Automatic Question Generation Using SpaCy

MSc Research Project
Programme Name

Priyanka Sujgure
Student ID: X20136706

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Priyanka Sujgure
Student ID:	X20136706
Programme:	Programme Name
Year:	2021
Module:	MSc Research Project
Supervisor:	Prof. Christian Horn
Submission Due Date:	20/12/2021
Project Title:	Automatic Question Generation Using SpaCy
Word Count:	XXX
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Automatic Question Generation Using SpaCy

Priyanka Sujgure
X20136706

Abstract

The availability of learning resources on online learning platforms throughout the world has increased significantly, with some open-source Massive Open Online Courses (MOOCs), ed-tech businesses, and schools migrating to an online learning-based approach. This environment provides a wealth of possibilities and enables constant access to information, but it hinders the capacity to grasp crucial concepts. Improving the quality of understanding is a major issue that requires long-term solution. One method to enhance the teaching/learning experience it is to ask questions. It stimulates critical thinking and effective learning; however, writing questionnaires for every piece of online information is time-consuming and requires an automated solution. The proposal seeks to address the issues by automating question formation just with textual inputs. It uses SpaCy library of the Natural Language Processing (NLP) to do so. The NLP pipeline of tokenization, Parts of Speech (POS) tagging, etc. has been implemented to form Fill in the Blanks (FIB) and True or False types of questions. The use of pre-trained models such as BERT is been done here. Experiment is performed out on two different datasets to evaluate the results. The experiment performed shows more than 50% good quality questions generated in both the question types .

1 Introduction

The traditional teaching/learning methodologies have undergone massive transformation resulting in an alternative form of "e-Learning" or "Online learning"; however, it comes with challenges of its own. As a result, pupils are having a harder time adapting to this paradigm shift. Prior to the pandemic 11% of the public expenditure was assigned for education (Schleicher; 2020). The online education setup expose students to a lot of information. Knowledge checks are essential in this learning technique. Exam-style questions are a basic teaching tool that may be used for a number of objectives. Questions have the capacity to impact student learning in addition to serving as an assessment tool. Despite these advantages, creating manual questions is a difficult and painful undertaking that needs training, expertise, and resources. This problem is exacerbated by the need to replace assessment questions on a regular basis to ensure their value diminishes or disappears after a few rounds of use (due to being shared among test takers), as well as the rise of e-learning technologies like massive open online courses (MOOCs) and adaptive learning, which necessitate a larger pool of questions (Kurdi et al.; 2020).

Manual question generation from a text for practice exercises, tests, quizzes, and other purposes has been a time-consuming task for academicians and instructors for as long as they can remember, and with the explosion of a large educational material available online, there is a growing need to scale this task. In addition, there has been a growing need in recent years for intelligent tutoring systems that employ computer-assisted instructional material or self-help practice activities to improve learning and objectively assess a learner's aptitude and successes. The proposed question-creation strategy would help students improve their capacity to self-analyze and learning experience. To improve the overall effectiveness in the teaching learning process and take advantage of the ever changing learning model.

1.1 Example of a Factual Question Generation

This section provides an illustrative example about how an automatic question generation takes place provided with the current natural language processing (NLP) technologies.

At the very outset to understand the logic behind how the question generation works a paper-model was worked out. Below is an excerpt which is considered to help understand the logic behind the questions. These paragraphs are from the famous book ”The Alchemist” by Paulo Coelho.

> *”It sounded like a Gypsy prayer. The boy had already had experience on the road with Gypsies; they also traveled, but they had no flocks of sheep. People said that Gypsies spent their lives tricking others. It was also said that they had a pact with the devil, and that they kidnapped children and, taking them away to their mysterious camps, made them their slaves. As a child, the boy had always been frightened to death that he would be captured by Gypsies, and this childhood fear returned when the old woman took his hands in hers.”*

Below are the questions that can be generated:

- It sounded like a ________ prayer.
- Gypsies had a pact with the devil and they ________ children.
- Gypsies spent their lives ______ others.
- The Gypsies travelled with flocks of sheep. True or false
- Gypsies saved the children from the devils. True of False

From the paper-model experiment it was clear that to generate a basic question like FIB, the idea is to look for the core meaning within the sentence that it is trying to convey. Once you find it you have to replace that specific word with a dash. In case of the true or false the logic isn't as simple as the gap-fill. It can be done by many methods such as negating the sentence, changing a few named entities to something different, changing the main verb,etc. From the above example it is clear that splitting of the sentences into verb and noun phrase and changing the nouns do help to make a good true or false type of question.

1.2 Research Question

How can Natural Language Processing (NLP) approaches be applied to textual data to automatically produce questions and evaluate it's performance of different question types on different data?
The aim of thesis is to evaluate the performance of question generator system developed along with different types of questions with different data.
Further sections in the paper are organized as follows: Section 2 describes the Related Work, Section 3 illustrates the Methodology followed in this research, Section 4 explains the Design Specification, Section 5 illustrate the Implementation aspects of research, Section 6 examines the evaluation results, Section 7 showcases Conclusion and Future Work.

2 Related Work

Prior work in the topic of automated question generator (AQG) has covered methodologies and techniques used, types of questions created. The below section gives an overview on selection of different approaches from the past material within the explicitly mentioned subheadings.

2.1 Technologies used in automatic question generation

The study (Khullar et al.; 2018) describes a system that creates several natural language queries automatically from complex statements. The focus is on producing wh type of questions. It uses spaCy's parser to understand the relationship among the adverbs and the pronouns. It looks for well-defined linguistic elements in the phrase, such as the root and relative clause verb's tense and aspect type, and the head-modifier link between distinct clauses to communicate appropriate sections of the sentence, etc. the data that goes into the question generator. The evaluation was done 4 human evaluator's as there is no standard system to evaluate a question generator (QG) system. Methods used in this study outperformed Heilman's study on the basis of the grammatical correctness, semantic adequacy, fluency, and uniqueness. The limitation of the study is that it can generate questions only from the sentences which have atleast one relative clause.

Study (Nwafor et al.; 2021) shows the implementation of multiple choice questions generation through Natural Language Processing (NLP) tools. In this research the techniques used with NLP are Term Frequency-Inverse Document Frequency (TF-IDF) used to convert text data into vectors and N-gram for word breaking and summarization. There is no mention of what kind of library is used for this research. Evaluation was performed on five different educational materials to test the system's effectiveness and efficiency to ensure that it is not perverse. The teacher-extracted keywords were compared to the system-generated keywords, demonstrating the system's ability to extract keywords from educational materials. This study isn't about whether multiple-choice tests are superior than other sorts of assessments but to employ NLP to construct questions.

(Mhatre et al.; 2019) in this research, a method for question creation that uses coreNLP and tree regular expressions to manipulate parse trees is developed. They have utilized spaCy for named entity recognition to identify entities for question word identification. From an input text document, a QG system that can create appositive and subject-object questions is proposed. The aim for the research is to generate factual questions using parse tree manipulation as well as named entity recognition. Now, a very good utilization of spaCy has been done in this research to generate factual questions. The aim of generating multiple questions from a single input sentence is met but the drawback is they fail to show its results on a larger dataset.

(Panchal et al.; 2021) study sets very high standards for the pre-processing for data used for question generation. It generates fill in the blanks, multiple choice and rule based wh-type questions. The steps carried out for the pre-processing are NLP based namely tokenization, named entity recognition (NER), parts of speech (POS) tagging. The evaluation is one wherein the user can evaluate the quality of the question on the basis of answerability, grammatical correctness of the question. However, upon analyzing the findings, it is clear that the Fill in the Blank questions only have a 59% success rate, while the Wh questions only have a 49% success rate. Limitations is of getting wrong answers as the data is trained on SQUAD dataset.

