MIT-WPU

।। विश्वशान्तिर्ध्रुवं ध्रुवा ।।

Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

CCL - NLP

MINI PROJECT SYNOPSIS

GROUP MEMBERS -

23 - Siddharth Gunjal (1032222796)

29 - Samarth Mane (1032222870)

30 - Hari Maheshwari (1032222872)

PROBLEM STATEMENT -

"Develop an NLP-based system that would automatically generate multiple-choice assessment questions from textual content to enhance educational resources with diverse and contextually relevant questions for assessment."

OBJECTIVES -

- **Automate Question Creation:** Develop an algorithm to automatically generate multiple-choice questions from a given text.
- **Content Validation**: Ensure that the generated MCQs maintain a high level of accuracy, relevance, and difficulty appropriate to the subject matter.
- **Optimize Distractors:** Create effective distractors (wrong answer choices) that are plausible yet clearly incorrect.

FUNCTIONAL FEATURES

- **Text Input**: Allows users to input text or upload documents for MCQ generation.
- **Key Concept Extraction**: Extracts key concepts, terms, and facts from input text.
- **Automatic Question Generation**: Automatically generates MCQs covering essential details.
- **Distractor Generation**: Creates plausible incorrect options to challenge learners.
- **Topic-Based Question Generation**: Focuses question generation on specified topics or keywords.
- **Batch Processing**: Processes multiple documents simultaneously to generate large MCQ sets.

INNOVATIVE FEATURES -

- **Adaptive Learning**: Adjusts question difficulty based on learner responses, Serving MCQs to individual knowledge levels.
- **Contextual Feedback**: Provides immediate feedback for each answer, explaining correct and incorrect choices to reinforce key concepts.
- **Semantic Similarity**: Ensures distractors are closely related to the correct answer, making questions more challenging and reducing guesswork.
- **Cross-Topic Questions**: Generates questions that connect concepts across different sections, enhancing integrative thinking and deeper understanding.

LITERATURE SURVEY -

S. No.	Method Used	Data Used	Result Achieved	Advantages	Limitations	Reasearch Gap
1.	NER + Semantic Similarity	News Articles, Wikipedia, Academic Texts	Generates questions with semantically relevant distractors	Produces challenging distractors	Relies on high-quality NER , it can miss subtle nuances	Difficulty in handling ambiguous entities
2.	Template based Generation + NLP	Academic Textbooks	Produces MCQs with consistent formatting	easy to customize	Not adaptable to varied content	Limited adaptability to different educational contexts
3.	GPT-based Text Generation + Distractor Generation	Open Educational Resources (OER)	Creative MCQs with plausible distractors	High variability in question generation	May produce nonsensical questions	Needs better control over question difficulty
4.	Machine Learning + Human-in-the-Loop	Diverse Educational Datasets	High-quality MCQs refined through human feedback	Combines machine efficiency with human judgment	Slower than fully automated systems	Finding the right mix between machine automation and human input.

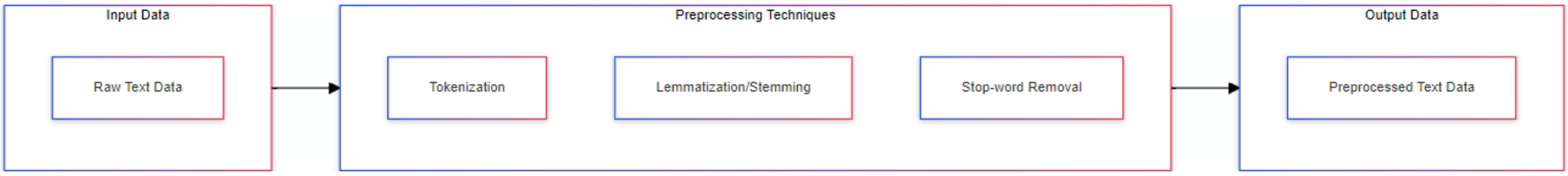
METHODOLOGY -

A) Text Preprocessing

Techniques Used :- Tokenization, Lemmatization/Stemming, Stop-word Removal

Why: Text preprocessing prepares the input data for analysis by breaking it down into manageable units (tokens), reducing words to their base forms, and removing common words that don't contribute to the meaning. This ensures that the subsequent analysis is accurate and focused on meaningful content.

