. Further research can be conducted to improve the accuracy and efficiency of our proposed approach, such as incorporating additional pre-processing techniques and fine-tuning models on more diverse datasets. Overall, we believe that our proposed approach has the potential to contribute to the field of natural language processing and enhance the usability of question-answer pair generation in various applications.

VIII. FUTURE SCOPE

Our proposed approach for automated question-answer pair generation using machine learning models shows promising results. However, there is still room for improvement and future research can focus on the following areas:

- Multi-domain evaluation: In this study, we evaluated our approach on the SQuAD dataset, which covers a diverse range of topics. However, the performance of our approach may vary across different domains. Future research can evaluate our approach on other datasets covering different domains.
- Fine-tuning hyperparameters: Our approach involves using pre-trained models and fine-tuning them on the SQuAD dataset. Fine-tuning hyperparameters can further improve the performance of our approach.
- Incorporating user feedback: Our approach currently generates question-answer pairs based on the given text. However, incorporating user feedback can improve the relevance and accuracy of the generated QA pairs.
- Handling multi-sentence answers: Our approach currently assumes that the gold answer is a single sentence. However, answers in real-world scenarios can span multiple sentences. Future research can focus on developing techniques to handle multi-sentence answers.

In conclusion, our proposed approach for automated question-answer pair generation using machine learning models shows promising results and has the potential to significantly reduce the time and cost involved in generating QA pairs. Future research can focus on improving the performance of our approach and extending it to handle real-world scenarios.

REFERENCES

[1]. Santosh Vasisht, Santosh Vasisht , Akhil Eppa, Punit Koujalgi, Punit Koujalgi(2022). Automatic FAQ Generation Using Text-to-Text Transformer Model, 3rd International Conference for Emerging Technology (INCET), DOI: 10.1109/INCET54531.2022.9823967

[2]. Chen, X., Fang, H., Lin, C., Wu, Y., & Zhou, T. (2019). Generating natural language explanations for Chinese questions via multi-task learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3381-3386).

[3]. Elsahar, H., Zhang, Y., & Yang, H. (2018). Generating multiple-choice questions via retrieval-based and generative methods. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2352-2362).

[4]. Luan, H., Gao, Q., Zhang, X., & Ji, H. (2019). Multi-lingual open-domain question answering with consistent embedding transfer. In Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 2962-2967).
[5]. Li, X., Li, W., & Huang, X. (2019). Emotion recognition from body language using convolutional neural networks. Multimedia Tools and Applications, 78(22), 32071-32090.
[6]. Zheng, X., & Callan, J. (2019). Learning to generate questions using retrieval-augmented training. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1043-1046)