(Devlin et al.; 2018) shows how does the BERT, i.e, Bidirectional Encoder Representations from Transformers can be used in the question and language interference. Two methods to apply the pre-trained models are feature-based and fine-tuning. Feature-based approach uses architectures specifically required for the tasks whereas the fine-tuning such as Generative Pre-trained model (OpenAI-GPT) has less task-specific parameters. It is trained just by tuning the parameters. BERT is used in two phases in the framework namley pre-training and fine-tuning. Evaluation using BERT is performed on 11 different tasks. Main contribution from this research is that they have extended the findings of the experiments to deep bidirectional architectures, enabling the same pre-trained model to successfully solve a wide range of NLP tasks.

(Kriangchaivech and Wangperawong; 2019) uses Recurrent Neural Network(RNN) to develop the question generator. The datset used is of Wikipedia articles. The training is done with the help of Stanford Question Answering Dataset (SQUAD) dataset consisting of 100,000+ questions. Model was able to generate questions consisting of 8 words per question. To compare the similarity of SQuAD questions with model-generated questions, the word error rate (WER) was utilized as a measure. Drawback of the study is that the results are lower compared to the results from the pre-trained models.

(Azevedo et al.; 2020) Question Answering(QA) and Question Generation(QG) have been the topic of extensive research. Study merges these two subjects focusing on using QG to improve QA outcomes. A program using existing Natural Language Processing (NLP) approaches that address these two topics separately have been created. The QG methodology analyzes a sentence's content using POS tagging and Named Entity Recognition (NER). To ensure loose coupling with the QA task, they employ Information Retrieval to rank sentences that may contain relevant information about a query, and Open Information Retrieval to assess the sentences.

(Indurthi et al.; 2017) investigates the topic of automatically creating question-answer pairs from a knowledge graph. Knowledge graphs are one that holds information regarding lot of entities and the relationship between them. The model treats subset of keywords as a sequence, and a sequence to sequence model based on RNN is proposed to produce a natural language query from it. F1 score is used as an evaluation metric. An extrinsic assessment by utilizing the produced QA pairs to train a QA system, and we find that employing automatically generated QA pairs in addition to manually created QA pairs increases the QA system's F1-score by 5.5 percent (relative).

In (Das and Elikkottil; 2010) study, combination of a Q/A system with a summarizer is done, claiming that using an automated summarizer with a Q/A system will improve the Q/A system's efficiency. This study investigates ways to create an extractive summarizer that selects the most essential phrases that might be a possible answer to a given query. Combining a summarizer with the Q/A system might have proved beneficial for a smaller data but it would fail if a descriptive answer were required. The summarizer helps in locating the main idea but restricts the number of questions that can be generated. Addition of the summarizer increases the efficiency and effectiveness.

(Heilman and Smith; 2010) Automated question generation from reading materials for educational practice and evaluation is addressed in the study. They over generate questions and rank them. They utilize manually defined rules to change declarative statements into questions. Those questions are then ranked by a logistic regression model trained on a small, specialized dataset comprising of labeled output from their system. Ranking substantially doubles the percentage of questions deemed acceptable by annotators, from 27% of all questions to 52% of the top 20% of questions.

(Blšták and Rozinajová; 2021) Machine learning techniques are used in this research. The SQuAD (Stanford Question Answering Dataset) dataset is used. It classified new phrases using multi-label classification and established a distance measure based on the similarity of their composite patterns. Following that, individual questions are developed for each of the classes. The QGSTEC dataset was used to analyze the study's findings. The research would have been better served by using pre-trained models.

Conclusion: After studying the various technologies used for question generation and data pre-processing the idea of using spaCy is considered for this research as spaCy has advanced features and supports a lots of functionalities. For pre-processing steps such as tokenization, POS tagging, etc. will be performed. As the data is huge and has a lot of noise such techniques are required to implemented. Here, after analyzing the performance of pre-trained model with a machine learning model the decision is to use the pre-trained one's they are easier to tune and better accuracy can be achieved.

2.2 Types of output generated in automatic question generation

(Cruz et al.; 2021) focuses on generating methods , for example, automatically produces multiple-choice, true-or-false, and fill-in-the-blank questions from any text. It implements the Automatic Distractors Generator algorithm to provides a number of incorrect but pertinent responses for multiple choice questions. The research's overall performance for generating accurate answers for fill-in-the-blank questions was 70%, which is excellent, but just 16% for multiple choice questions, which is discouraging.

(Agarwal et al.; 2011) shows generation of wh type of questions using the discourse connectives. They play an important role in understanding the text coherent. On basis of what sense the discourse connective creates the question is generated. For example, 'because' is a casual sense so 'why' is the type of question to be generated. Although the evaluation and other techniques are mentioned the model deployment is not mentioned in the paper.

(Agarwal and Mannem; 2011) shows how to create gap-fill questions for material in a document using an automated question creation system.A conventional biology textbook was used to test the system. It selects the questions which has enough data that has sufficient data to predict the key when it is blanked out along with generating distractors. There are two evaluators to evaluate the quality of the questions generated. They provided a score of 91.66% and 79.16% as good quality of questions.

(Murugan and Ramakrishnan; 2021) displays the development of distractors for Tamil, a comparable Indian language to Hindi. It also focuses on fill in the blank type of question. The study is remarkable in that it shows that automatic question generation can be done not just in English but also in other languages, benefiting students in learning both their native language and other languages of interest. POS tag, difficulty level, spelling similarity, semantic similarity, and word co-occurrence approach were used to create distractors.

(Chinkina et al.; 2020) demonstrates the automatic production of text-based and wh-type questions from a given text. The goal of the research was to evaluate the perceived quality of machine produced and human-written questions. It majorly focuses on the grammatical parts of the questions. Whether the question uses the correct grammar or not. It deploys a template-based approach which leads to increased computational time.

(Zerr; 2014) uses a template based approach for question generation based on Part of Speech pattern templates. The idea is to break down the sentences and form them in a way that it matches with the pre-build templates of POS pattern. This is achieved simply by shuffling the places of the phrase. Dataset considered is pretty simple. It uses stories data from kids books. The evaluators scored the system with average 58.36%. Limitation is that this experiment is performed on a very small and a pretty simple database.

(Kurdi et al.; 2020) is a review of the researches done in the field of Automatic Question Generation. A survey of 93 papers from the year 2015 to early 2019 is included in the study. Highest type of studies being done for assessments while only 1 study was for active learning. 17 out of the 93 studies were about generating fill in the blanks and only 1 to generate true or false type of question. Drawback of this study is that it fails to include research's done in languages other than English.

(Curto et al.; 2012) In the course of this study, the program "THE-MENTOR" was developed. Its goal is to automatically generate multiple choice questions from the supplied text to let visitors start a conversation about their vacations by utilizing question/answer and quizzes. The lexico syntactic patterns are learned by "THE-MENTOR." It understands and evaluates patterns using a pattern learning algorithm, categorizing them as strong, weak, or inflected. Only 24% to 30% of the questions generated are answered correctly. This research isn't flawless; it just creates one form of questions as an output.

Conclusion : From the above research it is clear that maximum research's use the fill in the blank type of question. Also, true or false kind of question is not something that many of the research's haven't focused upon. So, in this research the fill in the blanks and true or false kind of questions are selected to be implemented.

