MCQ Generation from Given Text

Shritej Jagtap (47) Saniya Sayed (42) Raj Patil (52) Sarvesh kasar(38)

1. Introduction:

Natural Language Processing has emerged as a transformative technology, enabling computers to understand and process human languages. One of the key applications of NLP is automatic question generation, which is beneficial in various domains, including education, assessment, and language learning. In this mini-project, we focus on developing a system that generates multiple-choice questions (MCQs) in Marathi from a given English text.

* 1. Aim:

This project aims to bridge the gap between English content and Marathi language learners by automatically translating key concepts into questions, providing a means for assessment and knowledge testing in a regional language. Given the increasing demand for educational resources in local languages, this approach can serve as a valuable tool for educators and students alike.

* 1. Objectives:

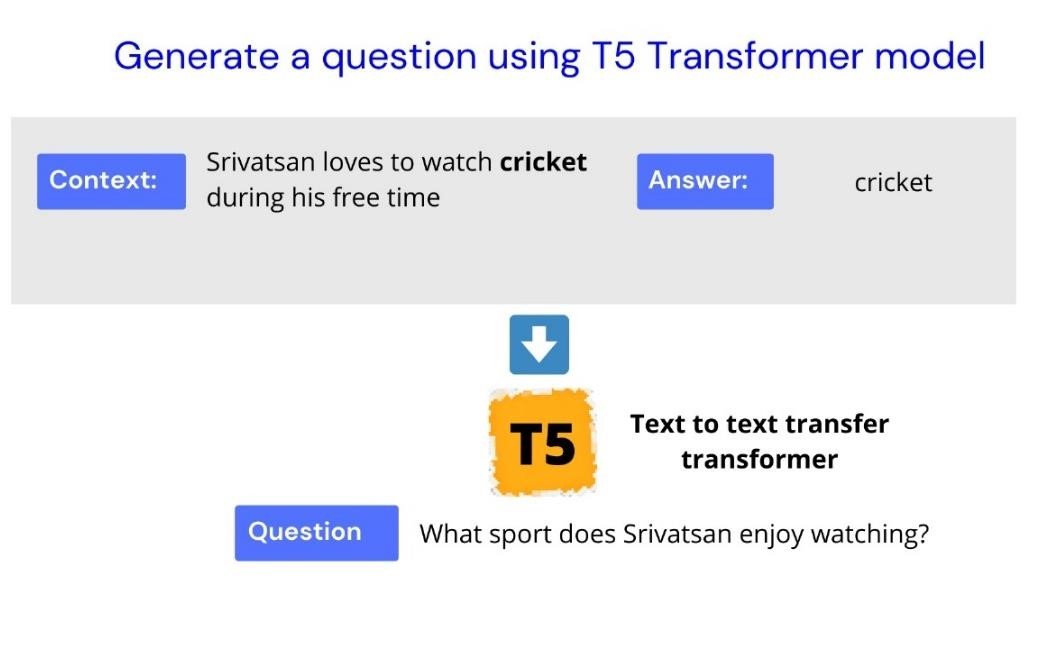
* + - **Automate MCQ Generation**: Develop an NLP-based system to generate multiple-choice questions in Marathi from given English text.
    - **Ensure Semantic Accuracy**: Maintain the meaning and context of the original English text in the translated Marathi questions.
    - **Promote Multilingual Learning**: Provide a tool for generating educational content in Marathi to aid language learners and educators.
    - **Enhance Assessment Methods**: Offer a scalable solution for creating MCQs, reducing manual effort in crafting questions for assessments.
    - **Leverage NLP Techniques**: Utilize translation, text processing, and question generation algorithms for effective question creation.

1. Methodology:

1. Data Collection

Input Multilingual Text: The system accepts a block of text in any supported language (e.g., Marathi, Hindi, Gujarati) which is translated to English using the Google Translator API as the basis for MCQ generation.

