IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Automatic Question Generation System

[1] Akash Gavali, [2] Kartik Bhagwat, [3] Dr. Hansaraj Wankhede, [4] Prof. Achamma Thomas

[1,2] Research Student, [2,4] Project Guide

[1,2,3,4] Department of Artificial Intelligence

[1,2,3,4] G. H. Raisoni College of Engineering, Nagpur

Abstract: This article presents a new rule for automatic query construction that addresses the precise identification of syntactic and semantic patterns in sentences. The design and use of the method is carefully explained. Although the main goal is to create text-based questions, machine learning has shown great success in reading comprehension, especially focusing on creating questions from sentences. The system developed in human evaluation shows that it is more useful than other systems, that it creates problems like people and creates problems in general.

Index Terms - Natural Language Processing, Machine Learning, Artificial Intelligence, Question Generation, T5 Model

INTRODUCTION

In recent years, natural language processing (NLP) has become very successful and has changed the entire paradigm of language applications. An important development in this field is the automatic generation of questions from text. This research project aims to contribute to the growing knowledge of NLP by introducing a new rule-based approach to automated querying. Our goal is not only the extraction of syntactic patterns, but also the analysis of the syntactic and semantic content in sentences.

This research stems from the growing need for intelligent machines that can understand and process human language. Question design is important in many applications, including learning, data retrieval, and agent interaction. By working on this process, we aim to improve the performance of natural language communication and facilitate interaction between humans and machines.

The main goal of our project is to create a system that can not only capture the complex phrasing of the sentence but also explore the nuances in its semantics. We consider a system that recognizes the relationship between elements and generates relevant questions that go beyond the superficial syntactic process. This research is at the intersection of linguistics and artificial intelligence to bridge the gap between human language understanding and computational models.

Beyond its original goals, our project took an unexpected turn when we investigated its effectiveness in improving reading comprehension. Designed specifically to build questions from description, this document provides a strong foundation for testing changes to our policy and its effectiveness. After the automated analysis showed good performance, it led to further research into the body's ability to decipher the overall meaning of the content.

Human testing forms an important part of our research and provides insight into the body's natural and human ability to create problems. Preliminary results show that our proposed method outperforms existing models in terms of query quality and demonstrates the ability to transform land-generated queries.

This research paper describes the design, implementation, and evaluation of our automated survey design. Through this research, we aim to contribute to the understanding of NLP and pave the way for better language understanding and interaction in artificial intelligence applications.

OBJECTIVE

This research paper describes the design, implementation, and evaluation of our automated survey design. Through this research, we aim to contribute to the understanding of NLP and pave the way for better language understanding and interaction in artificial intelligence applications.

The project is motivated by the growing demand for intelligent systems capable of understanding and processing human language effectively. Automated question generation has practical applications in various fields, including educational technology, information retrieval, and conversational agents. By automating this process, the project seeks to enhance natural language interfaces and facilitate more seamless interactions between humans and machines.

LITERATURE REVIEW

Onur et al. It offers a rule-based automatic question generation method that determines the syntactic and semantic structure of a sentence. The main goal is to generate general questions based on the semantic role of a word. The scheme adopts dependency based, Named Entity Recognition (NER) based and Semantic Role (SRL) tag based policy/layout. Chunking is used to distinguish between "who" and "what" questions [1]

Aleena et al. It focuses on the use of query-generating systems that can understand natural language for data processing and management. The system deals with data pre-processing, keyword extraction and natural language processing. It allows the development of fast, stable and stochastic systems suitable for learning. The system accepts the input as text, file or PDF file, removes punctuation and uses the appropriate language. The content is extracted using the TF-IDF algorithm and the content is checked on the wiki. The WordNet tool is used to create a three-dimensional query and clarify the expression [2].

Priti et al. Discuss the issues raised by Bloom's taxonomy, including the ins and outs of Stanford POS tags. Perform syntactic and semantic analysis, including some audio tagging and segmentation in syntactic analysis and namespaces in semantic analysis. To create a question, first the word "WH" will be drawn in the text [3].