3 Methodology

Research uses the CRISP-DM methodology where the business knowledge and data knowledge form an equal importance. Here the focus is on the educational sector. The online learning environment is hampering the understanding capacity to grasp crucial concepts by the learner.

3.1 Data Understanding

The dataset used is of "wikibooks" which is freely available to access on internet. It consists of 270,000 chapters of wikibooks in 12 different languages namely English, French, German, Polish, Italian etc. This was selected as it had lots of options for data to select from. English language text is selected that has 86736 rows and 5 columns. The dataset is in a .sqlite file which is accessed through the DB Browser software. Columns in the database are title, url, abstract, body_text, body_html. The "title" column consists of all the name of the books, "url" consists the link to the html page of each books content, "abstract" is the short summary about the content, "body_text" as the name suggests is the textual content and the last column is "body_html" consists of the html version of the contents of the page.

Sr.no	Name of the column	Content	Data Type
1	Title	Title of the chapter	Text
2	url	Link for the content	Text
3	abstract	Short summary of content	Text
4	body_text	Content of the chapter	Text
5	body_html	Html version of the data	Text

3.1.1 Dataset challenges

1. Dataset is available in sqlite and not csv format.
2. Consists of a lot of non-sensical data.
3. Much of the text in the data do not follow a standard structure. Due to this there are a lot of cleaning required to be performed.

3.1.2 Ethical concerns

The data received from the source is made public, with permission to use it for personal and commercial reasons granted. The use of the dataset does not conflict with the use. Data sources are specified.

3.1.3 Exploratory Data Analysis

While taking a look at the actual content of the "body_text" there seems to be a lot of abandoned or nonsensical texts. Below is a histogram in Figure 3.1.3 of the length of the text fields(note that its a logarithmic value). This is done with the help of R software.

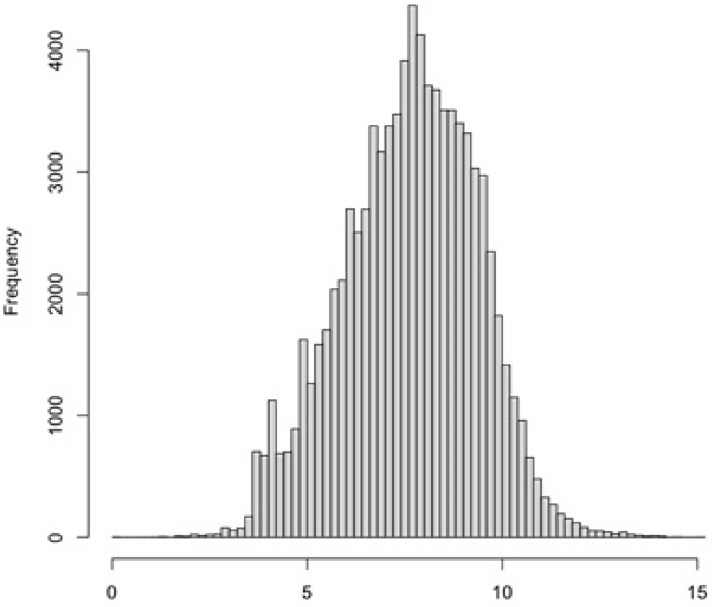


Figure 1: Histogram of the text data(log)

From the histogram, log(x)=7 means x is about 1,000 and log(x)=11 means x is about 60,000. It is useful to consider the data that has certain length(example 2000 to 50,000 characters). The lower bound is the important one to distinguish between something that is more a Wiki entry to a book. It can go higher as well and the upper bound is only for saving on the computing time. This is the reason a lower bound of 10,000 is been selected to select the data in the initial phase.

3.2 Data Preparation and Cleaning

Textual type of data consists a lot of noise. It is critical to effectively process this data in order extract meaning out of it. There can be a case where a lot of data that is irrelevant and does not make any sense. As wikibooks is an open source medium and everyone can contribute to the wikipages, there is very high probability of having poor data.

Two types of textual data pre-processing is done. First being the natural language processing (NLP) specific and second one specific to the Question Generation (QG). To make it easier for accessing the dataset the files are uploaded to a cloud storage service called "Mega". Out of the 12 languages available in the dataset "English" language data is chosen. As the data is huge, it takes a long time to process, which is why it is sampled. To implement the question generator, a sample of the first 100 records is selected.

1. Data Preparation: The dataset has 86736 rows and 5 columns, not all of which are helpful. Before going ahead with any of the critical pre-processing tasks initial

level analysis like null value check and summary of the data is done as the very first steps of data cleaning. The columns named abstract and body_html are not required because the abstract column has a very short summary of the chapter and the body_html has the html version of the actual content. As the data is already present in the body_text column, and the question generation is mostly based on the complete text, the wiki page html code is also ineffective. So, both the columns are dropped. Subsequently the data is analyzed to see whether there are any NA values and/or duplicate values.

2. Data Cleaning : Here, as observed after the basic cleaning steps there is a lot of non-sensical data present which needs to treated to make the data usable or else it will be just garbage data (i.e, data does not make any logical sense as there are missing texts and symbols) generating output which will have no significance. The raw data might be pre-processed in two steps. The first is a transformation that is particular to the raw data, while the second is a transformation that is specific to the question types as two different types of output questions are to be generated i.e, fill in the blanks and true or false. The following are the cleaning steps that must be followed in order to process the data in both of the previously stated stages:

    - Only having letters, numbers and punctuation's like full-stop, comma, hyphen,etc.in the sentence and eliminating everything else.
    - Deciding on the length of the paragraphs.
    - Deciding on the length of the output for the questions generated.
    - Further cleaning to remove garbage data.
    - Removal of punctuation marks like single quote, double quote and question mark.
    - Splitting sentences at verb phrase or noun phrase.

3.3 Logic Implementation or model implementation

After the research, it was clear to implement logic for the two different types of question generation namely fill in the blank and true or false. SpaCy library is majorly planned to be utilized to perform all the Natural Language Processing (NLP) steps. SpaCy (Khullar et al.; 2018) has better features compared to the NLTK library. So, the SpaCy will be used. From (Devlin et al.; 2018) it is clear that the pre-trained models are easier to be tune to generate good output. It's performance can be improved by fine-tuning the BERT parameters. The details for the implementation are discussed in the implementation section.

3.4 Evaluation

In this project textual data is used. Because this is textual material, human judgement is not always accurate. The aim is to analyze how many questions can be correctly created grammatically and logically using the proposed question generator system with various sorts of input data. A detailed discussion on its implementation is in the evaluation section.

4 Design Specification

The framework of the research consists of four major steps namely the Data understanding, Data Pre-processing/cleaning, Logic and model implementation, and lastly Evaluation as shown in Figure 2.

