Learning Quiz Game using NLP and Knowledge Representation

Tanishka Dhondge
Computer Science and Technology Department
Usha Mittal Institute of Technology
Mumbai, India
dhondgetanishqa@gmaill.com

Dhruva Kamble
Computer Science and Technology Department
Usha Mittal Institute of Technology
Mumbai, India
kambledhruva.1@gmail.com

Shubhika Dev
Computer Science and Technology Department
Usha Mittal Institute of Technology
Mumbai, India
shubhi17dev@gmail.com

Sumedh Pundkar
Computer Science and Technology Department
Usha Mittal Institute of Technology
Mumbai, India
sumendhpundkar@gmail.com

***Abstract*— The expanding application of natural language processing (NLP) in the learning and educational domains served as the inspiration for this research, which precisely integrates NLP and knowledge representation. In the interest of developing a game that is similar to a quiz, a semantic triple of knowledge representation is utilized in combination with the NLTK and Spacy modules of the Python programming language. The game includes multiple questions regarding the subject given as input. The project aids in assessing one's understanding of a particular subject and one's own knowledge of it. Data collection or extraction on the given input is the first step in the process, followed by the creation of a knowledge representation to help comprehend the relationships between the entities in the data. For name entity recognition, a few crucial phases include data cleansing, POS tagging, and semantic triples. The development of a question generator would require expertise in natural language processing, machine learning, and data analysis. It involves designing algorithms and making proper use of existing modules that can analyze text and generate questions based on the content of the text.**

***Keywords*— Knowledge Representation, Natural language Processing, Semantic Triples, POS Tagging, NER (Name Entity Recognition), NLTK (Natural Language Toolkit), SciPy (Scientific Python), RDF (Resource Description Framework).**

I. Introduction

In the field of education, there is an increased use of Natural Language Processing(NLP) and Knowledge representation (KR) to create smart tools for learning. These tools help aid the creation, analysis, and assessment of students and educators. Simply studying a subject is insufficient; having checkpoints to assess the level of understanding is also necessary. This promotes greater comprehension and gives a chance to augment knowledge of that particular subject. Assessing one's comprehension is just as crucial as learning. The development of this educational game facilitates understanding self-analysis.

Interactive learning through games can be fun and more effective. It is also a very important method to keep users engaged in the learning process. An effective learning curve can be achieved by choosing self-paced and self-curated learning processes. Kenneth C. Litkowski
CL Research
The project's main areas of focus are knowledge representation using the semantic triple and NER (Name Entity Recognition), and natural language processing. The backend method for creating the game requires work on Natural Language Processing. Several crucial operations begin from scratch and involve gathering information on the subject at hand, cleaning it up, turning it into semantic triplets, and identifying and adding the POS (part of speech) tags.

The goal is to develop an interactive game that functions like a quiz to assess users' knowledge of that particular subject. Any domain-independent text data can be provided as input which can be supplied as a pdf or Txt format file. A set of multiple-choice questions on the given input will be generated as a result. The users can quantify their understanding on the basis of the score they obtain after submitting the quiz. Through the creation of questions that concentrate on the most significant aspects, it may be used to assist individuals in finding the essential details in an immense amount of text rapidly. Users could use it to develop practice questions to assess their comprehension of a specific topic

The objective of the proposed learning game is to make a substantial impact on learners and facilitate learning using Natural Language Processing and Knowledge representation.

II. Literature Review

POS Tagging approaches offer a comprehensive comparison of rule-based, statistical, and hybrid approaches to Part-of-Speech (POS) tagging. The study evaluates their accuracy and performance in different languages and corpora, identifying the strengths and weaknesses of each method. While statistical approaches, such as Hidden Markov Models and Conditional Random Fields, generally perform better, hybrid approaches show promise. The survey also discusses the challenges of POS tagging, including ambiguity and unknown words, and explores potential solutions. The findings provide valuable insights for researchers and practitioners in the field of natural language processing.[1]

The Natural Language Processing of Semitic Languages such as Hebrew, Arabic, and Amharic provides a comprehensive survey of the challenges and opportunities presented by natural language processing (NLP) in Semitic languages. The study examines the unique linguistic features of Semitic languages and explores various NLP tasks, highlighting the state-of-the-art techniques for each. Additionally, the research identifies the current limitations and future directions of NLP research in Semitic languages such as the need for domain-specific resources, standardization, and scarcity of computational tools and resources which further complicates NLP tasks.[2]