D. R. CH and S. K. Saha presented a research paper on automatic marking of various questions in the text. Their methods include reading text from a repository, using NLP-based concepts to write text and count the frequency of words, and pattern matching for keyword selection. Word networks, pattern matching, domain ontologies, and semantic analysis contribute to the creation of intervention objects, which are mainly aimed at creating a functional framework for the production of UHT [4].

Ankita, K.A. et al. Discuss the complexity of some speech tagging and suggest reducing the complexity with hidden Markov models. The article discusses sentence-level tagging using HMM-based taggers. The so-called recognition site (NER) has also been proposed to simplify the POS tagging process [5].

The main focus in the paper by Amruta Umardand et al is the concept of question paper generation system. This article explores different ways to create problems, whether through random or automated processes. This automatic system is considered very good to use. The system allows issues to be stored and reused by collecting information and using security measures. This ensures the integrity and security of the problem while allowing other authors to create problems based on similar or different concepts [6]

In Kalpana et al.'s study, the aim was "What is a POS tag?" and clarify the functions of the part of speech. Part of speech tagging involves assigning a part of speech to a word. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunctions, and their subcategories. Taggers rely on many types of information, including single words, messages, and rules. Dictionaries group or classify certain words, and it is common for many words to belong to the same group. For example, "escape" can be used as both an adjective and an adverb. To resolve the confusion, the tagger uses the information that will appear. This way some of the speech tagging is done and appropriate characters are given [7]

In their work, Edward Loper and colleagues have introduced an innovative approach to streamline and adapt the practical aspect of computational linguistics. The NLTK toolkit serves as a flexible and comprehensive framework for natural language processing, covering both symbolic and statistical approaches. The toolkit is implemented through a set of modules, with each module defining an alternative data structure or task. Core modules within the toolkit characterize various systems utilized throughout. Examples of modules include

Chunk Parsing and Probabilistic Parsing, among others. The toolkit's modules offer valuable instances of well-structured code, providing insight into clean code organization and thorough documentation [8].
Ankita, K. A., and colleagues address the complexity of part-of-speech (POS) tagging regarding the level of computation required to determine POS tagging in sentences. The goal of their paper is to reduce the complexity associated with the number of comparisons made by the Hidden Markov Model (HMM) of POS tags. The authors also propose another method for name recognition (NER) using the Bloom filter. Although many POS taggers are available, researchers continue to search for methods that require less time and effort. In the HMM-based tagger discussed in this article, tags are assigned to entire sentences rather than to individual words. These symbols exploit the possibilities of transformation and emission. The proposed algorithm places two words in a block and treats them as a unit by collecting the text of the unit together [9]
In their study, Mohd Husain and colleagues discuss different ways to ask categorization questions, including target type, fill-in-the-blank, and “what” type questions. The sign of this problem involves extracting relevant information from the text through integration and transformation. Use this method to create explanations and facts. Although the main focus of our article is on "what" questions [10]
In their work, Min-Kyoung Kim and co-authors proposed an automatic query generator based on the generated query response. They created a legal framework that includes a center, called an authentication center, to generate questions and determine common answers. [11]
Sushmita Gangopadhyay and colleagues introduced the question-answer design process. Their approach combines an automated algorithm that processes SQuAD data with a combination of neural entity selection algorithms and recurrent neural networks. [12]
Mai Mokhtar developed an automatic query model based on deep learning. The approach involves an encoder-decoder model combining name recognition and Seq-2-Seq techniques, and the model is trained on the SQuAD dataset. [13]
In their work, Riken Shah and collaborators developed an automatic query system designed for intelligent instruction. The model was trained on information provided by Wikipedia and combined with social discovery techniques to create an intervention. However, this system is limited due to its reliance on prior knowledge base and limited subject matter. [14]

PROPOSED SYSTEM

The concept of automatic question and answer generation (QAG) systems utilizes a variety of techniques to generate precise and relevant question and answer pairs from a wide range of digital content for a variety of questions. This section provides an overview of how the system works. The QA generator uses the Hugging Face Transformers library to develop the T5 (text-to-text) model to generate answers to questions. The T5 model is based on the Transformer architecture and is pre-trained on large datasets, including functions such as translation, annotation, and query answering. When the T5 model is loaded, the QA generator receives the text for processing; This can be any type of digital content such as text, books, research articles or video tutorials. The QA generator then processes the input data in preparation for creating questions and answers.