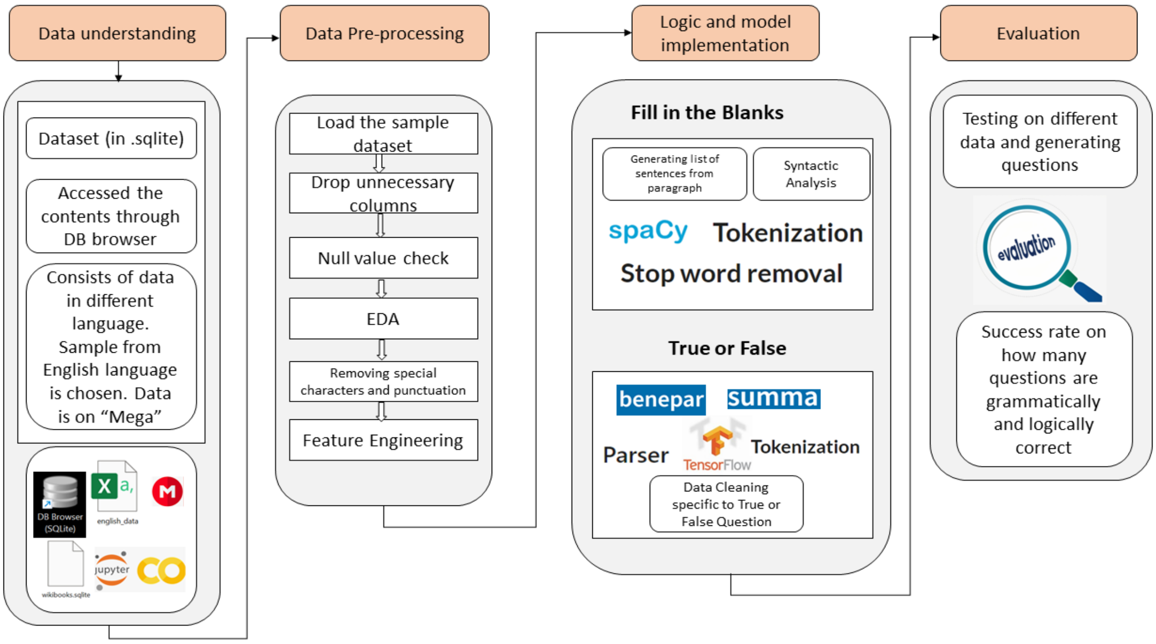

Data understanding
Dataset (in .sqlite)
Accessed the contents through DB browser
Consists of data in different language. Sample from English language is chosen. Data is on "Mega"
DB Browser (SQLite)
english_data
wikibooks.sqlite
jupyter
CO
Data Pre-processing
Load the sample dataset
Drop unnecessary columns
Null value check
EDA
Removing special characters and punctuation
Feature Engineering
Logic and model implementation
Fill in the Blanks
Generating list of sentences from paragraph
Syntactic Analysis
spaCy
Tokenization
Stop word removal
True or False
benepar
summa
Parser
TensorFlow
Tokenization
Data Cleaning specific to True or False Question
Evaluation
Testing on different data and generating questions
evaluation
Success rate on how many questions are grammatically and logically correct

Figure 2: Design flow

1. Data Understanding : Once the data is been exported from the .sqlite format to the .csv format using "DB Browser" software it is uploaded to "mega" and then loaded onto "Google Collab" to be perform cleaning and other actions. "Mega" is used only to ease the access towards the dataset. As it is stored onto cloud manual transfer of the dataset is not necessary.

2. Data Pre-processing : This is an important step in the projects as it involves Data cleaning, EDA, feature engineering. Steps like null value check, duplicate value removal, dropping NA values, dropping unnecessary data columns. As a part of EDA, a basic understanding of what the data actually contains is taken, wordcloud is also generated to understand most used words in a paragraph. Feature engineering involves cleaning the data to as much possible as clean and useful format which involves removal of unnecessary symbols, removing punctuation's, etc.

   Let's have a look at how they were accomplished. The elimination of unnecessary symbols, paragraph length, and further cleaning of junk data are all part of the initial step of data cleaning in order to turn raw data into a useful format.

   Firstly, the column of "body_text" is been selected as the actual text data is under that column. Next, except for the alphabets (upper and lower cases), numbers (0 to 9) and punctuation marks like comma, full-stop and hyphen are kept rest all the data is been deleted. This is appended into a list called "new_para" with the help of a regex function.

Second item mentioned is to decide the length of the paragraph. The reason behind doing this is to choose chapters worth some length. If they are too short it will not generate enough questions whereas, if the data is very large it will add up to the computational time. The data is been trimmed to 10,000 words. This limit is decided randomly and has no statics behind (as explained earlier).

Third step is specific to the fill in the blank type of question. To create a fill in the blank one needs to have a sentence, main idea of the paragraph and not an entire paragraph so, the paragraphs are broken down into separate sentences. This is accomplished using spacy's "sentencizer," a component of the pipeline for detecting sentence boundaries using rules. All these sentences are stored in the list. After which the sentences who has a length of more than 50 characters and less than 200 characters are only selected in an array. This is done as too long or too short sentences will not help.

The further cleaning is necessary as there are still few sentences which had garbage data for example *"1969.09.03 Numbers SC 3711-3715 Liapine 3746-3750 Scott 3634-3638 Michel 3661-3665 SG 3723-3727 Yvert 3522-3526"*. This sentence is does not make any sense as one can see. There are two ways to remove such data. First being to delete the entire chapter and second to apply some logic to remove only the sentences. The later is been performed to clean the data. One of the observation is that this data is inside square brackets. So, a regex function stating that if the data is in square brackets and has single characters with few symbols shall be removed. The regex is used to target any data that is inside square brackets and have random characters, numbers, symbols or a combination of all.

Next step performed for the cleaning of the data is specific to the true or false question generation. With the summa extractive summarizer library (BERT), summarization of the loaded data is done. Remove any sentences with single quotes, double quotes, or question marks from the summary sentences since they aren't appropriate for a true or false kind of questions.

For splitting the sentence at an appropriate phrase the parser is used in understanding where a word can be changed to make a statement false. The use of "Berkley Constituency parser" to split the sentence is done. This is explained in detail in the implementation section ahead. Now we have data that is ready to generate the questions with a little bit of logic implementation.

3. Logic and model implementation : The logic or model for both types of questions use spaCy, pre-trained models from BERT, summarizers, summa parser, tokenzation, tensorflow, benepar, OpenAI-GPT2, scipy and few other libraries. The tokenization to form word is done by removing the punctuation's, stop words and numbers. Pre-trained models are used here to get the outputs.
4. Evaluation : As it is a textual data it is to be tested on different subjects data and its performance is to be evaluated which is done in the further sections.

5 Implementation

The processes required for implementing the design logic and methodology utilized for output creation using the technology tool are detailed in this section. Setting up the environment "Jupyter Notebook" is often a first choice; however, the data size is significantly large and it needs to be processed as close as real-time. In order to achieve such high speed processing and turn around time the tool of chioce was chosen to be Google Collaboratory from Google Cloud Platform GCP. As previoulsy discussed, the dataset file is stored on cloud as to easily access the data (it can be eliminated and path . Python programming language is chosen for coding.

Initially, the idea was to perform the research using the NLTK library, however after more investigation/validation, it was discovered that the spaCy comprises advanced features which can be utilised within the project. SpaCy is object oriented while NLTK uses strings as input and output. SpaCy's performance is better than NLTK library(Khullar et al.; 2018) . The research can be majorly classified into two broad categories as "Fill in the Blanks" (Cruz et al.; 2021) and "True or False" question generation . The data to be used shall be textual and should not consist of any images or non-textual data; however, the data used in this case do consists a lot of non-textual data for which EDA (using R) and cleaning is performed (discussed ahead in this section). The programming language is python. Now, let see how the implementation is done.

