

Automated MCQ Generator using ACG Techniques & Spacy's NER Model

Pradeep Sudakar, Balamurugan R, Priya G

VIT, VELLORE CAMPUS, TIRUVALAM RD, KATPADI, VELLORE, TAMIL NADU 632014

Abstract

Health and education are the two most significant fields that appear to have no end. Numerous fields can be used to further categorize education. Being literate in today's society is merely a minimum requirement for survival; we also need the ability to reason. Being healthy is all about surviving and treating ailments like cancer, which can also be prevented through education. Old customs like Ayurveda may not be considered as an education standard. However, the educational component of the current generation will be the focus of our project.

As we all know, a child's education always starts with learning the alphabet, then words, and meanings, and finally, framing sentences. A child learns all the other fundamental subjects like English, history, politics, science, geography, and mathematics once they are at least somewhat comfortable with any language. These students learn how to apply logical and analytical thinking to their studies, problem-solve, and brainstorm. Comprehensive skills would be the fundamental area of testing in education for any subject at the beginner level. The student must be able to understand the passage and be able to answer all the questions asked with the help of this passage.

Given the current increase in demand for education, it is difficult for each individual faculty to assess the student's knowledge of all relevant skills. The teacher must be able to offer various sets of test questions to prevent student cheating in order to get better results. We must be able to use current technology to solve our issues because it takes a lot of time to create unique test questions for each student. With multiple choice questions and automated results that can be published at the end of the test, this project creates a quiz from a random passage. Students can access various collections of thorough exercises in this way without having to deal with the burden of instructors.

Key Words: MCQs, BERT, Wordnet, PKE, Deep learning

1. INTRODUCTION

E-learning has become a rapidly expanding area of education in the current generation. Our application enables our users, who are primarily students, to upload content and learn anything from anywhere they choose. A thorough quiz is a fundamental exam used to gauge how well primary students have read and comprehended the passages. Any language may be used to implement this. The students must be able to read the assigned passage, comprehend it, analyze it, and respond to a set of questions using the assigned passage as support. In terms of the application's fundamental ideas, we heavily rely on NLP library files and the Python programming language.

The outdated practices of writing exams on paper with a pen have not been utilized as much by this generation. Every educational department expresses interest in administering exams online to streamline the process. Since the population is rapidly growing, faculty members would also be under pressure to ensure that students did not copy answers on exams.

3. Methodology

3.1) Model Description

The MCQ Quiz generator begins by receiving a text file from the user that comprises a predefined paragraph written in proper language. When the Txt file is ready to be uploaded and scanned, it takes you to another page with all of the questions and alternatives created by the file. The following is a full description of the workflow.

1. *Input* – For convenience of creation, the txt file must include a single broad theme with several smaller subtopics solely in English.
2. *Text Pre-processing* — When the text is pre-processed, the natural language models will detect it. All non-alphanumeric characters (excluding full stops) are ignored in the natural language model's output.
3. *Named Entity Recognition + Entity Ranking* — To find entities in text data, Spacy's NER model is employed. These things include people's names, dates, locations, and amounts, among others, since they provide value to the inquiries that must be created.
4. *Incorrect Option Generation* - Word2Vec is the model utilized in this step. This model creates comparable entities that were recruited for option creation in the previous step. If a circumstance develops in which no corresponding entity can be discovered, words from the provided text can be chosen.

3.2) Module Decomposition

We have divided our project into three major modules namely:

1) User interface

UI is the point at which our users would interact with our site and we have developed the frontend using HTML, CSS and JavaScript. We are planning to have 3-4 major web pages and the design & functionalities to be basic and simple as we are targeting school student

2) Questions formation using NLP

This is the core module which contains basically all the algorithms and actual procedures involved in generating a meaningful quiz. As discussed we will be mainly using the tools like *Spacy's NER model*, *TF-IDF scores*, *WordToVec*, etc.,.

3) Quiz generation

This module is more like the output of our application which consists of the meaningful Quiz that's generated by our NLP algorithm from the given text file. We also decided to show the marks after the user finishes the quiz.