Automatic Question Generation(AQG) Based on Sentence Structure Analysis Using a Machine Learning Approach presents a survey of various approaches to automatic question generation, focusing on machine learning-based techniques that analyze sentence structure. The study compares the performance of different models, such as Recurrent Neural Networks and Support Vector Machines, in generating questions from the text. The research proposes a machine learning-based approach for automatic question generation, which uses sentence structure analysis to identify key components and generate questions based on specific criteria. The results show that machine-learning approaches can achieve high accuracy in question-generation tasks. [3]

A Survey on Natural Language Processing Tasks and Techniques offers a comprehensive overview of various natural language processing (NLP) tasks and techniques. The study examines NLP tasks, including text classification, sentiment analysis, machine translation, and information extraction, and investigates the different approaches employed, such as rule-based, statistical, and neural network-based methods. The research identifies challenges in NLP, such as linguistic ambiguity and insufficient annotated data, and offers potential solutions such as the use of unsupervised and semi-supervised learning, and the incorporation of domain-specific knowledge. [4]

Preprocessing of texts is a very important and primitive step in NLP. There are various pre-processing techniques that help to clear the noise, reduced the size of the data to only effective text size.Tokenization,Stopword removal,stemming are some techniques used to process data that help improve the accuracy of the NLP models. These help to improve the Information retrieval performance[5]

Word2Vec is a neural network proposed by Google that makes use of two algorithms to process data and find similar words in a large corpus of data.In order to use Word2Vec in larger data sets, two steps have been employed. Initially, the input is passed through Word2Vec and similar words are found by using linear calculation.Then K-clustering is used to form the K-cluster of these similar words using K-values. This method increases the speed and decreases the dimension of the data. This helps to increase the training ability to train the big data sets[6]

Term Frequency Inverse Document Frequency (TF-IDF) is a commonly used numeric value that determines the importance of a word in a collection of words or documents. It is a statistically calculated product of term frequency and inverse term frequency. TF-IDF is simple to use but it is not as effective when the text to be classified is distributed unevenly. An improved TF-IDF algorithm can be achieved when the ratio of the document to the total eigenvalue can be replaced by the positive and negative documents that can deal with confusion more effectively.[7]

III. Proposed System

The paper focuses on the proposed method and has two major areas of work. The first one deals with working on raw data. Converting the unstructured data into a knowledge graph and the second phase deals with the generation of multiple choice questions related to the provided input.

A. Phase 1: Working on raw data

Major implementation of Natural Language processing is carried out to process the raw, unstructured data, and semantic triples are used to determine the relationships between two entities. Prior to the process of determining the semantic triple, data is cleaned. Semantic Triple is also known as a subject-predicate-object triple, it is a three-part statement. Typically, the object is a value or another resource, the predicate is an attribute or relationship, and the subject is an entity or resource.

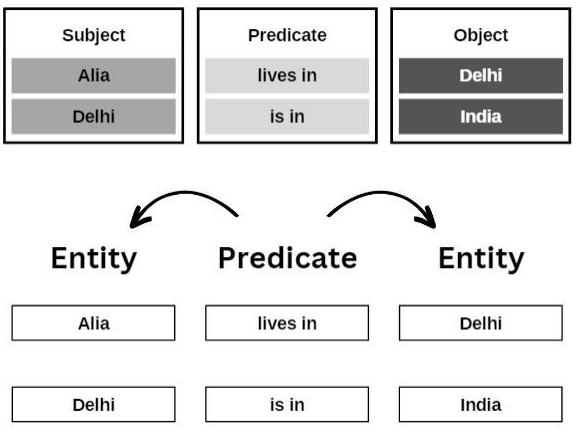

Fig 1. Semantic Triples

The semantic knowledge graphs are constructed using the final set of extracted triples. A knowledge graph often referred to as a semantic network, portrays a network of specific terms, such as things, events, circumstances, or concepts, and shows how they are related to one another.