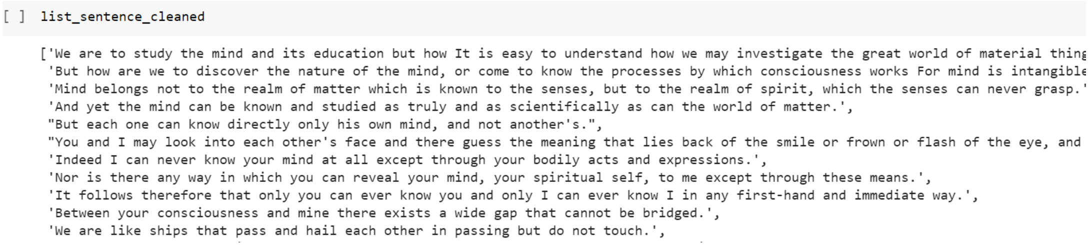
A detailed description about what libraries are required for the implementation of the project are a part of the configuration manual. To setup the working environment firstly, data is been collected and loaded into the Google collab where it is been processed using python. The data cleaning is divided into two basic types i.e, the normal data cleaning to drop unnecessary columns, removing duplicates mentioned and another being specific to the question generation data. As mentioned in the methodology section below are the details on the cleaning steps taken.

The further implementation can be classified into two broad categories i.e, the fill in the blank question generation and true or false generation. Let's see the details of implementing these two questions ahead.

5.1 Fill in the Blank Implementation

To generate the fill in the blank type of question the spaCy library is installed. Once the data has been cleansed as discussed in the earlier sections next steps could be executed. A general pipeline in case of NLP is spell check, Tokenization, POS tagging and NER tagging. The spell check is performed so as to have good communication, literacy and engagement. The reason why tokenization is done is to divide the sentences or strings into words or sub-strings. Tokenization is the process of breaking down a phrase, sentence, paragraph, or even an entire text document into smaller components like individual words or phrases. Next, Parts of Speech(POS) tagging is used to break paragraphs into sentences. It is a software that parses text and assigns tags to different parts of speech example nouns, verbs, and other parts of speech. Doing this explains how a word is used in the sentence. Sentences are broken down into tokens they are then sent for POS tagging.

While executing the first few tasks it was observed that there were a few words whose spelling was considered wrong. So, spaCy's "conceptualspellcheck" was deployed. That's when it come to the attention that the version of spaCy that was used was an older one because the "conceptualspellcheck" works with the version 3.0 and not version 2.0. Due to this there were changes made throughout the code to suit the updated version. Loading the clean data for further implementation. Now, the data available is in the form of paragraphs but to generate a question individual sentences are required and not the entire paragraph. As the data is stored in the form of paragraph it is difficult for the system to analyze where does the string start or end. For this, spacy's "sentencizer" is used. Output is a list of sentences as seen the below figure. It breakdowns the long paragraphs into smaller sentences making it easier to generate the tokens.



```
[ ]  list_sentence_cleaned

     ['We are to study the mind and its education but how It is easy to understand how we may investigate the great world of material thing
      'But how are we to discover the nature of the mind, or come to know the processes by which consciousness works For mind is intangible
      'Mind belongs not to the realm of matter which is known to the senses, but to the realm of spirit, which the senses can never grasp.'
      'And yet the mind can be known and studied as truly and as scientifically as can the world of matter.',
      "But each one can know directly only his own mind, and not another's.",
      "You and I may look into each other's face and there guess the meaning that lies back of the smile or frown or flash of the eye, and
      'Indeed I can never know your mind at all except through your bodily acts and expressions.',
      'Nor is there any way in which you can reveal your mind, your spiritual self, to me except through these means.',
      'It follows therefore that only you can ever know you and only I can ever know I in any first-hand and immediate way.',
      'Between your consciousness and mine there exists a wide gap that cannot be bridged.',
      'We are like ships that pass and hail each other in passing but do not touch.',
```

Figure 3: Output from Sentencizer - List of Sentences

Then, the list of sentences are broken into tokens. These tokenized words are stored in a word dictionary that is prepared. While creating a word dictionary all the punctuation's and numbers have been ignored because while generating a fill in the blank type question it is not sensible to ask for a punctuation mark. Stop words such as is, an, the, and an are not included in the list since the goal is to create factual questions, not to assess English grammar questions. The logic is straightforward now. Once the pivotal answer from the sentence is chosen with the help spaCy's "secret" library from the word dictionary. The sentence is parsed for the pivotal answer and replaced with a blank space. This experiment is done on both a single sentence as well as for the entire dataset at once.

5.2 True or False logic

A few logical steps to create a true or false type of question is by adding or removing negation, changing a named entity, changing the adjective or main verb, splitting complex sentences into simpler or short one's, and changing verb phrase or noun phrase. In this research, last type that is changing the noun phrase or verb phrase is deployed. Library installation is done along with the fill in the blank question and the data cleaned in the fill in the blank generation is considered for generation of true or false.
The same data is loaded into an array(here is named "text"). This data is summarized using the "summa extractive summarizer" library. This library accomplishes by first embedding the sentences, then using a clustering technique to discover the sentences closest to the cluster's centroids. For true or false single quotes, double quotes or question marks have been removed as they are not suitable as assertive statements are necessary in here and not interrogative type.

As the plan is to change the noun or verb phrase the sentences has to be split at an appropriate place. This is done with the help of ”Berkley Constituency parser”(Khullar et al.; 2018). It has pre-trained models and uses segmentation and tokenization from spaCy. The ”Parser” used, simply encodes a text sequence using a stacked encoder. Below figure Figure 4 shows the parse tree for a sentence from the dataset using the ”Parser”.

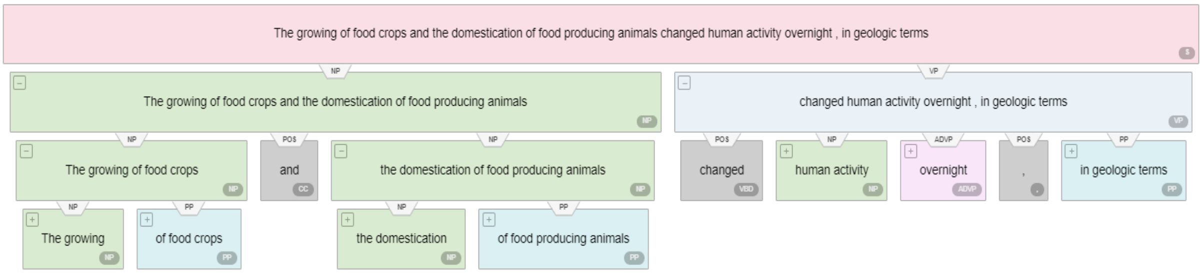

The growing of food crops and the domestication of food producing animals changed human activity overnight , in geologic terms
S
NP
VP
The growing of food crops and the domestication of food producing animals
NP
changed human activity overnight , in geologic terms
VP
NP
POS
NP
POS
NP
ADVP
POS
PP
The growing of food crops
NP
and
CC
the domestication of food producing animals
NP
changed
VBD
human activity
NP
overnight
ADVP
,
.
in geologic terms
PP
NP
PP
NP
PP
The growing
NP
of food crops
PP
the domestication
NP
of food producing animals
PP

Figure 4: Parse tree for a sentence from the dataset

It is initially split into noun phrases and verb phrases. As can be seen in the illustration, this may now be further parsed. So, we consider a sentence as *”The growing of food crops and domestication of food producing animals”*. The noun phrase is then fed into OpenAI GPT-2, which generates an alternative ending verb or noun phrase in order to build a different sentence from the original.

The Generative Pre-trained Transformer-2 (GPT-2) is an open-source AI (artificial intelligence) that translates text, summaries passages, and provides text output on a level that is sometimes indistinguishable from that of humans. Multiple sentences are created with the OpenAI GPT-2 after the models are in place and loaded, and they are filtered from the similar ones (with BERT). Since we only want to maintain the sentences that are distinct to the original ones they are given as an output (considering them as false statements).