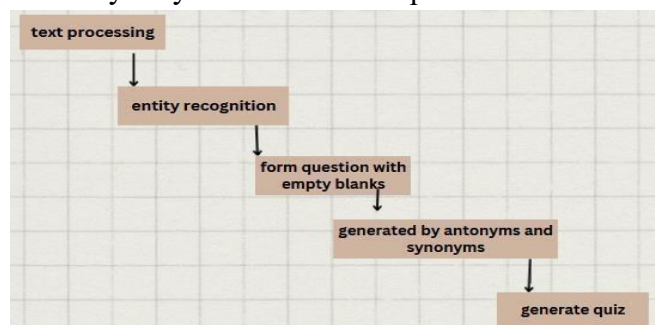
3.3) NLP Libraries

- ❖ **Genism** - It's an open-source library that is mainly used for representing documents as semantic vectors. It was designed so that it could process raw documents, and unstructured digital texts by making use of unsupervised machine learning algorithms
- ❖ **Word2Vec** - This model is used to create vectors of the words that are distributed numerical representations of word features. It uses a shallow neural network for word embeddings. Its input is a text corpus and its output is a set of vectors that represent words in that corpus. It turns text into a numerical form that deep neural networks can understand.
- ❖ **NLTK** - It is a platform that is used in building python programs that interact with human language data for applying in statistical natural language processing
- ❖ **Sent_tokenize** - This comes with the NLTK package, and it is used to split a document or paragraph into sentences.
- ❖ **Word_tokenize** - This also comes with the NLTK package, and it is used to split a sentence into tokens or words.
- ❖ **Sci-kit Learn** - The most helpful machine learning library in Python is Sci-kit Learn. It includes a variety of useful tools for machine learning and statistical modelling, such as classification and regression.
- ❖ **TfidfVectorizer** - This function is included with Sci-Kit Learn and is used to convert text into feature vectors that can be fed into the estimator. It computes a word occurrence frequency sparse matrix by mapping the most frequent words to features generated using an in-memory vocabulary that is a Python dict.

4. Model chosen

4.1) Model Diagram

Our solution is mostly based on natural language processing (NLP). When a pdf/txt file is uploaded, the text is initially analyzed and entities are recognized. Important verbs and nouns are discovered and translated into blanks in questions using NLP. Alternate options are generated by antonyms and synonyms and the final quiz is made.

