Survey paper on text segmentation with feature similarity for exam assessment using Machine Learning.

Aditya Singh [1], Pranav Kapse[2], Rahul Nagpal[3], Shivam Pandey[4], Ms. P. V. Ambekar[5]

Department of Computer Engineering, Savitribai Phule Pune University

[1]adisingh071997@gmail.com
[2]pkapse173@gmail.com
[3]rahul11310@gmail.com
[4]shivamail12pcm@gmail.com
[5]pvambekar_sits@sinhgad.edu

***Abstract*- The need of green computing in order to reduce the excess use of paper to assess the theoretical answer is a serious demand. We therefore intend to provide a solution by building a model which helps in evaluating the theoretical answers online to reduce the human efforts. The paper involves the use of machine learning, NLP, keyword extraction and matching aggregation for checking the similarity between the user answer and the specimen answer. The user written answer is tokenized into bag of words and the meaning of words are extracted and matched with the specimen answer for semantic analysis. The machine learning algorithm analyses the answer and gives the percentage of similarity between the two answers with this system we can automatically evaluate the theoretical answers easily and efficiently, thus reducing the use of paper**.

Keywords: **Machine learning, Natural Language Processing, Keyword extraction, TF-IDF, Semantic analysis.**

I. INTRODUCTION

The aim of the survey is to study about developing an online system for evaluating the theoretical answers. The set of question and answers are stored in the database with which the answer written by the candidate is matched based on its semantic analysis. The user answer is tokenized into keywords and their meaning are extracted which is further evaluated through the machine learning algorithm to check the similarity between the original answer and the candidate answer. The answer is assessed and the percentage similarity is given as an output.

II. METHODS

A. "*Knowledge Based Question Answering(KBQA)*".

Yunshi Lan et al., [1] proposed the use of a "matching-aggregation" framework to match candidate answers with questions. The method that proposed by Yunshi is able to attain state-of-art performance on datasets .Yunshi working on two datasets that are web questions and simple questions. This paper also overcomes limitation of existing neural network-based method for knowledge based question answers(KBQA).

Jaylalita Vishwkarma et al., [2] proposed a framework for restricted domain question Answering System using advanced NLP tools and software and that framework can be used to develop a Question Answering System for extracting exact and precise answer from restricted domain textual data set. question and answering system can be classified into three category are open domain, closed domain and restricted domain. Proposed system work on keyword and question matching and return precise answer of question. In this paper Jaylalita Vishwkarma worked on restricted domain question answering system. The proposed framework not only provides a simple and implementable framework for developing question answering System but also provides a proper flow of data for answer extraction.

Zhang Kunpeng et al., [3] has proposed the idea of NLP technology to promote the development of AI and many system to make people's work easier. The author also proposes the work of question-answer system described in thesis which uses NLP technology and information retrieval technology. As this system is based on text retrieval it is completely different from traditional search engine. The question answer system allows the user to input the question

in the form of natural language and accordingly the system can get a short and accurate answer to the user. According to the Zhang the research is still a follow up study with creative research and innovative ideas, he says the question answering system at present cannot answer as well as human beings and also comment that in future the question answer system will probably replace the search engine and help the people to retrieve the information efficiently. The topic-based model for combining textual data extracted from online discussion forums to other external source which helps to identify the strength and weakness of student and help to create profile based on the similarity.

B. **"***Keyword Extraction***".**

Victor Rolim et al., [4] proposed the adoption of natural language processing for extracting the data and given focus on keyword extraction. The approach is to extract keywords and achieve an accuracy and excluding logic. The author proposed a combination of external resources and keyword identification to improve the quality.

Nebjosa D. Gruji et al., [5] has proposed potential of natural language processing to predict words based on their associative relations on a Serbian language dataset. The author's approach investigated is to use a number of different written materials, and see if a neural network can extract word associations just by reading regular texts. According to author the results are evaluated using different data sizes and preparation techniques and gives 70% of accuracy. According to the author the most important factor is the size of the data so is the improvement factor as it enlarges the contextual richness of learning set. The research shows that by eliminating most frequent words the results can be improved.

