

NLP Project Phase #2

(Project High Level Architecture and Literature Survey)

Project Title : Automatic MCQs generator from text

Project ID : 8

Project Team :	Name	SRN
1.	B Pravena	PES2UG19CS076
2.	Spoorthi R	PES1UG19CS500
3.	Varna S	PES2UG19CS448

Outline

- Abstract
- Motivation Scope of the Project
- Suggestions from Review - 1
- Literature Survey (6 papers) (Individual information is mandatory)
- Summary: Learning from Literature Survey
- High Level Architecture /Detailed Flowchart (DIAGRAM)
- Work Done So far
- References

Abstract

Automatic generation of multiple choice questions based on a given piece of text.

Multiple-choice questions are the most popular assessment questions created whether it is for a school test or a graduate competitive exam. Given the increased volume of workload on teachers/assessment creators due to online learning, it would be very helpful if we can create an automated system to create MCQs.

Motivation and Scope of the Project

The use of automated systems in second-language learning could substantially reduce the workload of human teachers and test creators. Automatic multiple choice question (MCQ) generation from a text is a popular research area. MCQs are widely accepted for large-scale assessment in various domains and applications. However, manual generation of MCQs is expensive and time-consuming. Therefore, researchers have been attracted toward automatic MCQ generation. Many systems have been developed for MCQ generation.

Suggestions from Review - 1

- Provide the suggestions and remarks given during first review:
 - Size of input text

- Mention the feasibility on the same showing the progress.
 - Size of input limited to 1000 words based on average size taken in the literature survey

Individual Information on Literature Survey

S.N.	Name of the Student	SRN	Paper	Paper Title
1	B Pravena	PES2UG19CS076	1	A framework for automatic question generation from text using deep reinforcement learning
			2	Distractor Generation for Multiple Choice Questions Using Learning to Rank
2	Varna Satyanarayana	PES2UG19CS448	3	Intelligent Questions and Summary Generator - Examally
			4	An automated multiple-choice question generation using Natural Language Processing Techniques
3	Spoorthi R	PES2UG19CS500	5	Automated MCQ generator using natural language processing
			6	Literature review of automatic question generation systems

Paper-1

Literature Survey

Paper Details (Citation)	Objective of paper, Techniques/Methods	Detailed explanation along with results	Limitations
Ramakrishnan, Ganesh & Li, Yuan-Fang. (2018). A framework for automatic question generation from text using deep reinforcement learning.	<p>They have used a novel generator-evaluator framework that directly optimizes objectives that reward semantics and structure.</p> <p>Technique - deep reinforcement learning based framework</p>	<p>Generator generates a question-answer pair and explicitly factors in question syntax and semantics, identifies pivotal answers, recognizes contextually important words and avoids meaningless repetitions. Evaluator evaluates the generated pair. From reinforcement learning (RL) point of view, the generator is the agent and the generation of next word is an action. Two novel reward functions evaluate the syntax of the question and semantics of question-answer pair in its entirety.</p> <p>The evaluator makes use of task-specific scores, such as BLEU, GLEU, ROUGE-L, DAS that are naturally suited to QG and other seq2seq problems. Using evaluation metrics as the reward function during reinforcement based learning improves performance for all metrics</p>	<p>Adding QG-specific rewards significantly improves question quality in human evaluation, but less noticeable improvements in automatic evaluation.</p> <p>Performance of the model on automatic evaluation metrics does not improve with DAS as the reward function, though the quality of questions improves.</p>