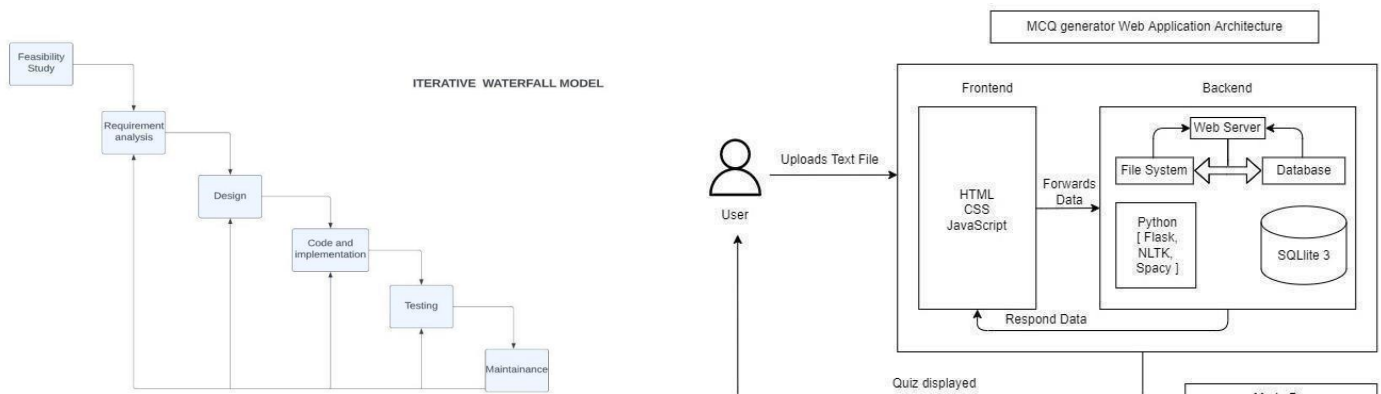


4.2) Process Model Diagram

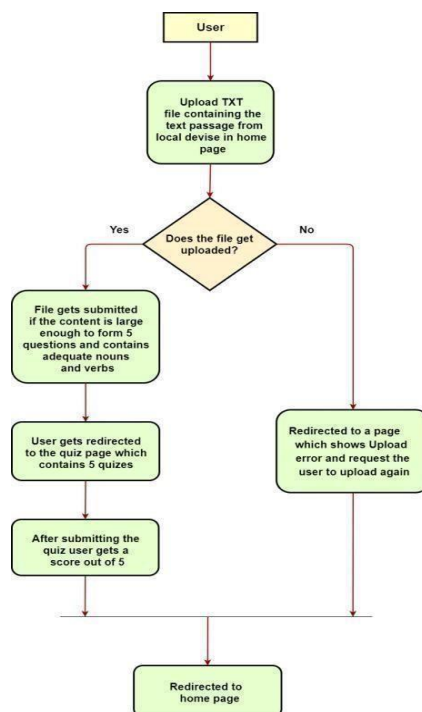
The suitable model for our project is the Iterative Waterfall model as requirements are well defined and errors can be detected and corrected in each phase we won't have to wait till the end to rectify the errors.

Advantages:

It is easy to understand and implement. Suitable for single projects where work products are well defined and their functioning is understood.



4.3) Data Flow Diagram



1.1.1 5. Architecture Diagram

Literature Review

1. AUTOMATED QUESTION GENERATOR SYSTEM USING NLP LIBRARIES

This paper was published on June 2020 by Priti Gumaste, Shreya Joshi, Srushtee Khadpekar, and Shubhangi Mali in the International Research Journal of Engineering and Technology (IRJET). Their system takes a paragraph and first pre-processes using semantic and syntactic analysis which basically includes POS tagging, Chunking, and Named Entity Recognition. Then the text is mapped with a “wh” question and quiz is generated. The accuracy of the model is said to be 70% but they have also mentioned a lot of improvements are needed like applying Bloom’s taxonomy, storing result in the database for re-usability, and increasing the precision which might help them to make their user select the difficulty level.

2. AUTOMATIC GENERATION OF MULTIPLE CHOICE QUESTIONS USING DEPENDENCY-BASED SEMANTIC RELATIONS

Naveed Afzal and Ruslan Mitkov published this paper via a Softcomputing journal. Their approach was to extract questions using semantic techniques rather than syntactic and surface relations. They first processed unannotated corpus with the help of NER and pattern ranking evaluation and semantic relation. Three main components of their approach are IE methodologies, semantic relation and using a distributional similarity measure. There have been no work previously that uses semantic relations based on information extraction methodologies in the context of MCQ generation but coming to the disadvantages it’s only restricted to the biomedical domain and it can’t be used to generate MCQ for other domains.

3. A QUESTION-GENERATION ENGINE FOR EDUCATIONAL ASSESSMENT BASED ON DOMAIN ONTOLOGIES

This paper was published in the 11th IEEE International Conference on ALT by Maha AlYahya. The approach he took was mainly based on strategy. He named his model THE ONTOQUE ENGINE. The 3 strategies were class-membershipbased strategy, individual based strategy, and property-based strategy and RDF keys. The main advantage of his work was along with the MCQ his model was also capable of generating True/False and fill-the-blanks questions but still as the name suggest it restricted to the ontology domain and it has some limitation of generating less number of question and difficulty was often noticed to be easy.

4. SHERLOCK: A SEMI-AUTOMATIC QUIZ GENERATION SYSTEM USING LINKED DATA

Dong Liu¹ and Chenghua Lin published this paper on 2014 in In International Semantic Web Conference. The architect of this model is composed of mainly data collection, training, similarity computation, RDF key and template based quiz render.