B. Phase 2: Question Generation

Using the semantic knowledge graph and the processed data to produce an interactive learning game is the second main area of development. NLP and data analysis are done using SciPy and other machine learning modules. The entity collection is ranked, and subsequently, select the necessary numbers of entities from this ranked list to construct question-and-answer pairs. Then, using questions that have been formed by framing sentences, we create alternate inappropriate answers by looking up synonyms and terms that are similar to the correct responses. This way each question will have five options of which only one will be the appropriate response, and the rest will be similar incorrect alternatives. These questions, the appropriate responses, and any additional incorrect responses are grouped together. The questions are all presented in order, offering a quiz-like game.

IV. Methodology

A. Data Extraction

One has to provide a text file containing information about one primary topic and possibly additional subtopics. If the input is in PDF format, then the pdf2text function takes in the path of a file and its extension as arguments. It identifies the file type and extracts its contents. If the file is a PDF, it uses the PdfReader module from the PyPDF2 library to read the contents of each page of the PDF and concatenate them into a single string. If the file is a text file, it reads its contents using the built-in open() function. The function then returns the text content as a string. The txt2questions function takes in the text content, the number of questions to generate, and the number of options per question as arguments.

B. Data Cleaning

Noise abounds in the raw, unstructured data. Data cleaning must be rigorous in order to improve precision and eliminate redundancy from the input data. Among the characters that are removed during data cleaning are stop words, blank spaces, acronyms, duplicate entities, and punctuation except for full stops. The full stop helps to identify the end of a sentence. The sub-modules and functions supported by the nltk module in the Python programming language facilitate all of these procedures.

C. Entity Extraction

Entity extraction is the initial procedure for acquiring semantic analysis. Tokenization is a step in the process that breaks down sentences and paragraphs into smaller, easier-to-understand components. The next step is POS (Part Of Speech) tagging, which refers to categorizing words in a text (corpus) in accordance with a certain part of speech, depending on the definition of the term and its context. The final phase involves chunking, which entails disassembling a sentence to determine its components in order to extract phrases from unstructured material (Noun Groups, Verbs, verb groups, etc.)

D. Name Entity Recognition

Named-entity recognition (NER), also known as (named) entity identification, is a subtask of information extraction that aims to find and classify named entities mentioned in unstructured text into pre-defined categories, such as names of people, companies, locations, time expressions, quantities, monetary values, percentages, etc. To carry out this process mainly two modules in Python Programming Language are used they are - nltk and spacy modules

E. Post Processing

The unstructured material is handled by a variety of procedures before being transformed into semantic triplets. The triplets are then cleaned and deduped once more.

F. Ranking the Entiies

The data in triplet form will be utilized in the second phase. The extracted entities are given rankings, which are calculated using the Term Frequency- Inverse Document Frequency (TF IDF) Score. It is a metric that could determine how significant or relevant string representations (words, sentences, terminology, etc.) are within a document compared to other texts. It is used in the disciplines of information retrieval (IR) and machine learning. Ranking the entities highlights the significance and frequency of each term in the given document.

G. Mask Sentences with Rank Entities

The highly ranked entities from the entire text are generally the keywords or notable subtopics. The entities selected for future applications are those with higher rankings. The purpose is to formulate questions using these key terms. Scikit-Learn and other packages of Python are used to achieve this. These open-sourced libraries are primarily used to perform data analysis.

H. Generate Incorrect Alternatives

The Word2Vec model used in gensim is employed to produce incorrect responses or additional alternatives for this developed question. With the help of a huge text corpus, the word2vec technique employs a model of neural networks to learn word associations. It is employed to turn words into distributed representations of numerical vectors. Word2vec turns text into vectors that represent the relationships and semantics of words. Usually, choose between 10 and 15 entities to improve the game, and from this group, at least four related entities are provided as alternatives.

The same approach is used to create a variety of questions, each with five possible responses. Ultimately, all of these

questions, correct answers, and incorrect options are merged together.

V. Workflow of the Proposed System

The project's first objective is to construct a knowledge graph, and in order to accomplish this, the operations listed below are performed in a sequence.