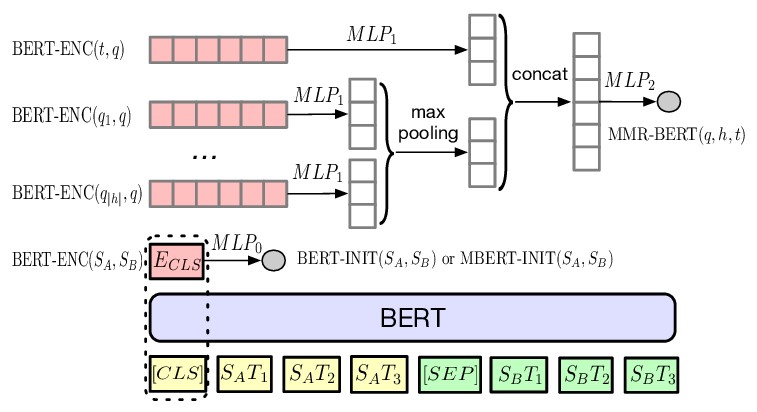
1. Text Processing and Key Concept Extraction
   * Tokenization: The translated text is broken down into sentences and words using the NLTK library's sent\_tokenize and tokenization methods.
   * Part-of-Speech (POS) Tagging: The spaCy library is used to identify key parts of speech (nouns, verbs, adjectives, etc.) from the text, focusing on extracting key nouns and concepts.
   * Named Entity Recognition (NER): Important entities such as names, places, dates, and other nouns are identified using spaCy for better question formation.
   * Key Concept Identification: The system uses PKE Multipartite Rank and Sentence Transformers to extract key phrases that serve as answers for the generated questions.
2. Question Generation
   * + Summarization: A T5-based model (T5ForConditionalGeneration and T5Tokenizer) is utilized to summarize long text into concise, meaningful paragraphs that are more suitable for MCQ generation.



* + - Question Formation: For each identified key concept, a T5-based question generation model forms "Wh-" questions (e.g., Who, What, Where) in English.
    - Distractor Generation: Sense2Vec is used to generate plausible distractors based on semantic similarity.