The site also enables the user to manually edit some of it after the quiz is generated and also evaluation takes place but the accuracy of the system is reported to be only 55 % and they have mentioned there's scope for lot of improvement like allowing user to opt different difficulty level and providing with larger data set for increasing the present accuracy.

5. THE DESIGN OF AUTOMATIC QUIZ GENERATION FOR UBIQUITOUS ENGLISH E-LEARNING SYSTEM

This paper was published in Technology Enhanced Learning Conference (TELearn) on 2017 by Li-Chun Sung, Yi-Chien Lin and Meng Chang Chen. Their core concept was based on SemNet (semantic network). The algorithm first lexically marks all the components as verb, adjective, noun and then tries to understand the link between each phrase and form a network. It uses a dynamic knowledge extension strategy. Coming to implementation they have used just a basic java platform which is unattractive and old-style and they also stated a more extensive knowledge base is required for future improvements but still the precision of model is proclaimed to be good.

6. SEMANTIC ATTRIBUTES MODEL FOR AUTOMATIC GENERATION OF MULTIPLE CHOICE QUESTIONS

Fattah I, Aboutable A and Haggag M published this paper on 2014 in the International Journal of Compute Applications. The main techniques used in their model was Semantic Role Labeling (SRL) and Named Entity Recognition (NER). A huge sentence knowledge base was also included and the questions were formed with the help of text similarity approach. They have almost combined 8 different algorithms of string based similarity and also introduced a classification to identify the question difficulty level. Their N-gram algorithm found to produce the highest level of questions difficulty. The main advantage of their work is including a lot of approaches and combining them properly which made their system precision way too good and they have also stated that the model could be further improved using corpus-based similarity and knowledge base similarity algorithms.

7. AUTOMATIC GENERATION OF MULTIPLE CHOICE QUESTIONS FROM DOMAIN ONTOLOGIES

This paper was published in IADIS International Conference elearning 2008, Amsterdam, the Netherlands in January 2008. This paper helps the reader understand how quiz based on multiple choice questions can be generated using OWL technology which is based on Standard Web Based technology. The quiz generated is independent of lexicons. The implementation has been explained using various strategies that are related to ontology based problem generation. They include class based strategies, property based strategies and terminology based strategies. However, the results produced from implementation has showed best outputs only in the case of providing strong domain ontologies. This is one drawback in this paper. There are some corrections to be made in NLG system. Real life testing also needs to be done in some elementary schools.

8. AUTOMATED QUIZ GENERATOR

This paper was published as a part of Advances in Intelligent Systems and Computing book series where it was first made online on 21st October 2017. This paper is basically an extension of an implementation proposed by Michael Heilman. The add ons in this paper include two things. One is trying to publish related multiple choices for the corresponding question that has been generated. Other is the ability to rank these questions based on a priority level which depends on the input text that the corresponding question has been generated. Interestingly, the implementation has an additional feature where it can get help from Wikipedia articles for question and multiple choice generation. This system is completely dependent on a database called DBpedia which consists of structured content extracted from Wikipedia. There are some drawbacks where the quiz generation can't be generated with accuracy. These factors are basically external factors like wrong coreference resolution and wrong entity recognition.

9. AUTOMATED QUESTION PAPER GENERATION SYSTEM

This paper was published in April 2016 by Rohan Bhirangi and Smita Bhoir in the International Journal of Emerging Research in Management & Technology. In this, they have proposed automatic question paper generation based upon the Shuffling Randomization Algorithm and Role-Based hierarchy model. They have integrated an efficient question marking system in the revised version of the shuffling algorithm. The algorithm uses a basic randomization algorithm which has a flag system to mark selected questions. After a question paper is generated, it is converted into an encrypted pdf and can be distributed with a single click. The input for the algorithm is the number of questions of each course. The probability that the same question paper will be generated is $1/362880$.