Shweta Ganiger et al., [6] has proposed that the automatic text summarization is most important area in text mining as there are many techniques for text mining summarization. There are two type of summarization techniques extractive and abstractive text summarization. The main aim of text summarization is to obtain the meaningful text from the original text document. Keywords plays an important role in building a text summarization, there are many keywords extractions algorithm. In this paper author implemented the most popular keyword extraction algorithms like TF-IDF, TextRank and Rake algorithm. In this paper we have studied keyword extraction algorithms for a single document. The three keyword extraction algorithms were implemented and compared. To test the efficiency of these algorithms we are testing on multiple documents. Proposed emerging technique Natural language processing in today's era and how it is useful in establishing machine which is capable of translating between linguistic pair. Diellza give two classifier 'Rule-based' or parts of Speech (POS) which helps in identifying feature of language from large text collection.

C. "*Based on Natural Language Processing (NLP)*".

Diellza Nagavci Mati et al., [7] also explained the graph based label Propagation for projecting POS across different languages i.e. for that also which do not have annotated data. Bensik explained that how raw text can be used to generate spellcheck dictionaries and Biemann proposed the Chinese-whispers algorithm to find rare used words. How NLP has prospective in increasing the benefits of BPM practices at different levels.

Josep Carmona et al., [8] provides NLP techniques that facilitate the automation of particular tasks. Also this paper overcomes the previous limitation that provides open-source BPM datasets to both academia and industrial application. NLP based BPM method to automatically synchronize and transform different business process representations with less time and high efficiency.

Xue han tells et al., [9] that how NLP process works for semantic analysis. NLP is the ability of a computer program to understand human language as it spoken the semantic analysis of natural language content starts by reading all of word in content to capture real meaning of any text it identifies text elements and assigns them to their logical and grammatical. In this paper Xue explained NLP pipeline method to increase performance of dependency parsing.

D. "*Based on Information Retrieval"*

Oliver Clark et al., [10] explained how the duplication question on stack overflow benefit the software development community. Oliver analysed the duplicate question from two perspective, first we analysed the experience of the user who post the duplicate question and second comparing the contents of duplicates to determine the degree of similarity. Oliver followed the data filtration, data extraction and tokenization of text approach for the identification of duplicate, which is very useful and useable in this project. Oliver also explained some future work like developing more precise technique for similarity and another technique for sentiment analysis to grapple and delivering quantative measure of

duplicates. A new model that will allow users to easily retrieve information from CSV files by natural language, a language that users are familiar with and use in everyday life. Users can specify conditions for data retrieval and processing to create the information they need. This will help non-technician users easily retrieve information without the need to learn any additional computer languages or programs.

Chalermpol Tapsai et al., [11] proposed that wrong position of words in the sentence will cause the wrong meaning. Evaluation of the model is performed by 98 testers. By inputting 1,137 natural language statements to the model, the results showed that the models were effective in retrieving and processing data accurately with very high values of precision, recall, and F-score which were all higher than 0.9. There are only 18 statements or 3.2% of all statements that produce errors in the outputs.

Reshma E U et al., [12] attempts to introduce the basic concept of NLIDBs and different architectures of NLIDBs. By using natural language to retrieve information from database is an easier way. The natural language interface is capable for translating the natural language query given by the user into an equivalent one in database query language. The computers can't understand the natural language so they need an interface that is the reason for developing a natural language interface to database. Thus natural language interfaces to databases (NLIDB) were developed for converting natural language to SQL query and get the corresponding result from the database. When a user give a natural language query to NLIDB then first it automatically understands the natural language not just semantically but syntactically too and then convert the intermediate natural language into a query that is accepted by the database management System and produce corresponding result from the database.