6 Evaluation

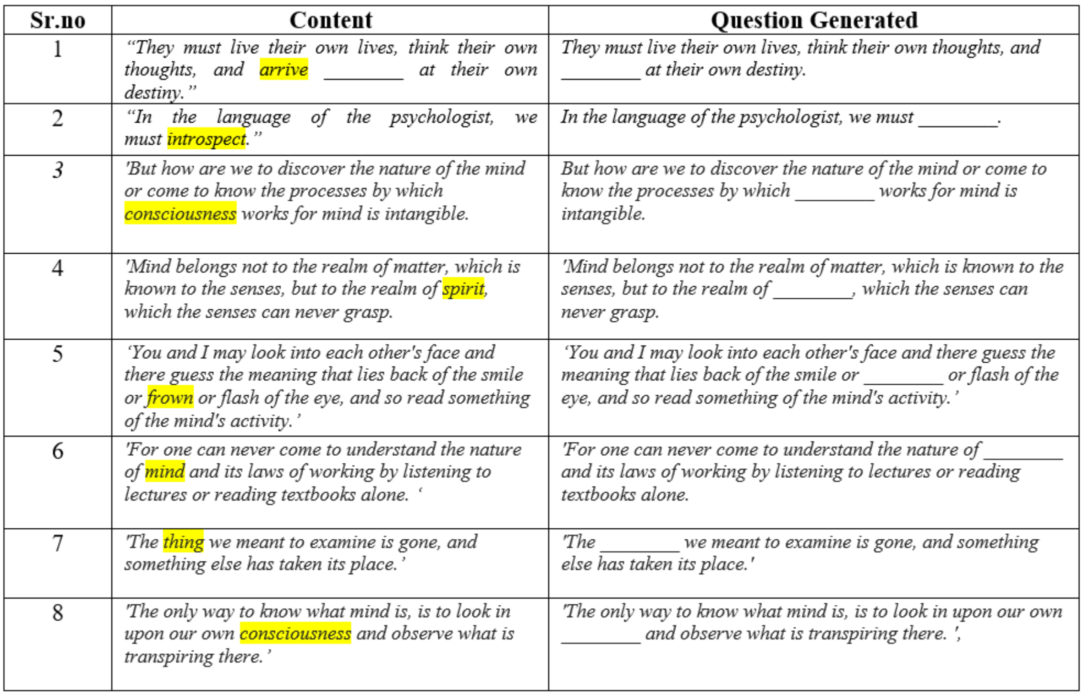
In this section, evaluation, data from different sources are considered to test the performance of the question generator system. The outputs and results are mentioned in this section. The type of evaluation used is intrinsic evaluation method. Human evaluation and automatic evaluation are the two types of intrinsic evaluation methodologies (Amidei et al.; 2018).

To begin setting up the environment, various data is necessary for performing assessments. The dataset used as a base for the question generator is very complex and has multiple challenges. Maximum challenges are to clean the data which have been taken care of. So, to perform evaluation different data which has less garbage data needs to be considered. Here, two data from the ”Project Gutenberg” (a library for ebooks) is considered. The very first experiment is carried out on a literature book and second is on a geography book. The reason behind choosing this subjects was to evaluate if they do generate questions which can be helpful for educational systems and as it is available freely it does not involve any ethics considerations.

6.1 Experiment 1 : Experiment on book by G.H. Betts ”The Mind and Its Education” [1]

For carrying out the first experiment, data from a book named ”The Mind and Its Education” by G.H. Betts is chosen. This book is available on internet so is used as a data source for performing evaluation of the question generator system.

The tabular image shows some of the output questions generated by the question generator system developed in the study. Most of the questions seem to grammatically and logically correct. The left side is the data from the book and the right side shows the output questions generated. Not all of the questions generated are good quality but most of them can be said to have a good quality. The total number of questions generated is a little high they cannot be all shown here. So, only a few examples are shown in the tabular image. In total there are 81 fill in the blanks and 58 true or false type of questions generated from 11 paragraphs fed from the book as an input.



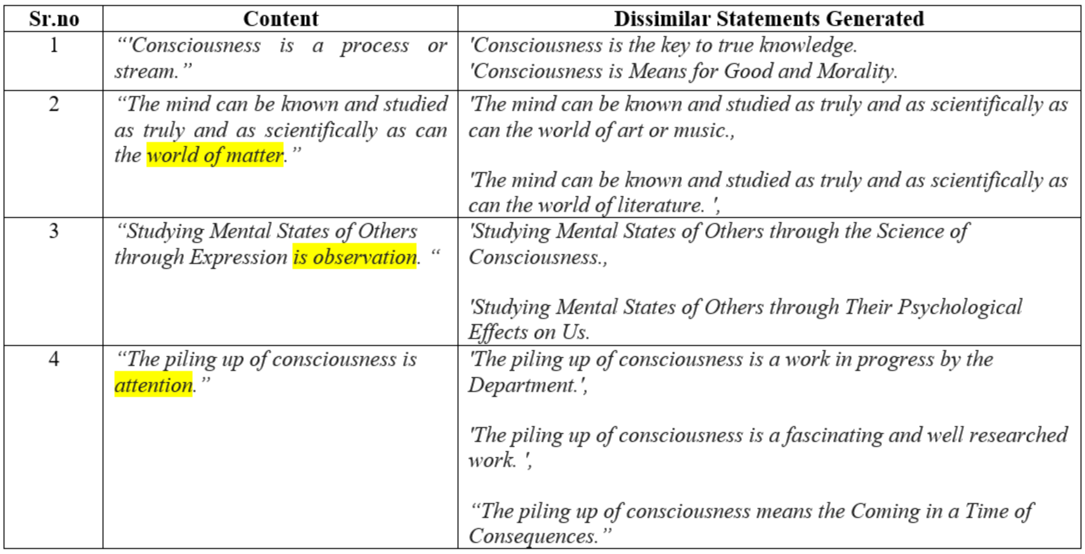
Sr.no	Content	Question Generated
1	*“They must live their own lives, think their own thoughts, and arrive ________ at their own destiny.”*	*They must live their own lives, think their own thoughts, and ________ at their own destiny.*
2	*“In the language of the psychologist, we must introspect.”*	*In the language of the psychologist, we must ________.*
3	*'But how are we to discover the nature of the mind or come to know the processes by which consciousness works for mind is intangible.*	*But how are we to discover the nature of the mind or come to know the processes by which ________ works for mind is intangible.*
4	*'Mind belongs not to the realm of matter, which is known to the senses, but to the realm of spirit, which the senses can never grasp.*	*'Mind belongs not to the realm of matter, which is known to the senses, but to the realm of ________, which the senses can never grasp.*
5	*‘You and I may look into each other's face and there guess the meaning that lies back of the smile or frown or flash of the eye, and so read something of the mind's activity.’*	*‘You and I may look into each other's face and there guess the meaning that lies back of the smile or ________ or flash of the eye, and so read something of the mind's activity.’*
6	*'For one can never come to understand the nature of mind and its laws of working by listening to lectures or reading textbooks alone. ‘*	*'For one can never come to understand the nature of ________ and its laws of working by listening to lectures or reading textbooks alone.*
7	*'The thing we meant to examine is gone, and something else has taken its place.’*	*'The ________ we meant to examine is gone, and something else has taken its place.'*
8	*'The only way to know what mind is, is to look in upon our own consciousness and observe what is transpiring there.’*	*'The only way to know what mind is, is to look in upon our own ________ and observe what is transpiring there. ',*

Figure 5: Fill in the blanks output

Next, is the second type of questions generated. In the true of false generator the system splits the sentences and then a new word or phrase is been added to change the sentences from its original form. These sentences which are dissimilar from the one’s in the original text are considered as false statements. The true or false generator produces good quality output for small or short sentences but for few it changes the meaning to an extent that the main idea behind the sentence gets disturbed. But, as observed from the second example in the table below it actually generated a dissimilar sentence which

[1] https://www.gutenberg.org/cache/epub/17299/pg17299-images.html

is treated as false in this case as a correct one. This makes it difficult to say if that statement really is a false one. For a single sentence multiple sentences with different endings are been generated. These dissimilar sentences are nothing but one's treated as a false statements.