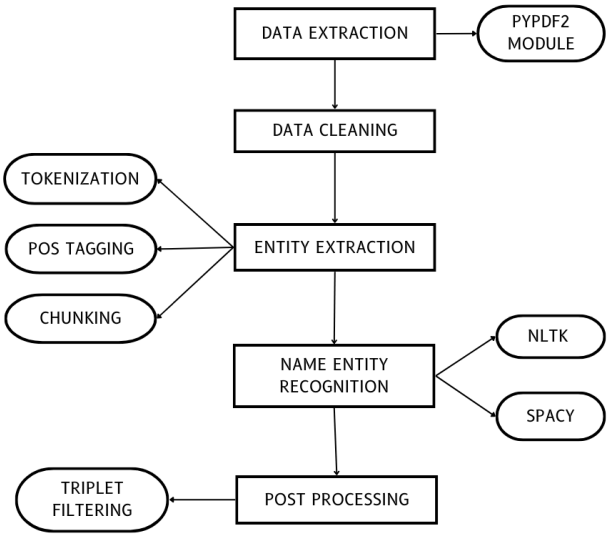

Fig 2. Workflow of the first phase

After the triplets are produced, the subsequent step in the process of creating questions and options to form a quiz is achieved using this set of triplets.

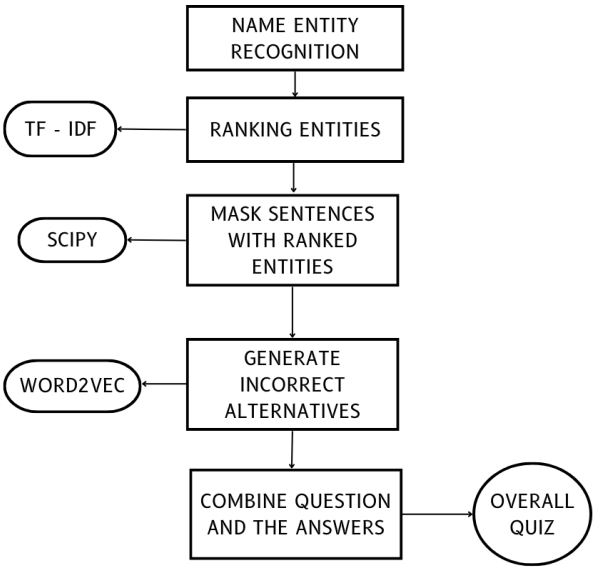

Fig 3. Workflow of the second phase

VI. Result and Discussion

The web application is the final output of the project. It is built using the Flask and Jinja2 Frameworks. The very first site of the application asks for a PDF file as input.

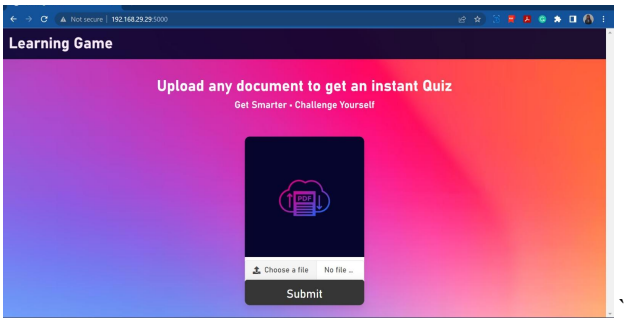

Fig 3. The Landing Page of the Web Application

Once the PDF is provided as the input, the application reads the file and converts it into multiple-choice questions.

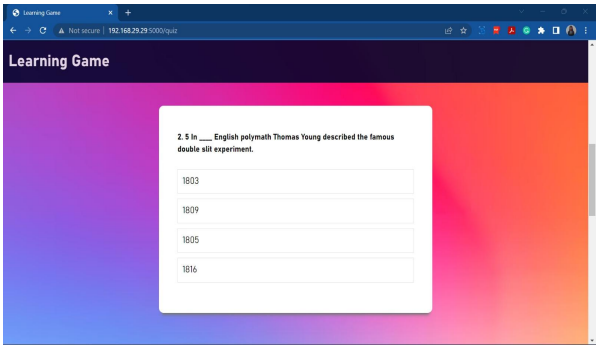

Fig 4. The display of Multiple Choice Questions

The user has clicked on the option she/ he thinks is correct. If the selected option is the correct answer, it reflects through the green color.