E. "*Based on TF-IDF Algorithm".*

Zhang Chi et al., [13] explained a news keyword extraction algorithm that combines the TextRank and TF-IDF algorithm, and enhances the weight of headlines. . This paper takes English news text as the research object of keyword extraction method. Zhang Chi combined TF-IDF and the TextRank algorithm to extract keywords from text by constructing word graph model, counting word frequency and inverse document frequency, and considering the weight of the positioning of headlines. The experimental results show that the combination of TextRank and TF-IDF algorithm and the selection of appropriate title weight can effectively improve the efficiency of keyword extraction .The Outcomes shows that the integration of TF-IDF and the TextRank algorithm significantly outperforms the traditional algorithm in performance parameters and extraction effect.

Caizhi Liu et al., [14] improves the TF-IDF algorithm, and introduces the weighting factor E(t), which reflects the degree of inter-class dispersion, the degree of intra-class dispersion, and the degree of association between feature words and categories. According to Caizhi Liu this paper presents a vector representation of feature words based on the deep learning tool Word2vec, and the weight of the feature words is calculated by the improved TF-IDF algorithm. experimental results show that the improved TF-IDF algorithm has a higher classification accuracy compared with the traditional TF-IDF algorithm.

III. CONCLUSION

In this paper, we have studied the use of Natural Language Processing technology in the development of Artificial Intelligence and also studied a model that will allow users to easily retrieve information from CSV files by natural language. Keywords plays an important role in building a text summarization, there are many keywords extractions algorithm. In this paper author implemented the most popular keyword extraction algorithms like TF-IDF, TextRank and Rake algorithm. The question answer system allows the user to input the question in the form of natural language and accordingly the system can get a short and accurate answer to the user. We have also studied the similarity of text by extracting keywords and tokenizing it in order to reduce the duplicacy in the sentence.

IV. REFERENCES

[1] Yunshi Lan, Shuohang Wang, and Jing Jiang 2019"Knowledge Base Question Answering With a Matching-Aggregation Model and Question-Specific Contextual Relations".

[2] Jaylalita Vishwakarma , Prof. Mayank Bhatt 2017""Implementation of Question and Answering Retrieval System in Natural Language Processing".

[3] Zhang Kunpeng 2019"Research on the Optimizing Method of Question Answering System in Natural Language Processing".

[4] Vitor Rolim, Maverick Ferreira, Anderson PinheiroCavalcanti 2019" Identifying students' weaknesses and strengths based on on-line discussion using topic modeling".

[5] Nebojsa D. Gruji , Vladimir M. Milovanovi 2019“Natural Language Processing for Associative Word Predictions”.
[6] Shweta Ganiger , K. M .M .Rajashekharaiah 2018“Comparative Study on Keyword Extraction Algorithms for Single Extractive Document”.
[7] Diellza Nagavci Mati , JauminAjdari , BujarRaufi ,Mentor Hamiti , BesnikSelimi 2019“A Systematic Mapping Study of Language Features Idesntification from Large Text Collection”.
[8] Han van der Aa, Henrik Leopold, Jan Mendling, Josep Carmona 2018“Challenges and Opportunities of Applying Natural Language Processing in Business Process Management”.
[9] Xue Han, Yabin Dang, Lijun Mei, Yanfei Wang, Shaochun Li, Xin Zhou 2019“A Novel Part of Speech Tagging Framework for NLP based Business Process Management”.
[10] Durham Abric, Oliver E. Clark, Matthew Caminiti, KeheliyaGallaba, and Shane McIntosh 2019“Can Duplicate Questions on Stack Overow Benet the Software Development Community?”.
[11] Chalermpol Tapsai 2018“Information Processing and Retrieval from CSV File by Natural Language”.
[12] Reshma E U , Remya P C 2017“A Review Of Different Approaches In Natural Language Interfaces To Databases”.
[13] Lu Yao, Zhang Pengzhou, Zhang Chi 2019“Research on News Keyword Extraction Technology Based on TF-IDF and TextRank”.
[14] Cai-zhi Liu, Yan-xiu Sheng , Zhi-qiang Wei and Yong-Quan Yang 2018“Research of Text Classification Based on Improved TF-IDF Algorithm”.