Sr.no	Content	Dissimilar Statements Generated
1	“'Consciousness is a process or stream.”	'Consciousness is the key to true knowledge. 'Consciousness is Means for Good and Morality.
2	“The mind can be known and studied as truly and as scientifically as can the world of matter.”	'The mind can be known and studied as truly and as scientifically as can the world of art or music., 'The mind can be known and studied as truly and as scientifically as can the world of literature. ',
3	“Studying Mental States of Others through Expression is observation. “	'Studying Mental States of Others through the Science of Consciousness., 'Studying Mental States of Others through Their Psychological Effects on Us.
4	“The piling up of consciousness is attention.”	'The piling up of consciousness is a work in progress by the Department.', 'The piling up of consciousness is a fascinating and well researched work. ', “The piling up of consciousness means the Coming in a Time of Consequences.”

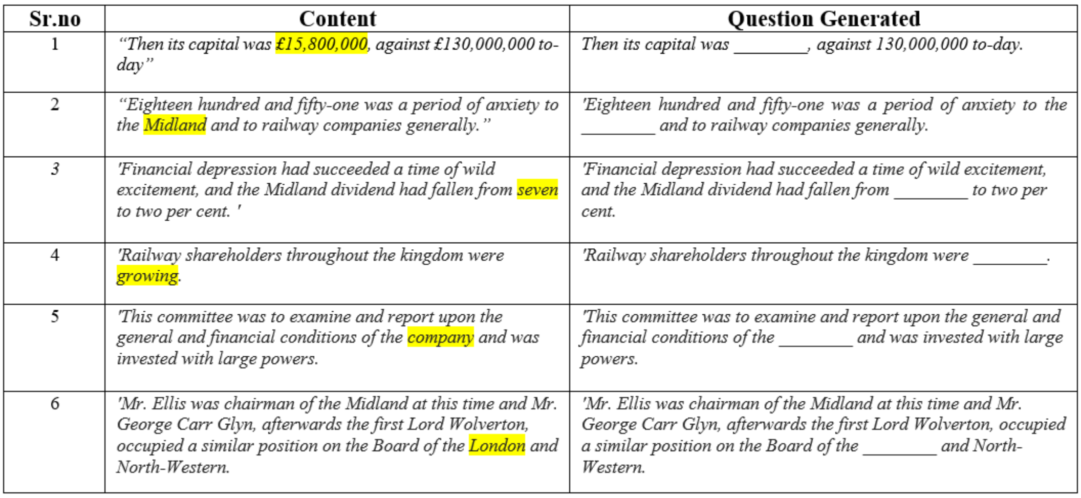
Figure 6: True or false output questions

6.2 Experiment 2 : Experiment on Geography book by Joseph Tatlow named ”Fifty Years of Railway Life in England, Scotland and Ireland”[2]

Next experiment is performed on geography data. This is also taken from the Project Gutenberg site . Here, only one chapter is selected i.e, the third chapter on ”The mildland railway and ”King Hudson” consisting of over 14 paragraphs. It generated 52 fill in the blanks and around 63 true or false questions which is higher than the earlier data experiment. The highlighted text is the pivotal word chosen to generate the fill in the blank. It can be observed, that the generated output is quite good quality ignoring a few bad quality output statements.
The code eliminates all the unnecessary characters. This step was included as the main data used for developing the model had many characters which were not used correctly. Due to that the symbol ”£” got removed along with the unnecessary characters which is might make it difficult for one to understand what do the numbers in the question represent. To overcome this the code will be tweaked for accepting the currency symbols. For the rest of the questions generated they are grammatically and logically correct.

Taking a look at the overall generated output it can be said that the result is a positive, as the questions generated are logically and grammatically correct. Although, this is not true in case of each and every sentence, but it can be said that the geography data does produce some good quality questions. It is important that in order to generate

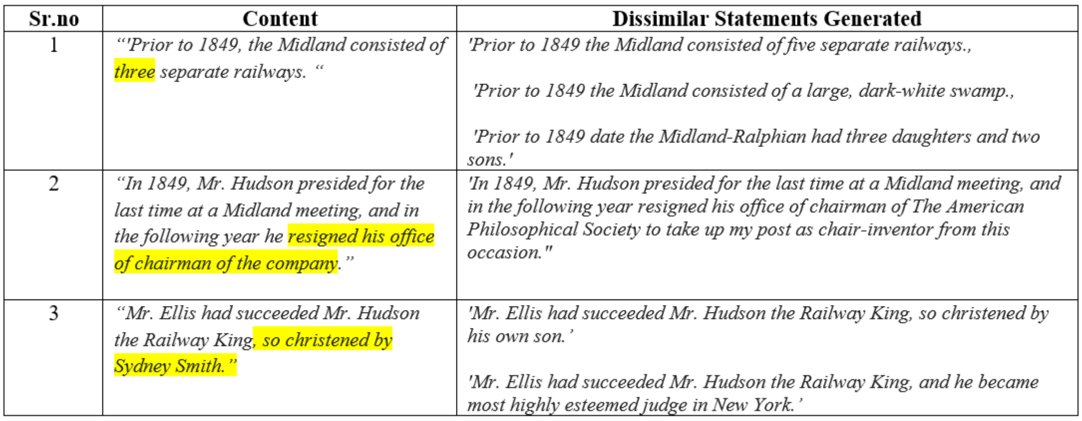
[2] https://www.gutenberg.org/cache/epub/20220/pg20220-images.html



Sr.no	Content	Question Generated
1	*“Then its capital was £15,800,000, against £130,000,000 to-day”*	*Then its capital was ________, against 130,000,000 to-day.*
2	*“Eighteen hundred and fifty-one was a period of anxiety to the Midland and to railway companies generally.”*	*'Eighteen hundred and fifty-one was a period of anxiety to the ________ and to railway companies generally.*
3	*'Financial depression had succeeded a time of wild excitement, and the Midland dividend had fallen from seven to two per cent. '*	*'Financial depression had succeeded a time of wild excitement, and the Midland dividend had fallen from ________ to two per cent.*
4	*'Railway shareholders throughout the kingdom were growing.*	*'Railway shareholders throughout the kingdom were ________.*
5	*'This committee was to examine and report upon the general and financial conditions of the company and was invested with large powers.*	*'This committee was to examine and report upon the general and financial conditions of the ________ and was invested with large powers.*
6	*'Mr. Ellis was chairman of the Midland at this time and Mr. George Carr Glyn, afterwards the first Lord Wolverton, occupied a similar position on the Board of the London and North-Western.*	*'Mr. Ellis was chairman of the Midland at this time and Mr. George Carr Glyn, afterwards the first Lord Wolverton, occupied a similar position on the Board of the ________ and North-Western.*

Figure 7: Output of the questions generated from the history book data

good quality questions a very good quality data input is required.
From the true or false type of questions, it is realized that the OpenAI-GPT2 algorithm makes quite a good changes. The details about how many of the questions are really helpful will be discussed in the result section. For both the experiments quite a good number of questions have been generated from the developed system.