10. AUTOMATIC MULTIPLE CHOICE QUESTIONS GENERATION FROM TEXT

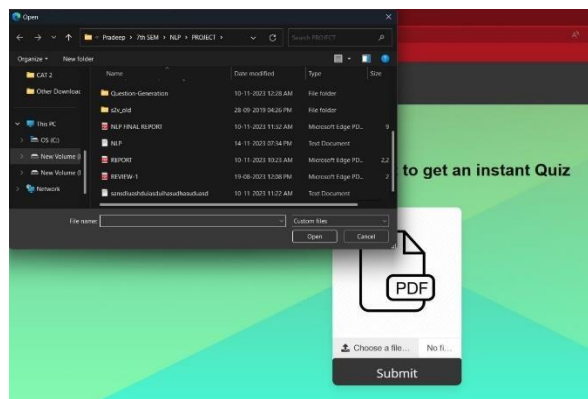
This paper was published in May 2021 by Pranali Patil, Nikeeta Patil, Kratika Kumari, Devendra Ingale, A.R.Uttarkar in the International Journal of Scientific Research & Engineering Trends. In their approach, the input content is first summarized using the BERT calculation. To make decisions over questions, the distracters are created using WordNet which is a lexical dataset for English. Python keyword extractor (PKE) which is an opensource python toolkit is used to extract the keywords from the content, and then only the keywords which are present in summarized text are kept. WordNet is used to get the correct sense of the word. The outcome of this approach is that using BERT gives better results than other text summarisers and question generators.

1.2 output

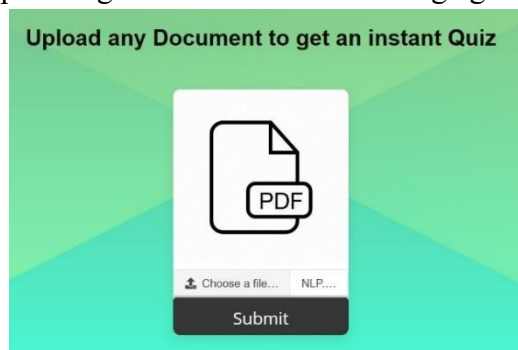
Here you can upload any type of document you want



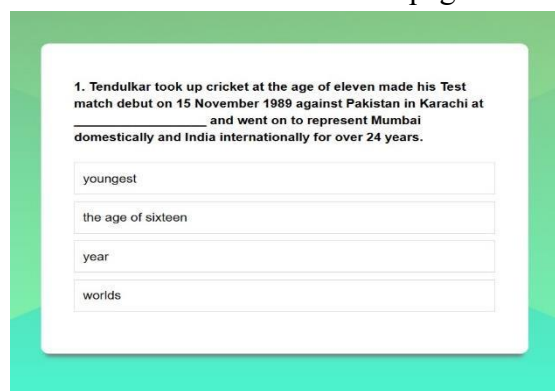
It will direct you to the specific place where you want to select the file .shown below



After uploading it will show like the image given below



After clicking the submit button it will move to the next page which is the MCQ page



The option turns green if it is the correct option

1. Tendulkar took up cricket at the age of eleven made his Test match debut on 15 November 1989 against Pakistan in Karachi at _____ and went on to represent Mumbai domestically and India internationally for over 24 years.

youngest

the age of sixteen

year

worlds

The option turns red if it is the incorrect option

2. The same year Tendulkar was a part of the team that was one of the jointwinners of the 2002 ICC Champions Trophy.

the year

years

worlds

world

At the end, there is a Submit button where you can end your test

5. In ____ Time included Tendulkar in its annual list of the most influential people in the world.

2001

2010


1998

2000

Submit

DONE BY: PRADEEP SUDAKAR 20BCE0834 & BALAMURUGAN 20BCE2469

After completion of the test, the marks will be displayed



You got 5/5 right!

Upload another document

DONE BY: PRADEEP SUDAKAR 20BCE0834 & BALAMURUGAN 20BCE2469

Results

1. Analysis of Result

Based on the analysis of the result, we were able to effectively complete the work of reading the text file that had been uploaded, determining the number of questions required, and then generating a system of relevant questions with closely connected alternatives. After the user has completed the quiz, they may review the results.