Paper-2

Literature Survey

Paper Details (Citation)	Objective of paper, Techniques/Methods	Detailed explanation along with results	Limitations
Liang, Chen & Yang, Xiao & Dave, Neisarg & Wham, Drew & Pursel, Barton & Giles, C. (2018). Distractor Generation for Multiple Choice Questions Using Learning to Rank. 284-290. 10.18653/v1/W18-0533.	<p>To investigate how machine learning models, specifically ranking models, can be used to select useful distractors for multiple choice questions.</p> <p>Techniques - Feature-based models and NN based models.</p>	<p>Feature-based Models -:</p> <ul style="list-style-type: none"> • Feature Description - Given a tuple (q, a, d), feature-based model first transforms it to a feature vector $\phi(q, a, d) \in \mathbb{R}^d$ with the function ϕ. • Classifiers - Logistic Regression, Random Forest, LambdaMART <p>NN-based Models -: Uses GAN which has 2 components - generator G and discriminator D. G aims to capture the conditional probability of generating distractors given stems and answers $P(d q, a)$. D estimates the probability that a distractor sample comes from the real training data rather than G. Decomposable Attention model is used to measure the similarities between q and d whereas cosine similarity is used between a and d. 2- stage cascading is then used on the above.</p> <p>First Stage Ranker reduces the candidate size for the later stage while achieving a relatively high recall. The proposed ranking models perform better than unsupervised similarity based methods.</p>	<p>NN performs worse than feature-based models. The main reason is that NN is solely based on word embeddings. Given the limited training examples (SciQ and MCQL), it is difficult to learn a robust end-to-end NN-based model.</p>

Literature Survey

Paper-3

Paper Details (Citation)	Objective of paper, Techniques/Methods	Detailed explanation along with results	Limitations
A., J. & D., A. & A., W.. (2020). Intelligent Questions and Summary Generator - Examally. International Journal of Computer Applications. 177. 1-6. 10.5120/ijca2020919663.	<p>This paper presents an intelligent question generation and lesson summarization system which allows the student to get a summarization or a set of questions based on a specific lesson. Students study more efficiently. Save teacher's time by automatic generation of questions based on text or audio recording of lesson.</p> <p>Makes use of extraction based summarization and TF IDF and POS tagging is used for keyword identification</p>	<p>Language model and the Acoustic model is used to extract text from discussions or text is extracted from a pdf file.</p> <p>Preprocessing includes stemming, lemmatization, stop word removal and POS tagging.</p> <p>Sentence is split into subject, verb and object.</p> <p>Fill in blanks question, yes or no questions, MCQs are generated based on provided template.</p> <p>Fill in the blank questions formed by definition of the keyword,</p> <p>yes/no questions formed by taking opposite meaning of verb.</p> <p>MCQ distractors formed by taking words most similar to keyword from knowledge base.</p> <p>Works with an accuracy of 85%</p>	<p>Distractor is produced based on knowledge base therefore can be an easy give away of right answer as the only word present in the passage can be the answer</p>

Literature Survey

Paper-4

Paper Details (Citation)	Objective of paper, Techniques/Methods	Detailed explanation along with results	Limitations
Nwafor, Chidinma & Onyenwe, Ikechukwu. (2021). An Automated Multiple-Choice Question Generation using Natural Language Processing Techniques. International Journal on Natural Language Computing. 10. 1-10. 10.5121/ijnlc.2021.10201.	<p>Paper presents Automatic multiple-choice question generation. The paper is based on the generation of Multiple Choice Question Generation through Natural Language Processing Techniques.</p> <p>The manually extracted keywords by the teacher were compared to the auto-generated keywords and the result shows that the system was capable of extracting keywords from lesson materials in setting examinable questions.</p> <p>The NLP processes are applied using TFIDF and N-gram.</p>	<p>The text is split into sentences. The split sentences are tokenized, from which the corpus are built as TF-IDF and N-gram mode. The text is tokenized and stop words are removed. And word normalization is done.</p> <p>Dataset had keywords manually marked by teachers to set the gold-standard (keywords extraction). Auto generated keywords are compared to the gold standard of the document.</p> <p>we represent each sentence in a lesson material as a document</p> <p>Term Frequency-Inverse Document Frequency (TF-IDF) is used for stop-words filtering in various subject fields including text summarization and classification.</p> <p>TF- number of times term occurs</p> <p>IDF - how important the term is</p> <p>Precision answers the question – The keywords extracted from the lecture materials are they important words according to the gold standard marked by the teachers?</p> <p>recall answers the question - are all the important words in the lecture materials extracted?</p> <p>the number of multi-choice questions (MCQs) to be generated is dependent on the number of keywords extracted</p>	Makes use of teachers/human intervention to validate and train the model's output.