Sr.no	Content	Dissimilar Statements Generated
1	*“'Prior to 1849, the Midland consisted of three separate railways. “*	*'Prior to 1849 the Midland consisted of five separate railways.,* *'Prior to 1849 the Midland consisted of a large, dark-white swamp.,* *'Prior to 1849 date the Midland-Ralphian had three daughters and two sons.'*
2	*“In 1849, Mr. Hudson presided for the last time at a Midland meeting, and in the following year he resigned his office of chairman of the company.”*	*'In 1849, Mr. Hudson presided for the last time at a Midland meeting, and in the following year resigned his office of chairman of The American Philosophical Society to take up my post as chair-inventor from this occasion."*
3	*“Mr. Ellis had succeeded Mr. Hudson the Railway King, so christened by Sydney Smith.”*	*'Mr. Ellis had succeeded Mr. Hudson the Railway King, so christened by his own son.'* *'Mr. Ellis had succeeded Mr. Hudson the Railway King, and he became most highly esteemed judge in New York.'*

Figure 8: True or False output of geography data

6.3 Result

Number of questions generated from both the experiments are calculated together if they are correct or not. Summary of the questions generated is as shown in the below table (this is done manually).

	Fill in the blank (Total 133)	True or false (Total 121)
Correct	72	81
Incorrect	61	40
Success rate	54.13%	66.94%

The above mentioned success rate might vary as every time the code is refreshed it generates a different output. Number of generated questions is definitely higher than any human could generate.

6.4 Discussion

Generating output from a data like "wikibooks" is a very tough challenge but still the developed system has a success rate of more than 50%. More than 50% of positive outcomes from both the questions and the ability to generate more than 100 questions per chapter is amazing. A few limitations for this study will be discussed. Overcoming these limitations will help in improving the positive outcomes of the question generator. Following are the findings from the experiments carried out:

- Removing unnecessary characters : As the base data had too many characters and symbols which were not required, they were ignored and only the letters, numbers and punctuation's were taken into consideration. The limitation of this was observed while performing experiment 2. The symbol "£" is a special symbol and it got removed in the cleaning steps due to which when the question was generated it was difficult to predict the logic behind the numbers in the question.
  To overcome this issue the code is to be tweaked to accept symbols for currencies.

- Generation of true or false questions with an "or" : We have a sentence *"Once acidic or basic substances have been added to pure water, the concentration of the ions will change".*
  False Sentences generated
  "a) Once acidic or basic substances have been added to pure water, the concentration of the ions will rise."
  "b) Once acidic or basic substances have been added to pure water, the concentration of the ions will be higher."
  "c) Once acidic or basic substances have been added to pure water, the concentration of the ions will be higher than normal."

  Now, the sentence is providing two possibilities but giving a very generic output i.e, concentration of ion will change. Instead it should had been specific to either acidic or basic property. So, sentences with an "OR" can be eliminated.

The prototype developed here shows that performs efficiently for more than 50% , however to be able to deploy this question generator in real-time it's efficiency is to be further increased. The system requires improvements going ahead. A few are already mentioned such as using a very high quality data, eliminating sentences with "or",to understand the main idea behind the paragraphs and then generating questions.

7 Conclusion and Future Work

After the analysis conducted on the developed system it could can be concluded that it meets the objectives of the research project and yields results more than 50%. With a few exceptions, the quality of the questions that were created was good. The developed system generated grammatically and logically correct questions although the dataset was a big challenge to deal with. As the data for training is challenging and demands a lot of pre-processing the output question may sometime be created as logically incorrect. Utilization of spaCy throughout the project has helped understand many aspects of the data, and as there are many features available it makes the tasks easier. Cosine similarity helped in not having repetitive questions as output. Although the developed system can be considered successful it is has its own limitations due to data quality, cleaning and techniques deployed. Limitations of not being able to consider sentences with "or" does impact the success rate of the deployed system as it creates confusing questions. To improvise in the future a feedback system can be used to analyse if the generated question is good quality or not and iterations to be carried out. Ahead, a question generating model that can produce all potential questions that can be answered with a single line as the next phase.

References

Agarwal, M. and Mannem, P. (2011). Automatic gap-fill question generation from text books, *Proceedings of the sixth workshop on innovative use of NLP for building educational applications*, pp. 56–64.

Agarwal, M., Shah, R. and Mannem, P. (2011). Automatic question generation using discourse cues, *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 1–9.

Amidei, J., Piwek, P. and Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018.

Azevedo, P., Leite, B., Cardoso, H. L., Silva, D. C. and Reis, L. P. (2020). Exploring nlp and information extraction to jointly address question generation and answering, *IFIP International Conference on Artificial Intelligence Applications and Innovations*, Springer, pp. 396–407.

Blšták, M. and Rozinajová, V. (2021). Automatic question generation based on sentence structure analysis using machine learning approach, *Natural Language Engineering* pp. 1–31.

Chinkina, M., Ruiz, S. and Meurers, D. (2020). Crowdsourcing evaluation of the quality of automatically generated questions for supporting computer-assisted language teaching, *ReCALL* **32**(2): 145–161.

Cruz, R. R. D. L., Khalil, A. and Khalifa, S. (2021). Automatic multiple-choice and fill-in-the-blank question generation from arbitrary text, *Future of Information and Communication Conference*, Springer, pp. 244–257.

Curto, S., Mendes, A. C. and Coheur, L. (2012). Question generation based on lexico-syntactic patterns learned from the web, *Dialogue & Discourse* **3**(2): 147–175.

Das, R. and Elikkottil, A. (2010). Automatic summarizer to aid a q/a system, *International Journal of Computer Applications* **975**: 8887.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Heilman, M. and Smith, N. A. (2010). Good question! statistical ranking for question generation, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 609–617.

Indurthi, S. R., Raghu, D., Khapra, M. M. and Joshi, S. (2017). Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 376–385.

Khullar, P., Rachna, K., Hase, M. and Shrivastava, M. (2018). Automatic question generation using relative pronouns and adverbs, *Proceedings of ACL 2018, Student Research Workshop*, pp. 153–158.

Kriangchaivech, K. and Wangperawong, A. (2019). Question generation by transformers, *arXiv preprint arXiv:1909.05017* .

Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes, *International Journal of Artificial Intelligence in Education* **30**(1): 121–204.

Mhatre, K., Thube, A., Mahadeshwar, H. and Shrivas, A. (2019). Question generation using nlp, *International Journal of Scientific Research & Engineering Trends* **5**(2): 2395–566.

Murugan, S. and Ramakrishnan, B. S. (2021). Automatic morpheme-based distractors generation for fill-in-the-blank questions using list wise learning-to-rank method for agglutinative language, *Engineering Science and Technology, an International Journal* **Vol. 26**.

Nwafor, C. et al. (2021). An automated mulitple-choice question generation using natural language processing techniques, *International Journal on Natural Language Computing (IJNLC) Vol* **10**.

Panchal, P., Thakkar, J., Pillai, V. and Patil, S. (2021). Automatic question generation and evaluation, *Journal of University of Shanghai for Science and Technology* **23**: 751–761.

Schleicher, A. (2020). The impact of covid-19 on education insights from education at a glance 2020, *Retrieved from oecd. org website: https://www. oecd. org/education/the-impact-of-covid-19-on-education-insights-education-at-a-glance-2020. pdf* .

Zerr, J. (2014). Question generation using part of speech information, *Final Report for REU Program at UCCS* pp. 19–23.