The first phase consists of several tasks such as tokenization, sentence segmentation, and removal of stopped words.

1. Tokenization: This involves splitting the input into separate words or tokens, which are then fed into the T5 model for processing.
2. Sentence segmentation: This process requires segmenting the input text into separate sentences, which are the input to the generated query.
3. Stop word removal: In this step, English stop words such as "the", "is" and "and" will be filtered out. Removing pauses is important because they add little meaning to the text and add noise to the question formation process.

The server receives requests containing text available from a variety of sources, such as books, research articles, or educational videos. Then use the T5 model to create question-answer pairs. T5 is a Transformer-based language that learns from big data and is used in many languages, including text.

The system uses a pre-learned T5 model (called “t5-basic question generator”) to generate questions and answers. This model represents a fine-tuned version of T5 specifically designed to improve question generation ability by training question and answer data.

.

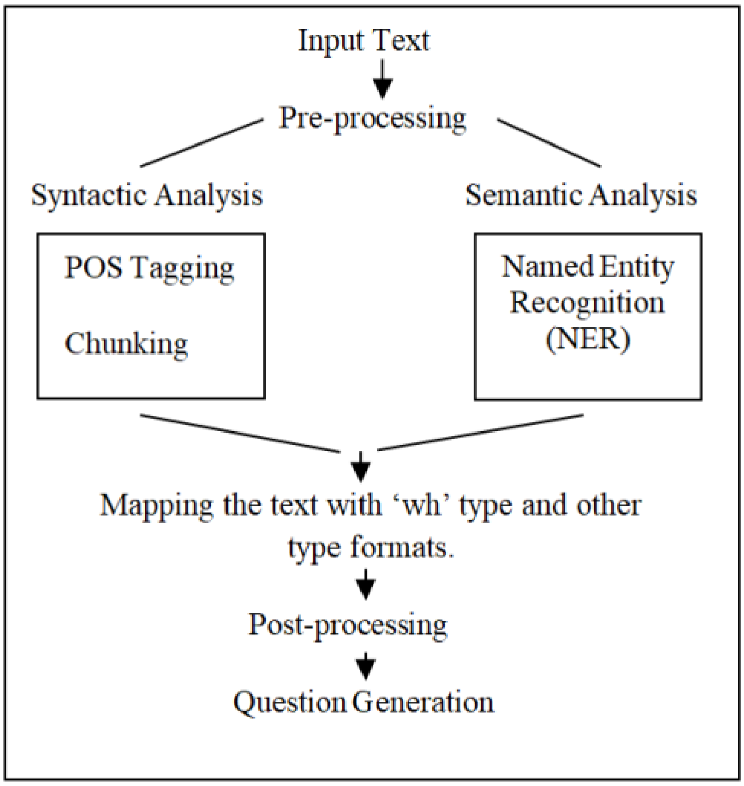

Fig. 1: Flowchart of proposed system

T5-Base-Question-Generator is a fine-tuned T5 model that focuses solely on generating questions from text. The model is trained from a large database of text and question-answer pairs using a link-ranking model to generate questions. The model simply uses the decision architecture and a probability model-based sampling technique to iterate the query one token at a time.