Literature Survey

Paper-5

Paper Details (Citation)	Objective of paper, Techniques/Methods	Detailed explanation along with results	Limitations
International Research Journal of Engineering and Technology (IRJET) Automated MCQ Generator using Natural Language Processing .	<p>In this paper, the text is first summarized using the BERT algorithm, and accordingly sentence mapping is done for generating MCQs. In order to generate choices for the questions, distractors are generated using wordnet (A lexical database for English).</p> <p>As the BERT algorithm has much better performance over other legacy methods as well as it can process a large amount of data in less time, it will enhance the speed of generating MCQs from given text.</p> <p>Techniques: BERT algorithm, Wordnet, PKE, Deep learning</p>	<p>Sentence selection plays a crucial role in the automatic MCQ generation task. Hence for summarizing the text, BERT Algorithm is used. BERT (Bidirectional Encoder Representations from Transformers) is a neural network-based technique for natural language processing. It is a pre-trained open-sourced model from Google. Rapid Automatic Keyword Extraction (RAKE) is a wellknown keyword extraction method that uses a list of stopwords and phrase delimiters to detect the most relevant words or phrases in a piece of text.</p> <p>For generating distractors, the wordnet approach is used.</p>	<p>Since our proposed system is based on Google's BERT Model, the accuracy of the system will increase in the future as the performance of the model is improved.</p>

Literature Survey

Paper-6

Paper Details (Citation)	Objective of paper, Techniques/Methods	Detailed explanation along with results	Limitations
International Journal of Scientific and Research Publications, Volume 5, Issue 1, January 2015 Literature Review of Automatic Question Generation Systems .	This paper represents the study of automatic objective question generation through NLP or Statistical pattern and review of research work carried out so far. Paper reviews different methodologies with respect to the phases of question generation.	Automatic question generation on the basis of <ul style="list-style-type: none"> the discourse connectives Using Software Agents for Technical Institutions Semantic Based Automatic Question Generation Question Bank Based on Predefined Templates,etc <p>many algorithms created and different methodology is used to generate the automatic question generation system. NLP is used to process the text and NER and SRL is used to identify the semantic relation.</p>	As there are different algorithms/methodologies, there is still scope for proposing methodologies by identifying complexities and type of question like one word answer and True or False need to generate.