Our research use natural language processing (NLP) techniques to scan the data corpus, evaluate it, and produce questions based on the user's content. For the production of questions and alternatives, we have used NLP packages and libraries like as nltk, spacey, and word tokenizer. The website features a basic interface that makes it userfriendly and straightforward to navigate. The application can be useful for self evaluation and for conducting assessments mainly for kids.

2. Comparison with existing system

The existing systems does not provide a proper implementation though, the algorithm has been suggested in some of the research papers. Most of the present day applications are limited to a certain field of study such as English literature and psychology whereas, our application is not limited to any field. User can provides a set of questions for any topic and an instant quiz will be generated.

Existing systems usually limit the number of questions generated to three to five, whereas our system allows the user to specify the desired number of questions while uploading the file and also allows the number of options generated to be determined by the user, making our application more user-friendly and interactive. Options for every question is also generated which are all synonyms and antonym of actual answer making the choice difficult for users. Our quiz's results would be created immediately after hitting the submit button, and the correct answers would be displayed to the user, providing a superior learning experience than existing models.

Conclusion

As previously stated, our software successfully creates a multiple-choice quiz with alternatives from a text file submitted by users using NLP algorithms. Once the quiz has been submitted, it also displays the quiz's results.

However, as this project focuses mostly on offering the most relevant alternatives to make it helpful for learners, an extension of this project would be altering the present algorithm into a more sophisticated model for providing a relevant option for increasing the difficulty level of the quiz.

It was a fantastic learning experience to work on this project. We gained a better understanding of NLP and AI principles. This experiment provided us an idea of how much time and work goes into developing a language-based NLP model. We also learnt about a variety of Python modules that helped us come up with interesting queries. We put in a lot of effort to learn about existing systems and provide the finest and most engaging model possible. At the end of the project we got a good knowledge on NLP & AI concepts which will be definitely useful for us in future works.

REFERENCES

- [1] Gumaste, P., Joshi, S., Khadpekar, S. and Mali, S., 2020. AUTOMATED QUESTION GENERATOR SYSTEM USING NLP LIBRARIES. *International Research Journal of Engineering and Technology (IRJET)*, 7(6), pp.4568-4572.
- [2] Afzal, N. and Mitkov, R., 2014. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing*, 18(7), pp.1269-1281.
- [3] Al-Yahya, M., 2011, July. OntoQue: a question generation engine for educational assessment based on domain ontologies. In *2011 IEEE 11th International Conference on Advanced Learning Technologies* (pp. 393-395). IEEE.
- [4] Liu, D. and Lin, C., 2014, October. Sherlock: A semi-automatic Quiz Generation System using Linked Data. In *International Semantic Web Conference (Posters & Demos)* (pp. 912).
- [5] Sung, L.C., Lin, Y.C. and Chen, M.C., 2017. The design of automatic quiz generation for ubiquitous English e-learning system. In *Technology Enhanced Learning Conference (TELearn 2017)*, Jhongli, Taiwan (pp. 161-168).
- [6]
- [7] Fattoh, I., Aboutable, A. and Haggag, M., 2014. Semantic attributes model for automatic generation of multiple choice questions. *International Journal of Compute Applications*, 103(1), pp.18-24.
- [8] Papasalouros, A., Kanaris, K. and Kotis, K., 2008. Automatic Generation Of Multiple Choice Questions From Domain Ontologies. *e-Learning*, 1, pp.427-434.
- [9] Bongir, A., Attar, V. and Janardhanan, R., 2017, September. Automated quiz generator. In *The International Symposium on Intelligent Systems Technologies and Applications* (pp. 174-188). Springer, Cham.
- [10]
- [11] Bhirangi, R. and Bhoir, S., 2016. Automated question paper generation system. Computer Engineering Department, Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra, India.
- [12] Patil, M.P., Patil, M.N., Kumari, M.K., Ingale, M.D. and Uttarkar, M.A., 2021. Automatic Multiple Choice Questions Generation from Text