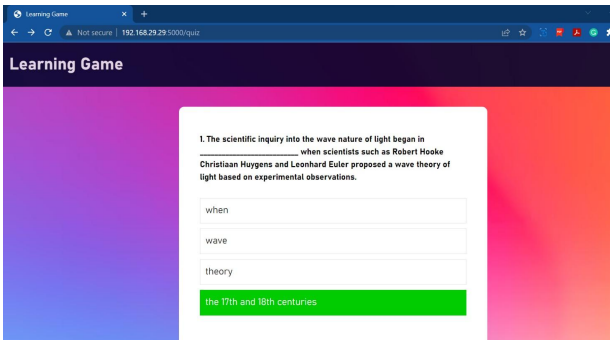

Fig 5. Selecting the correct option

If the selected option is incorrect, then the background changes to red color, hence indicating an incorrect answer.

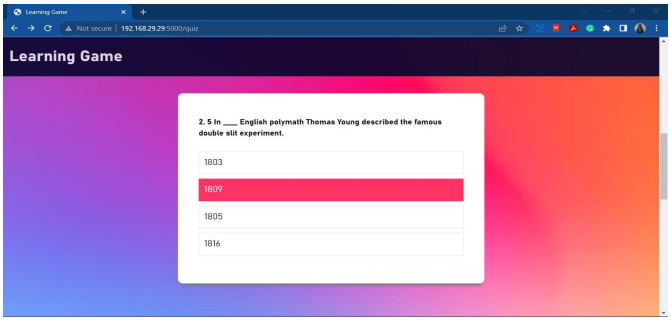

Fig 6. Selecting the incorrect option

Once all the questions are answered by the user, the final page appears which instantly displays the result of the quiz.

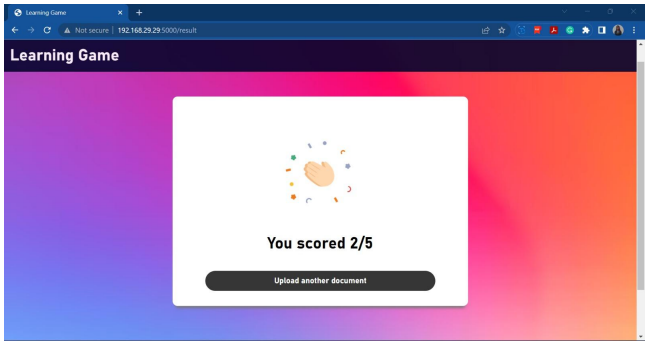

Fig 7. Displaying the result

VIII. Future Scope

NLP is a vast domain where continuous updation is required, and the current project can be widely implemented in the educational department. The current project can be updated and beforehand provided with all the chapters of a particular book, the user can simply type the chapter's name and she/he can get an instant quiz. Currently, there are no limitations on the size of the input which can be modified to avoid extreme strain on the server and model. The current project generates a quiz - multiple choice questions pattern, which makes it objective, by rigorous training, the project can be used to generate subjective questions and preferred answers to that auto-generated question. Apart from English, other prominent language interpretations can be a valuable addition, as it will help to understand vast data.

XI. Conclusion

In this project we have combined Knowledge Representation and Natural Language Processing. The Auto Question Generation methods are used to develop a quiz. The other notable algorithms used include TF-IDF (Term Frequency Inverse Document Frequency), Word2Vec, and machine learning modules. The key concept of the project is to convert the text format into a questionnaire. The interface of the project is kept simple to comprehend and very straightforward. The project can be further modified and can be implemented as an educational tool.

References

[1] POS Tagging Approaches: A Comparison, Deepika Kumawat, Vinesh Jain, 2015
[2] Named Entity Recognition, Natural Language Processing of Semitic Languages (pp.221-245), 2014
[3] Automatic question generation based on sentence structure analysis using a machine learning approach, Miroslav Blšták, Viera Rozinajová, 2022
[4] Automatic Question Generation A syntactical Approach, Husam Deeb Abdullah Deeb Ali, 2012
[5] Preprocessing Techniques for Text Mining - Vairaprakash Gurusamy, Subbu Kannan - October 2014
[6] Using Word2Vec to process big text data - Long Ma, Yanqing Zhang, 2015 IEEE International
[7] Improvement and Implementation of Feature Weighting Algorithm TF-IDF in Text Classification,- Dai, W., International Conference on Network, Communication, Computer Engineering (NCCE 2018), vol. 147
[8] A Systematic Review of Automatic Question Generation for Educational Purposes, Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattle, International Journal of Artificial Intelligence in Education 30(3), Nov 2019
[9] Question-Answering Using Semantic Relation Triples, Kenneth C. Litkowski CL Research