Summary :Learning from Literature Survey

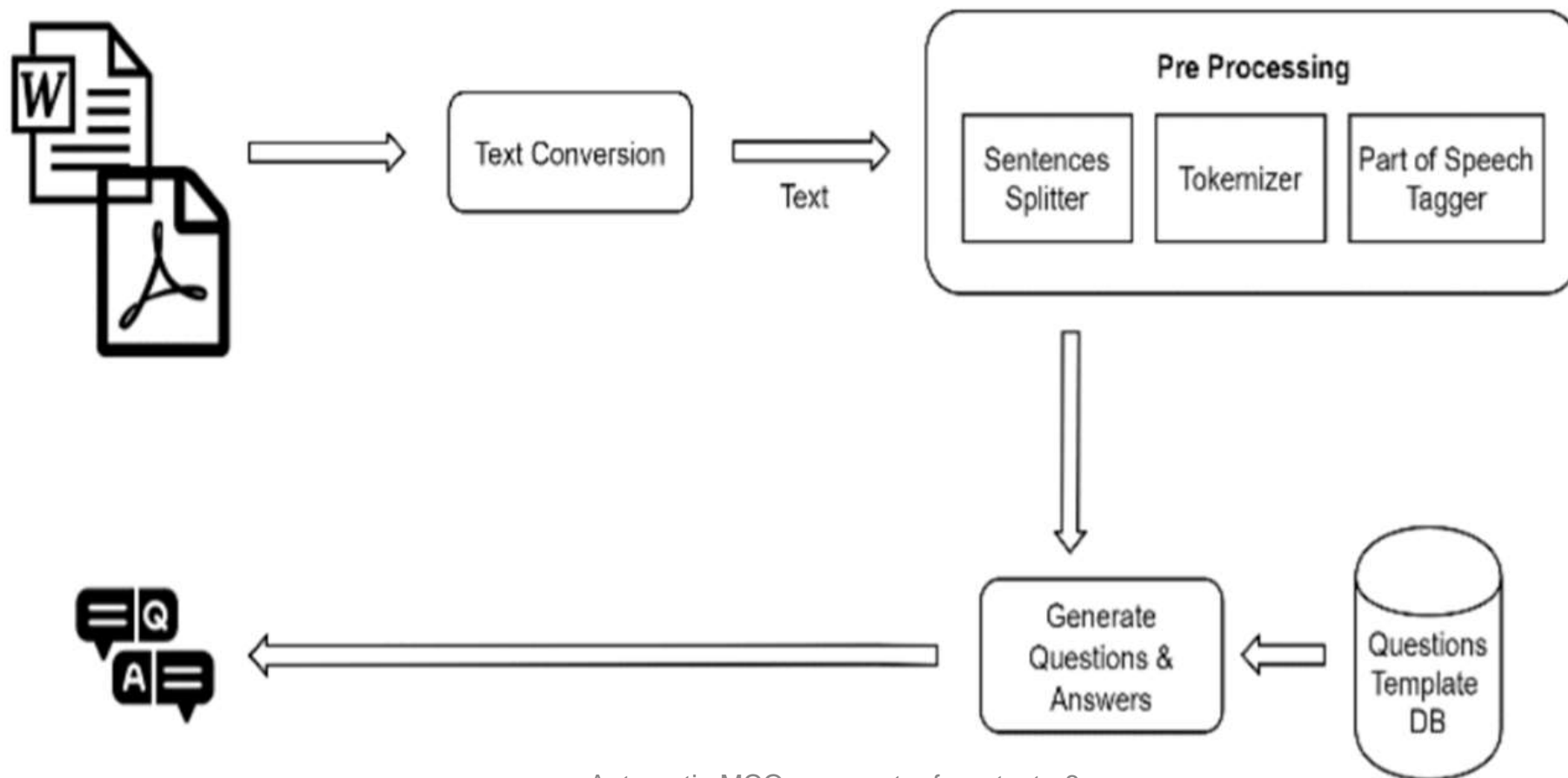
- A conclusion should then state clearly the main conclusions of the review and give a clear explanation of their importance and relevance:
 - TF IDF gives more accuracy in keyword extraction compared to POS tagging
 - Instead of producing blatantly wrong distractors (like opposite words), we produce distractors based on similarity of words in the document or in knowledge base or based on the question
 - Follow template for question generation

Summary :Learning from Literature Survey

- Give a glimpse of the proposed methodology:
 - preprocessing includes: stemming, lemmatization, tokenization, stop word removal, text normalization
 - TF-IDF to extract keywords
 - Distractor generation based on similar words in knowledge bases such as Probase and WordNet or,
 - Distractor generation can also be done by finding most similar words in the document using Glove or word2vec
 - Form questions based on a predefined template

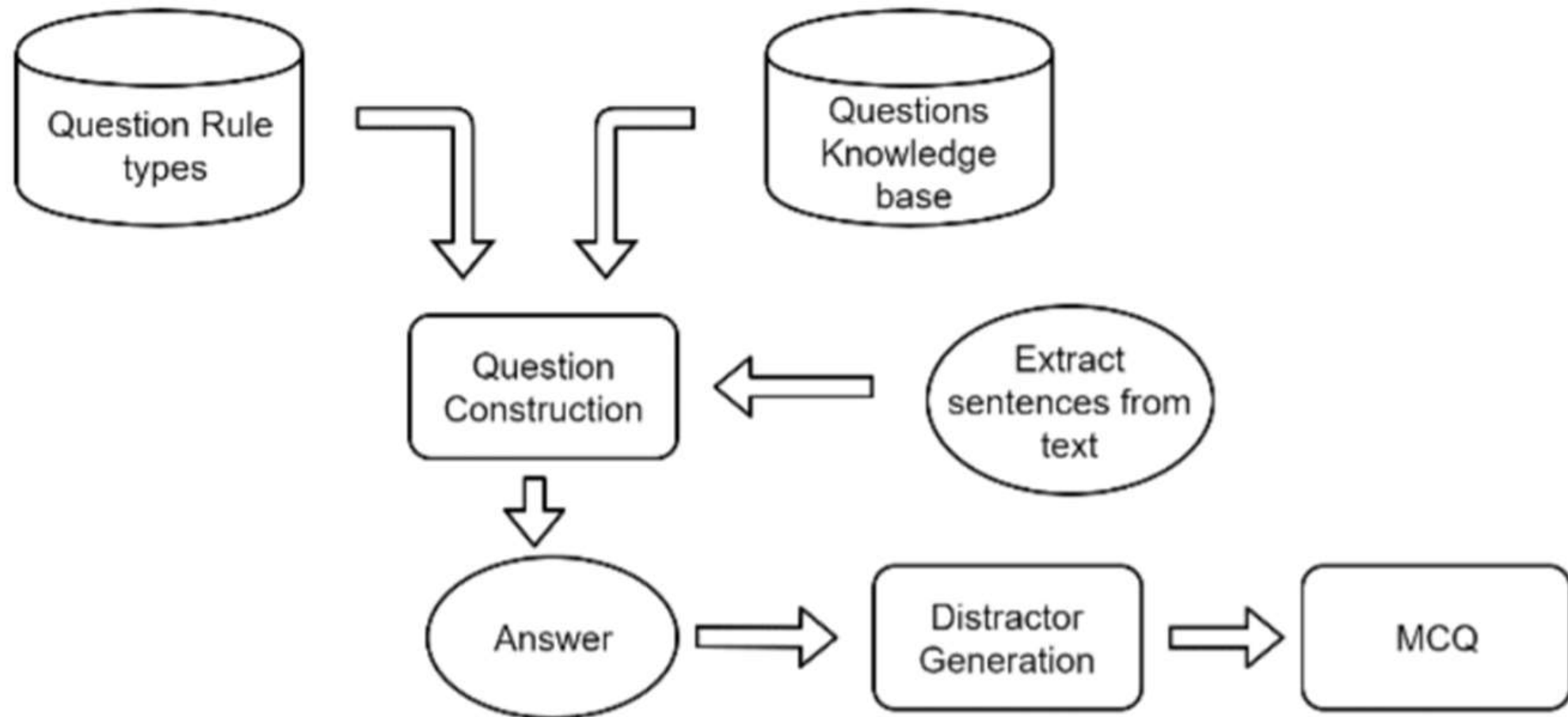
High Level Architecture Diagram/ Detailed Flowchart

Text Extraction & Question Generation Process



High Level Architecture Diagram/ Detailed Flowchart

MCQ Generation Process



Work Done so far

- Chose a dataset - SciQ
- We have done Data Cleaning on our dataset

References

- <https://aclanthology.org/W18-0533.pdf>
- <https://arxiv.org/pdf/1808.04961v2.pdf>
- <https://arxiv.org/ftp/arxiv/papers/2103/2103.14757.pdf>
- <https://www.ijcaonline.org/archives/volume177/number32/jayakody-2020-ijca-919663.pdf>
- <https://arxiv.org/pdf/2004.09853.pdf>
- <https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73>
- <https://allenai.org/data/sciq>
- <https://towardsdatascience.com/the-three-main-branches-of-word-embeddings-7b90fa36dfb9#:~:text=Word2Vec%20takes%20texts%20as%20training,that%20two%20words%20appear%20together.>

Thank
You