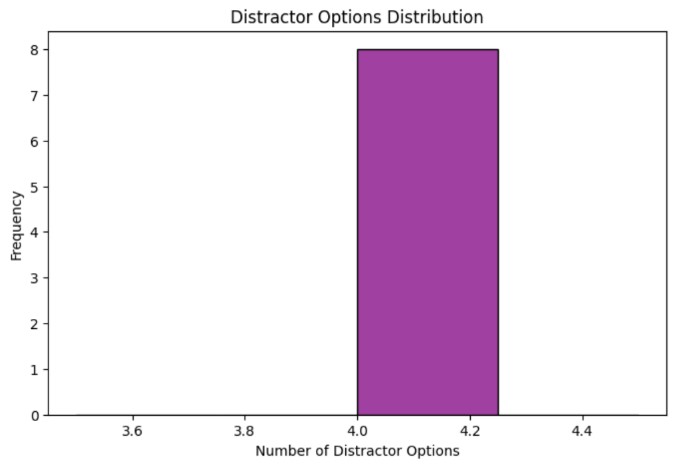
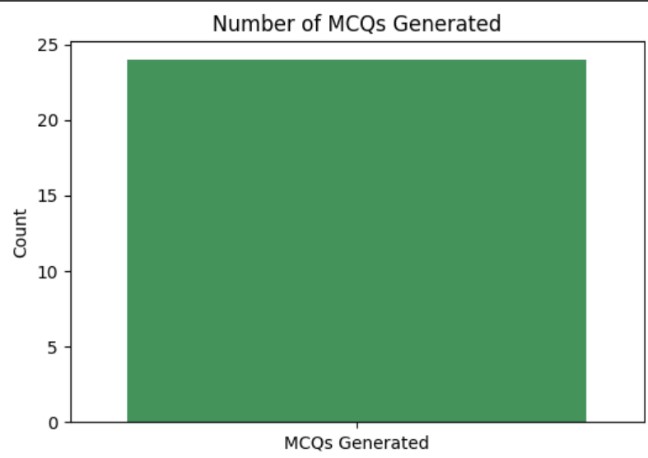
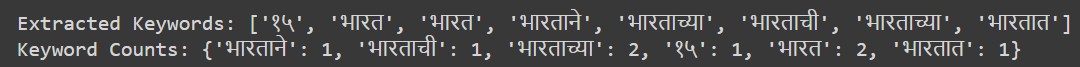
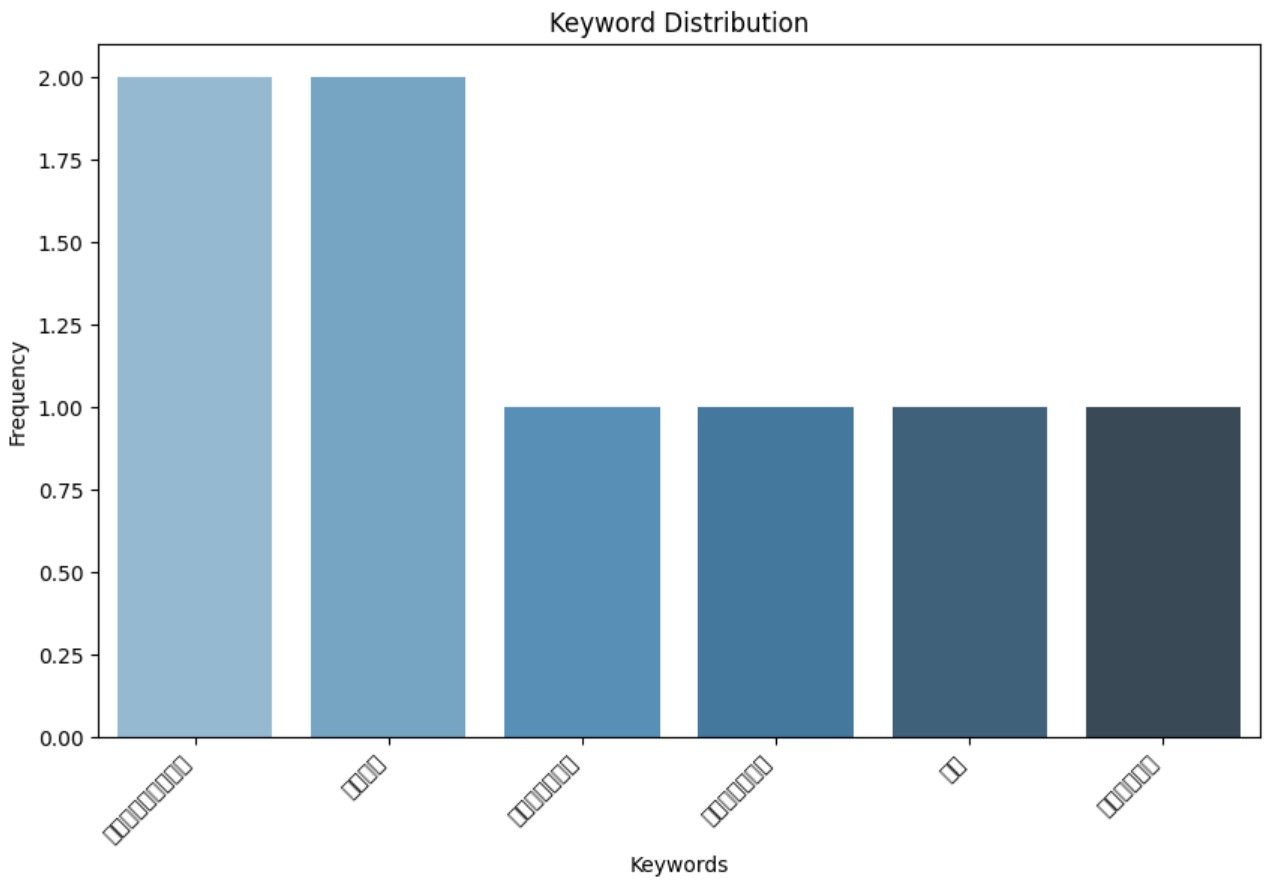
WordNet provides hypernyms/hyponyms for concept words to serve as alternative distractors.

* + - MMR (Maximal Marginal Relevance) is employed to ensure that the distractors generated are semantically relevant but distinct from the correct answer.



1. MCQ Structuring
   * + Final MCQ Formatting: The MCQs are structured with one correct answer and three distractors. The questions are clearly formatted, ensuring linguistic consistency in the output text.
     + Translation to Target Language: Once the MCQs are generated, they are translated back into the target language (e.g., Marathi, Gujarati) using the Google Translator API.This methodology uses advanced NLP models like BERT-based Sentence Transformers, Sense2Vec, and T5 transformers to ensure high-quality, multilingual MCQ generation, with accurate distractors and wellformed questions.

3) Data Visualization



Conclusion:

This project presents a comprehensive methodology for generating highquality multilingual MCQs from input text using advanced NLP techniques. By leveraging machine translation, tokenization, POS tagging, NER, and key concept extraction, the system ensures accurate translation and concept identification. The use of T5-based models for question generation, combined with Sense2Vec and WordNet for distractor creation, results in well-formed MCQs with plausible alternatives. The final MCQs are translated back into the target language, ensuring linguistic accuracy. Overall, the methodology provides an efficient solution for multilingual MCQ generation, enhancing educational content across diverse languages.

Front End:

