Automatic Question and Distractor Generation

B Pravena
CSE Dept.
PES University
Bangalore, India
bpravena11@gmail.com

Varna Satyanarayana
CSE Dept.
PES University
Bangalore, India
varna.satya@gmail.com

Spoorthi R
CSE Dept.
PES University
Bangalore, India
spoorthi9740@gmail.com

***Abstract*—Multiple-choice questions are the most popular form of assessment questions for school tests and graduate competitive exams. Given the increased volume of workload on assessment creators due to online learning, this work will be very helpful to create automatic generation of multiple-choice questions. This paper proposes a model to automatically generate questions and their respective distractors given any reading passage. The given passage is first summarized to retain the most important target words. Then, the questions are generated by using the T5 transformer model which is trained on the SQuAD dataset. Later, by making use of Sense2Vec, BERT and Gensim a model has been proposed which automatically generates 3 different distractors. This work's novelty is that the distractors are based on the passage and not the context from where the answer was taken from . Sense2Vec is used to find words or phrases that are contextually similar to the answer, BERT is used to produce a distractor by masking the answer in it's context and predicting possible words and the Gensim model is used to find the vector representation of the answers and assign a distractor based on the closest keywords from the given passage.**

***Index Terms*—NLP, Question Generation, transformer model, Sense2Vec, BERT, Gensim, target word**

I. Introduction

The use of automated systems in 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. Multiple choice questions (MCQs) are widely accepted for large-scale assessment in various domains and applications. However, manual generation of Multiple choice questions (MCQs) is expensive and time-consuming. Therefore, research is being focused towards automatic their automatic generation.

Within natural language processing, question answering (QA) is a heavily researched field, while the inverse task (question generation) receives much less attention. For education, being able to generate semantically correct and relevant questions is a challenging task. Most papers tend to concentrate only on the task of generating a question from a given context and answer, while other elements such as process of selecting the answer and generating the incorrect answers (distractors), required for multiple choice questions (MCQs) receive much less attention. The erroneous answers to a multiple-choice question are known as distractors.

Choosing the distractors for a specific question is a crucial task because it may make answering the question relatively simple if the distractor is unrelated to the question, or it can make answering the question more difficult if the distractors are good alternative solutions to the question. It also shouldn't be the same response in a different form or tense, as this leads to several right answers.

In this paper, for the process of question generation, the T5 Transformer model has been used. Its a text-to-text transfer model trained on the Stanford Question Answering Dataset (SQuAD) dataset. This dataset has more than 100,000 questions based on articles taken from Wikipedia. The dataset consists of questions, answers and context of passage. This is split into train and valid dataset. We check if there are any answers with length of 7 or more words and remove them from the train and valid dataset.This work uses the T5 transformer model built on the pretrained T5 - base model. The inputted text is first summarized by calculating the frequencies and extracting important text. To extract the keyword we apply Rake on the summarized text. We then map the final keywords to their respective context. This is then passed to the question generation model and then based on the context and answer (target word), questions are generated. The questions along with their answers, the context with respective interests are then passed onto the distractor generator. Target words are the words or phrases identified in the passage that can be answers around which we can form a question. Distractors and questions are generated based on the found target words.

Sense2Vec is a nueral network model that was trained on all reddit comments posted between 2015 and 2019. It is used to give the vector space representations of documents. This model makes use of sense2vec to get contextually similar phrases or words or answers to generate distractors. The Gensim library is used to get the vector representation of phrases which is used to find the most similar phrases from the passage so that they can be used to find the distractors for each answer. This is done to introduce an element of confusion into answering the question. The BERT model is used to predict distractors based on the context of the answers. As BERT is a bidirectional encoder, it considers the context of the answer by looking at the sentence on either side of it. Rather than calling BERT bidirectional, it can be called as non directional as it considers all the words of the sentence together to predict the masked word. This is what makes it perfect to generate good distractors.

.

II. Related Work

In [1], a novel framework of generator - evaluator has been used. The generator produces question-answer pair and specifically ensures that the generated questions are semantically, and syntactically right. It also ensures to pull out important keywords which are of significance in a given context so as to avoid illogical repetitions. The evaluator assess the generated question-answers pair using task specific scores such as GLEU, BLEU, DAS. These scores are naturally suited for question generation and other sequence to sequence tasks. The technique of Deep Reinforcement Learning has been used. So, from this view we can tell that generator is the agent and production of new word is the action. The evaluation metrics are fed in to the model as rewards. This work made use of two novel reward functions to evaluate the syntax of the generated question and then the semantics of the question-answer pair. But, the work noticed that the human evaluation helps improve question quality better than automatic evaluation.

In [2], the work made use of ranking models to generate sensible distractors for generation of multiple choice questions. The work has made use of Feature - based models and Neural Network based models. The feature based models handles description of features and then classifiers. So given a tuple of question, answer and distractor to it, using the function it generates the feature vector of that tuple. It makes use of the Logistic Regression, Random Forest, LambdaMART classifiers. Neural Network based model makes use of GAN which tries to capture the conditional probability of producing the distractors given the question and answers. Decomposable Attention model checks for similarity between questions and distractors whereas cosine similarity checks between answers and distractors.After this, 2 stage decomposing is performed on it. This work makes use of SciQ and MCQL datasets. It was noticed that, the Neural network model performed worse than the Feature based model mainly cause its based on word embeddings. It is also tough to learn a powerful end-to end Neural Network based model becuase it has limited training samples to learn from.

The objective of [3] is to intelligently generate questions and summarise lessons to make teaching and learning more effective. The input can be a passage or an audio recording of the lesson which can be either summarised or can generate questions based on it. The questions are generated by following a template to form the question. Different types of questions such as fill in the blanks, WH questions, etc can be generated. This made use of different NLP techniques such as POS tagging, TF IDF and extraction based summerisation for keyword identification. The distractors are produced by picking out the most similar word to the answer from a knowledge base. Question rule types and question knowledge base is considered for question construction by applying it to the extracted sentences from the text. The answer is also taken from the extracted sentence using which distractor is generated by referring to the knowledge base. An MCQ is therefore formed. Extraction of keywords from text and forming questions based on it and different pre-processing techniques used was considered in forming this paper's model. The disadvantage of this paper lies in the method used for distractor generation as a knowledge base is used to find the most similar words, this can become an easy give away of the right answer and decreases the topics of the passages that can be given as the input.

[4] presents an automatic question generation model where manually extracted keywords are compared to the automatically extracted keyword and the result shows that the system was capable of extracting keywords from lesson materials in setting examinable questions. The NLP processes used to extract keywords are TF-IDF and N-gram techniques. 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. The dataset that was used comprised of keywords manually marked by teachers to set the gold-standard (keywords extraction). Auto generated keywords are then compared to the gold standard of the document. Each sentence in the inputted lesson material is taken as a document. Term Frequency-Inverse Document Frequency (TF-IDF) is used for stop-words filtering in various subject fields including text summarization and classification. TF is the number of times a term occurs, IDF is how important the term is. The number of multi-choice questions (MCQs) to be generated is dependent on the number of keywords extracted. Precision and recall is used to check the performance of their model. This paper focuses more on how the keyword is extracted from a given passage more than the following question and distractor generation. The concept of basing the number of questions generated based on the number of extracted keywords generated and producing the questions based on the keyword and it's context is the motivation for this paper's model.

[5] his paper mainly approached on a BERTSUM model for text summarization, where the BERT algorithm has more accuracy when compared to other methods. Here, for keyword extraction ,the python library RAKE is used. Sentence mapping is done for generating MCQ's. And for the generation of distractors, Wordnet approach is used. So the proposed system creates the automated questions with the help of NLP concept.

[6] This paper proposed different algorithms and methodologies used with respect to the phases of question generation. Here, NLP is used to process the data . NER and SRL are used for identification of the semantic relation. Also the paper presents the review of work for generation of questions automatically from a given text.

III. Motivation

Generating multiple choice questions automatically is now a popular field of study as generating them manually is expensive and even time consuming. Introducing automated learning systems could remarkably reduce the effort of teachers and test producers. This is nowadays a common research topic. MCQs are frequently used in a variety of fields and applications, for large-scale evaluation. Many systems have been created to generate MCQs.

It was noticed from related works, that human evaluation is the best form to evaluate quality of questions that is, the performance of the model. This was because the accuracy metrics didn't provide as much information about the quality of questions.

It was also noticed that SQuAD was the best dataset to make use to train the model as it has a huge number of samples comapre to SciQ and MCQL datasets.

Most models used for MCQ generation make use of word prediction models where they produce words that are contextually similar to the answers with respect to the answer's context. This work's novelty lies in using target words and target phrases already present in the given passage to find the most similar target words or phrases to answers which are used as distractors for their respective questions. This is used to introduce an element of confusion into answering the question as one cannot just match the words found in the passage and to identify the correct answer. Many related work also made use of a knowledge base in order to generate distractors for a given question but this method limits the type of questions and passages that can be provided as the input to our model, we therefore made use of predictive models instead as they see how well the generated answer fits the context of an answer and not how right the generated answer is.

IV. Proposed Methodology

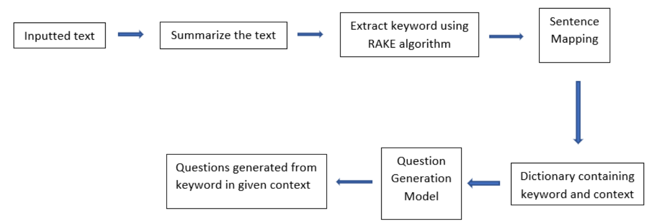

Fig. 1. Question Generation.

The above Fig. 1. shows the flow by which the given body of text is transformed so as to produce question - answer pair. The detiled explain follows below.

The input has to be in the form of a .txt file. This text is read by the model and is summarized by extracting only the most important sentences. The important sentences are determined based on the frequency of words used in them. So, greater the frequency of a particular word in the inputted text file, the more is the importance of the sentence containing that word.
On this summarized text, the Rapid Automatic Keyword Extraction (RAKE) algorithm is applied. This algorithm detects the important phrases in the body of the text by analyzing the frequency of the words and its co-occurrence with remaining words in the body of text. These important phrases are known as target or focus or key words. These are used to identify sentences in the passage that can be used to form a good, useful question. It is also helpful to find suitable distractors for those questions formed.

The given body of text is then tokenized based on location of keywords or focus words. So based on the position, it breaks down into different contexts that can be passed to a model to generate the questions. The keywords are then mapped to the context they are taken from and stored in a dictionary.

For the process of generating questions, the T5 transformer model is used which is a part of Hugging Face library. For training the T5 transformer model this work has made use of the SQuAD dataset what has over 100,000 questions to train on. It has 3 columns namely - context, answer and question. Records that have answers with more than or equal to 7 words are removed in both the train and valid dataset.

The pre-trained t5-base model implemented in Transformer library is used. The T5Tokenizer and T5ForConditionalGeneration are also used in this work. The t5_tokenizer makes use of encoder to find out what layers can be squeezed out. ¡pad¿ is used so as to fill in places where batch sizes are less compared to other required batch sizes. This is then decoded back to original statement. The T5 model is tuned on calculating the loss. It makes use of the AdamW optimizer which decouples the decaying of weight from gradient updation unlike Adam optimizer which faces convergence problems .

Since, the dataset is huge, we make use of GPU to train the model. To test the model, context is passed to model and it can be observed that it generates the questions based on this context.

To this model, the context and answers are passed. This then generates questions. The genrated questions along with their respective answers are then passed on to the next module "Distractor Generation" to generate appropriate multiple choice questions. As seen in Fig. 2, three different ways are used to produce three different distractors. By considering the generated target words, question answer pairs and context answer pairs as the input, distractors are generated using the Sense2vec model, Gensim model and BERT.

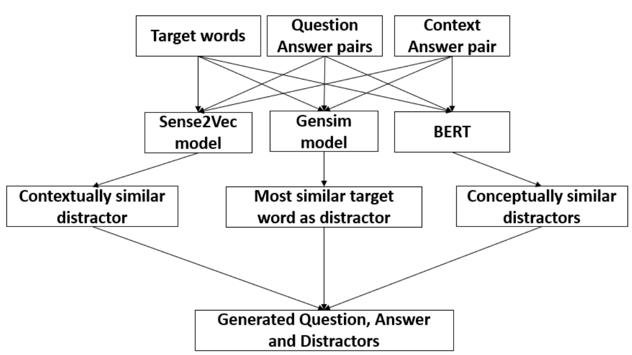

Fig. 2. Distractor generation

Distractors are the incorrect choices given for a multiple-choice question. Choosing the distractors for a given question is an important task as it can make answering the question very easy if the chosen distractor is unrelated to the question or makes answering harder if the distractors are good possible answers to the question. It should also not be the answer to the question in a different form or tense as that leads to having multiple correct options.

Distractor generation is based on the context of the answer and the answer itself for a given question. This model produces 3 distractors for each question. In order to do so, we consider the extracted keywords from the given passage, the context of the answer and the answer itself, and question-answer pairs. Any questions generated where the answer is already part of the question are removed.

The first distractor is generated by making use of the Sense2Vec. Sense2Vec is a neural network model and is used for getting the vector space representations of documents. Sense2Vec is an extension of word2vec but it makes use of contextual information which makes it better than word2vec. It was trained on the Reddit vector dataset which contains all comments on Reddit from the year 2015 to 2019. We pass each extracted keyword from the passage and obtain a list of tuples with words/phrases with the best matching sense based on available frequency counts and sense. The returned word which is not the same as the answer and is not the same as the answer when stemmed and lemmatized is taken as the distractor. In case the answer is a phrase and no best matching sense is returned for the phrase, the answer is split and contextually similar words for each word are joined together to get the distractor. In case no similar word is obtained for a word, the same word is assigned as the distractor. This distractor is later removed from the distractor list for that answer.

The next distractor is generated by considering the most similar keywords or target words in the given passage. This is to introduce more confusion into answering the questions. When contextually similar distractors are generated, even though answering the question might be difficult, it can be done by identifying what phrase had been used in the passage rather than actually answering the question. This can be solved by making one of the distractors a target word from the passage that is the most similar to the answer to the question. This model makes use of the Gensim library to identify the most similar target words from the passage. Gensim's doc2vec is trained on the 'text8' dataset. The vector representation of each answer is then calculated by the trained model. The target word with the closest euclidian distance is taken as one of the distractors for the answer. This allows us to add one more element of confusion while picking out the right answer in the MCQ.

The third distractor is generated by making use of the pre-trained (Bidirectional Encoder Representations from Transformers) BERT model. This is done by taking the context an answer belongs to and masking the answer from that sentence. The BERT model was chosen as , the prediction of the masked word is done by considering the context on either side of the sentence, therefore producing more contextually similar words. The sentence from which the answer is taken from (the answer's context) is considered as input to the BERT model in which the answer is masked by the [MASK] tag. The predicted words are sorted to find the most conextually correct word as the distractor.

Any redundant distractor produced where the same phrase or word is produced but in different tense is removed and any additional distractor produced by the BERT model is used instead as it produces the most contextually correct answers.

V. Experimental Results

Correct Question-Answer-Distractors Generated	Incorrect Question-Answer-Distractors Generated
Q: where did russia invade? 1 . russian 2 . crimea 3 . ukraine **A: ukraine**	**Q: who blasted wimbledon's ban on russian and belarusian players?** 1 . win record 2 . at 3 . world number 4 . taking part **A: world number**
Q: who were banned from taking part in this year's grand slam event? 1 . russian 2 . lithuanian players 3 . belarusian players 4 . Jews **A: belarusian players**	**Q: who was banned from taking part in wimbledon on wednesday?** 1 . ukrainian 2 . soviet 3 . russian 4 . Illogical **A: russian**
Q:What gets trapped in the batter as the cake bakes? 1.gas bubbles 2.Ants 3.baking powder 4.synovial fluid **A:gas bubbles**	**Q:The proteins and starches in the flour turn into a sturdy structure?** 1.flour turn 2.Sauce 3.Mix 4.cornmeal flip **A:flour turn**
Q:What do the proteins and starches in the flour turn into? 1.Fluffy 2.sturdier actual structure 3.sturdy structure 4.Mixture **A:sturdy structure**	**Q:What does gluten form in flour?** 1.Water 2.Protein 3.Flour 4.semi-solid state **A:water**

Generated MCQs

The number of questions generated by this work depends on the size of the inputted text and the number of keywords found out using the RAKE algorithm.

In this work, it takes a considerably large time to train the model but substantially lesser time to apply it on the inputted text.

The generated multiple choice questions by this work, are inspected by human evaluation. This turns out to give almost 90% semantically right questions but about 60% sensibly right question - answer pairs and their respective distractors. It was found that manual checking of produced questions had to be done for the proposed model as no quantifiable automatic method to check the performance of the model could be done.

VI. Conclusion

From the work done, it can be concluded that this model has great scope to generate MCQ's that have a higher level of confusion among the options to choose, compared to other models. Further, scope of this project lies in making more meaningful question answer pairs by identifying better target words. The question generation can also be further improved upon by producing more syntactically correct and meaningful sentences. The distractor generation can also be improved upon by correcting the tense of the produced distactors. Distractors can also be made to be more meaningful as even if the answer is grammatically and semantically correct when seen as the answer to the given question, it can be unrelated to the given passage therefore can become an easy giveaway to the right answer.

REFERENCES

REFERENCES

[1] Ramakrishnan, Ganesh & Li, Yuan-Fang. (2018). A framework for automatic question generation from text using deep reinforcement learning.

[2] 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.

[3] A., J. & D., A. & A., W.. (2020). Intelligent Questions and Summary Generator - Examally. International Journal of Computer Applications. 177. 1-6. 10.5120/ijca2020919663.

[4] 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.

[5] Pritam Kumar Mehta, Prachi Jain, Chetan Makwana and Dr. C M Raut,"Automated MCQ Generator using Natural Language Processing",International Research Journal of Engineering and Technology(IRJET) Volume: 08 Issue: 05 — May 2021

[6] Sheetal Rakangor and Dr. Y. R. Ghodasara,"Literature Review of Automatic Question Generation Systems",International Journal of Scientific and Research Publications, Volume 5, Issue 1, January 2015 ,ISSN 2250-3153

[7] Kaggle.com. 2022. Reddit vectors for sense2vec Spacy. [online] Available at: ¡https://www.kaggle.com/datasets/poonaml/reddit-vectors-for-sense2vec-spacy¿

[8] Analytics India Magazine. 2022. Guide to Sense2vec - Contextually Keyed Word Vectors for NLP. [online] Available at: ¡https://analyticsindiamag.com/guide-to-sense2vec-contextually-keyed-word-vectors-for-nlp/¿.

[9] Medium. 2022. BERT Explained: State of the art language model for NLP. [online] Available at: ¡https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270¿