High Level Diagram :-

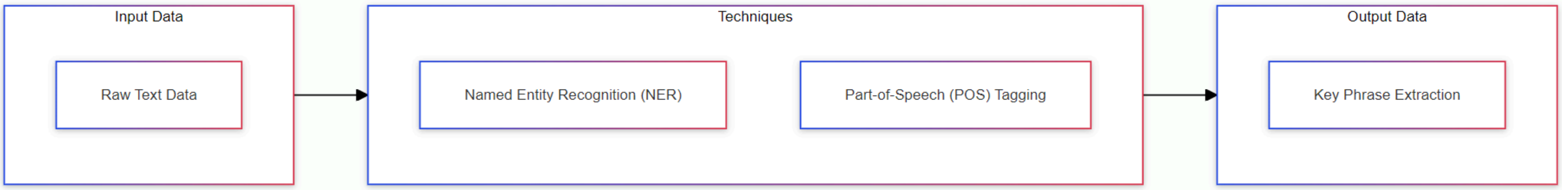


B) Text Analysis and Concept Extraction

Techniques Used :- NER , Part-of-Speech Tagging, Key Phrase Extraction

Why: NER identifies important entities such as names, dates, and places, while Part-of-Speech Tagging helps in understanding the grammatical structure of sentences. Key Phrase Extraction focuses on extracting significant phrases and terms from the text, which are essential for formulating relevant questions.

High Level Diagram :-

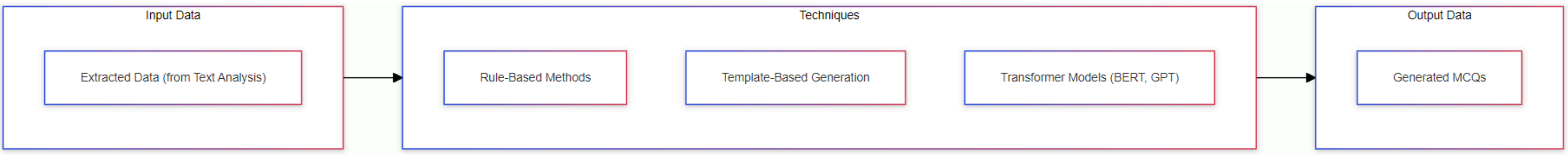


C) Question Generation

Techniques Used: Rule-based Methods, Template-based Generation, Transformer Models (e.g., BERT, GPT)

Why: Rule-based and template-based methods ensure that questions adhere to a specific format and are relevant to the extracted concepts. Transformer models enhance question generation by leveraging deep learning to understand context and generate more sophisticated and varied questions.

High Level Diagram :-

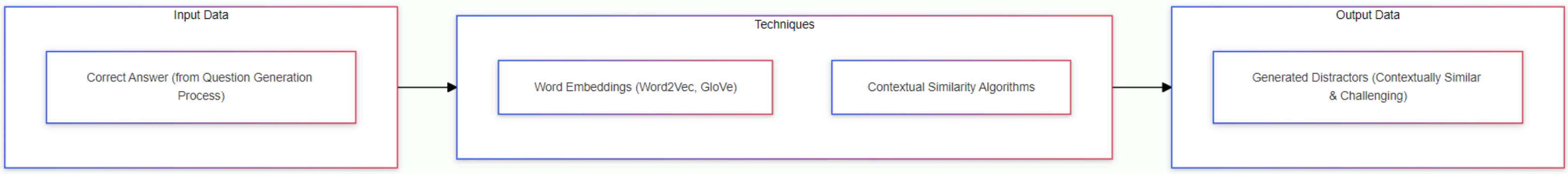

Input Data
Extracted Data (from Text Analysis)
Techniques
Rule-Based Methods
Template-Based Generation
Transformer Models (BERT, GPT)
Output Data
Generated MCQs

D) Distractor Generation

Techniques Used: Word Embeddings (e.g., Word2Vec, GloVe), Contextual Similarity Algorithms

Why: Word embeddings capture semantic relationships between words, enabling the system to generate distractors that are contextually similar to the correct answer. This makes the distractors plausible and increases the challenge for learners.

High Level Diagram :-


Input Data
Correct Answer (from Question Generation Process)
Techniques
Word Embeddings (Word2Vec, GloVe)
Contextual Similarity Algorithms
Output Data
Generated Distractors (Contextually Similar & Challenging)

REFERENCES -

[1] https://ieeexplore.ieee.org/document/6482468

[2] https://arxiv.org/abs/0906.4550

[3] https://ieeexplore.ieee.org/document/9412734

[4] https://ieeexplore.ieee.org/document/9382153

[5] https://doi.org/10.1109/ACCESS.2021.3060940

[6] https://ieeexplore.ieee.org/document/9992001