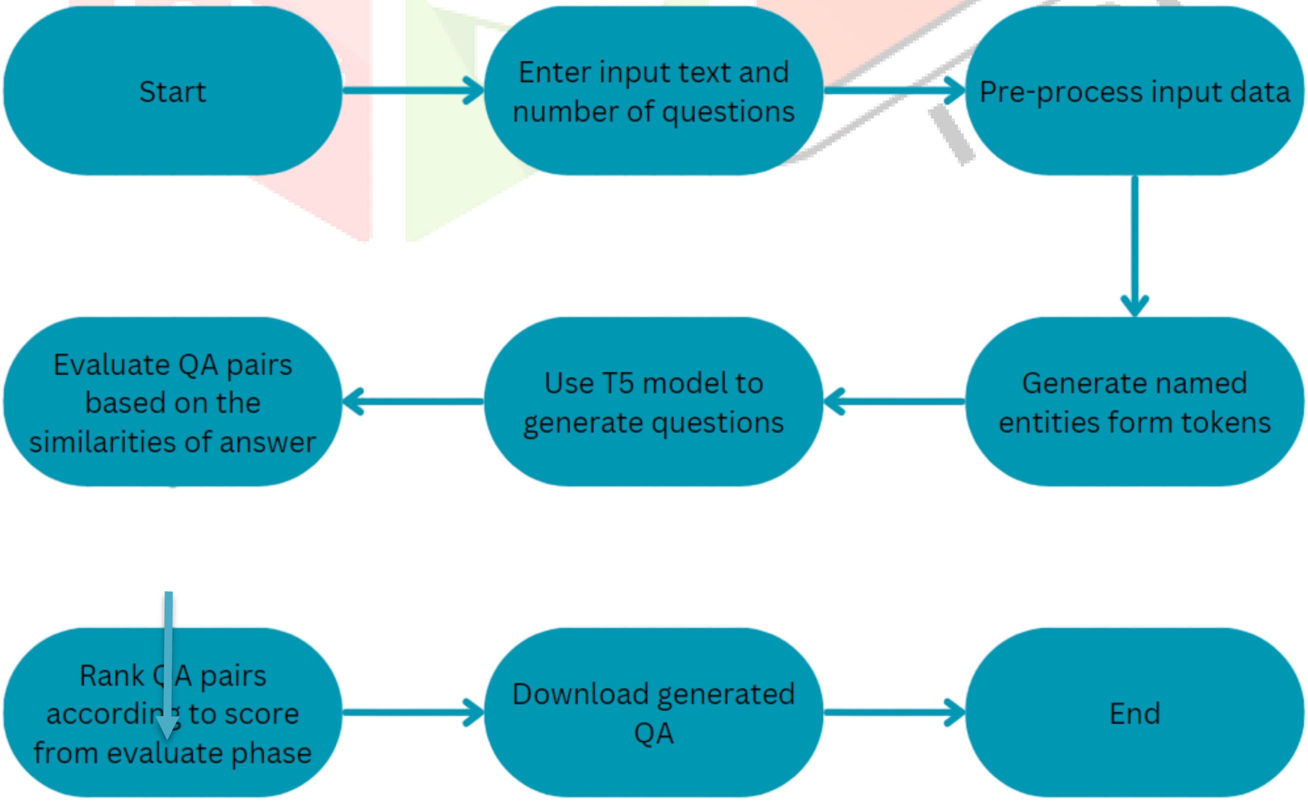

Fig. 2: QA Generation Flow

ADVANTAGES OF PROPOSED SYSTEM

- **Time Efficiency:** Automated question generation can save a significant amount of time compared to manual question creation. This is particularly useful for educators, content creators, or anyone involved in developing assessments.
- **Consistency:** The system ensures consistency in question formulation, avoiding variations in wording or difficulty levels that might arise with different human authors. This can contribute to fair and standardized assessments.
- **Scalability:** Easily scale the generation of questions to accommodate a large number of topics, subjects, or difficulty levels. This scalability is especially valuable in educational settings with diverse content requirements.
- **Personalization:** The system can be designed to adapt to individual learner needs by generating questions that match the learner's proficiency level. This personalization enhances the learning experience.
- **Adaptability:** The system can be trained on different datasets or fine-tuned to specific domains, making it adaptable to various educational or training contexts. It can be used for a wide range of subjects and topics.
- **Immediate Feedback:** If integrated into an educational platform, the system can provide instant feedback to learners, aiding in their understanding of concepts and helping them identify areas for improvement.
- **Resource Savings:** Automated question generation can reduce the need for extensive question banks or libraries, as the system can generate diverse questions on-demand. This can lead to resource savings in terms of storage and maintenance.
- **Language Proficiency Assessment:** The system can be used for assessing language proficiency by generating questions that evaluate various language skills, such as grammar, vocabulary, and comprehension.
- **Customization:** Users can customize the system to generate questions based on specific criteria, such as learning objectives, Bloom's taxonomy levels, or other pedagogical considerations.
- **Innovative Learning Tools:** Integration with other educational technologies or platforms can lead to the development of innovative learning tools, enhancing the overall educational experience.
- **Data-Driven Insights:** The system can provide valuable data on learner performance and areas of difficulty, which can be used to inform instructional design and improve educational strategies.
- **Continuous Improvement:** Through feedback mechanisms and iterative updates, the system can continuously improve its question generation capabilities, ensuring relevance and effectiveness over time.

RESULT

The goal of our project was to develop a natural language processing (NLP) model for question-answer generation, specifically focusing on generating questions with one-word answers. We employed state-of-the-art NLP techniques and a carefully curated dataset for training and evaluation. To assess the performance of our model, we used precision, recall, and F1 score as our primary evaluation metrics. These metrics provide a comprehensive view of the model's ability to generate accurate and relevant one-word questions.

```
Along with sons and daughters-in-law, who did mike have?
Grandchildren

Why did mike leave his village?
Tired

What did morris own?
Shop

What did mike do when he couldn't feed his family?
Decided

What type of farmer was mike?
Poor

Where was the largest jewelry shop located?
Village

Along with sons and grandchildren, who else did mike have?
Daughters

Who owned the largest jewelry shop in the village?
Morris
```

Fig. 3: Final Output

FUTURE SCOPE

- **Integration with test creation websites:** One avenue for future development is integration with test websites. This development will support teachers in understanding the questions and answers on these platforms. Thus, teachers and professors can try to create tests for students, thereby increasing the accessibility and effectiveness of the learning process.
- **Different types of input:** Another great area for future research is ensuring the system accepts different types of input, including images, URLs, and other formats. This extension will improve the performance of the system and enable better resolution of problems arising from different access types.
- **Bloom's Taxonomy Mapper:** Future improvements may include the development and integration of Bloom's Taxonomy Mapper module. This change divides the questions according to Bloom's classifications such as knowledge, understanding, application, analysis, connection and evaluation. This report will provide a better understanding of the complexity of the problem and make the system useful for learning.

- **Multi-word Answer Generation:** Explore ways to extend your model to generate questions for answers that are more than one word. This could involve more complex linguistic structures and could be beneficial in a wider range of applications.
- **Context-aware Question Generation:** Enhance your model to consider context when generating questions. This could involve incorporating information from the surrounding text or dialogue context to create more meaningful and contextually relevant questions.
- **Fine-tuning for Specific Domains:** Investigate the possibility of fine-tuning your model for specific domains or industries. This could make your system more versatile and applicable to different fields such as medicine, law, finance, etc.
- **User Feedback Integration:** Consider integrating user feedback mechanisms into your system. This could involve collecting feedback on the generated questions and using it to improve the performance and user satisfaction over time.
- **Interactive Question Generation:** Explore the development of an interactive system where users can provide input or guidance during the question generation process. This could lead to more user-specific and contextually relevant questions.
- **Evaluation Metrics:** Develop and experiment with new evaluation metrics to assess the quality of the generated questions. This is crucial for understanding how well your system performs and for comparing it with other question generation models.
- **Real-world Applications:** Explore practical applications of your system in real-world scenarios. For example, it could be used in educational settings for generating quiz questions, in customer support for automating responses, or in chatbots for more interactive conversations.
- **Integration with Conversational Agents:** Consider integrating your question generation model with existing conversational agents. This could enhance the conversational capabilities of such systems, making them more dynamic and engaging.
- **Ethical Considerations:** Investigate the ethical implications of your question generation system. Consider issues related to bias, fairness, and potential misuse. Propose methods to mitigate biases and ensure fairness in question generation.
- **Multilingual Question Generation:** Extend your project to support multiple languages. This could involve training your model on multilingual datasets and addressing the challenges associated with generating questions in different languages.

CONCLUSION

During the development of the system, we extensively researched various methodologies employed in papers focusing on automating Question Generation. The system is designed to take a textual paragraph as input and generate questions whose answers are contained within the paragraph. Prior to question generation, the input text undergoes preprocessing, encompassing both syntactic and semantic analyses. Syntactic analysis involves Part-of-Speech (POS) tagging and Chunking, while Semantic Analysis incorporates Named Entity Recognition (NER). Following preprocessing, the system maps appropriate 'wh' question words with the text, leading to the generation of questions. Future enhancements for this system include improving question generation accuracy, implementing a storage mechanism for generated questions in a database for reuse, applying Bloom's taxonomy to create questions of varying difficulty levels. Additionally, the system aims to accommodate input from PDF documents, expanding its applicability. Given the relatively low level of automation in the education sector, this paper serves as a foundational step towards future automation in this domain.

REFERENCES

[1] O. KEKLIK, "AUTOMATIC QUESTION GENERATION USING NATURAL LANGUAGE PROCESSING TECHNIQUES,"2018.Accessed:Nov.20,2022.[Online].Available: http://openaccess.iyte.edu.tr/bitstream/handle/11147/6938/T001801.pdf?sequence=1&isAllowed=y

[2] Aleena, Vidya – "Implementation of Automatic Question Paper Generator System", International Research Journal of Engineering and Technology (IRJET), Feb 2019. https://www.irjet.net/archives/V6/i2/IRJET-6I297.pdf

[3] Priti G, Shreya J, Srushtee, Subhangi – "Automated Question Generator System: A Review", International Journal of Engineering Applied Science and Technology (IJEAST), Dec 2019, Vol. 4, Issue 8, Pages 171-176; ISSN No. 2455-2143. DOI: 10.33564/IJEAST.2019.v04i08.027

[4] D. R. Ch and S. K. Saha, "Automatic Multiple Choice Question Generation from Text: A Survey," IEEE Transactions on Learning Technologies, vol. 13, no. 1. Institute of Electrical and Electronics Engineers, pp. 14–25, Jan. 01, 2020. doi: 10.1109/TLT.2018.2889100.

[5] Ankita and K. A. Abdul Nazeer, "Part-of-speech tagging and named entity recognition using improved hidden markov model and bloom filter," in 2018International Conference on Computing, Power and Communication Technologies,GUCON 2018, Mar. []2019, pp. 1072–1077. doi: 10.1109/GUCON.2018.8674901.

[6] Amruta Umardand, Ashwini – "A survey on Automatic Question Paper Generation System", International Advanced Research Journal in Science, Engineering and Technology (IARJSET), Jan 2017. https://www.researchgate.net/publication/313812481 _A_Survey_on_Automatic_Question_Paper_Generation_System

[7] Kalpana B. Khandale1, Ajitkumar Pundage, C. Namrata Mahender – "Similarities In Words Using Different Pos Taggers.", IOSR Journal of Computer Engineering (IOSRJCE),(PP 51-55). https://www.iosrjournals.org/iosrjce/papers/Conf.17003/Volume-1/10.%2051- 55.pdf?id=7557

[8] Edward Loper and Steven Bird – "Nltk: The Natural Language Toolkit.", July 2002. https://www.researchgate.net/publication/220482883 _NLTK_the_Natural_Language_Toolkit

[9] Ankita, K. A. Abdul Nazeer – "Part-Of-Speech Tagging And Named Entity Recognition Using Improved Hidden Markov Model And Bloom Filter", International Conference on Computing, Power and Communication Technologies (GUCON), 2018. https://ieeexplore.ieee.org/document/8674901

[10] Mohd Shahid Husain – "Automatic Question generation from Text", International Journal for Innovations in Engineering, Science and Management (IJIESM), April 2015. https://www.researchgate.net/publication/27688 6742_Automatic_Question_Generation_from_Text

[11] Min-Kyoung Kim, Han-Joon Kim, "Design of Question Answering System with Automated Question Generation", 2008 IEEE 28th International Symposium on Computer-Based Medical Systems.

[12] Susmita Gangopadhyay, Ravikiran S.M, "Focused Questions and Answer Generation by Key Content Selection", 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)

[13] Mai Mokhtar, "Automatic Question Generation Model Based On Deep Learning Approach", International Journal of Intelligent Computing and Information Sciences, 2021

[14] Riken Shah, Deesha Shah, Prof. Lakshmi Kurup, "Automatic Question Generation for Intelligent Tutoring Systems", 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)