International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 2752-2756

Text Segmentation with Feature Similarity for Exam Assessment using Machine Learning

Deepak Panchal, Rajesh Kumar Singh, Avtar Shishodia, Anuj Panwar, Devansh Garg
Meerut Institute of Engineering And Technology, Meerut
deepak.panchal.cs.2016@miet.ac.in
rajesh.singh@miet.ac.in
avtar.shishodia.cs.2016@miet.ac.in
anuj.pawar.cs.2016@miet.ac.in
devansh.garg.cs.2016@miet.acin

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.

A. 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.

B. METHODS

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

Yunshi Lan gave the usage of framework known as "Matching-Aggregation" which is used to match the answers of candidates with given questions. State-of-art performance can be successfully attained on datasets by the method that was proposed by Yunshi Lan .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 Vishw karma come up with a framework which worked for restricted domain x question Answering System, it used advanced NLP tools and softwares and a Question Answering System can be developed by using this framework for extracting precise and accurate answer from a restricted domain textual data set. The classification of question and

answering system can be viewed into three categories, that are open domain, closed domain and last restricted domain. To return accurate answer of a given question the proposed system work on keyword and question matching. In this paper Jaylalita Vishwakarma worked on restricted domain question answering system. This projected framework not solely provides an easy and implementable framework for the event of question responsive system except for answer extraction it conjointly provides correct flow of knowledge.

Zhang Kunpeng has proposed the idea of NLP technology to promote the development of AI and many system to make people's work easier. As described in thesis, the author also proposed the work of question-answer system in which NLP technology and information retrieval technology is used. The proposed system is a text retrieval based system which makes it completely different from the traditional search engine. The question answer system allows the user to input the question in the form of natural language and accordingly the system gets a short and precise answer to the user. According to the Zhang still the research is 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.

2. ***"Keyword Extraction".***

Victor Rolim recommended the idea of using NLP tools for extraction of data and given focus on keyword extraction. The approach is to extract keywords and achieve an accuracy and excluding logic. To improve the quality a mixture of exterior resources and keyword naming is recommended by the writer. Nebjosa D. Gruji has suggested that NLP has the ability to foretell words hinge on their related similarity on a Serbian language dataset. The writer's path is to explore the use a variety of written materials just by reading the usual content, and to look if a neural network can abstract association of words. The writer concede that the accuracy of 70% is obtained as a results when evaluation is done by using variety of arrangement of techniques and different data amount. 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 recommended that there are many approaches for the text mining characterization, and the most important area in text mining is the automatic text summarization. Abstractive and extractive text summarization are the two types of the summarization approaches. To achieve the meaningful words from the actual text file is the main target of text summarization. Keywords plays a crucial part in the construction of text summarization, the keywords extractions algorithm are of several types. TF-IDF, Text Rank and Rake algorithm are the most prominent keyword extraction algorithms that has been carried out by the writer in this paper. The keyword extraction algorithms have been examined in this paper for a single record. The implementation and comparing of the three keyword extraction algorithms were completed. Multiple records are being tested for the analysis of the efficiency of these algorithms. 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.

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

Diellza Nagavci Mati 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 usage and advantage of BPM practices at distinct levels.

Josep Carmona provides NLP techniques that makes easy the automation of certain tasks. Also this paper overcomes the previous limitation that provides open-source BPM datasets to application of both academia and industries. Automatic synchronization and transformation of different business process representations is done in NLP based BPM method with less time and high efficiency.

Xue han tells that how NLP process works for semantic analysis. NLP is concerned with the interaction among computers and human languages. The reading of all the words is the starting of semantic analysis of natural language followed by the identification and assignation of text elements logically and grammatically to capture the real meaning of any text is done. In this paper Xue explained NLP pipeline method to increase performance of dependency parsing.

4. ***"Based on Information Retrieval"***

Oliver Clark has described 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 quantitative measure of duplicates. A new model that will permit the users to easily get information from the CSV files with the help of natural language, it is a language that users are friendly with and use in every day to day life. The needed information can be created by users just by giving some conditions for data retrieval and data processing. With the help of this, even without the need to learn any additional computer languages or programs, the non-technician users can easily recover information.
Chalermpol Tapsai suggested that incorrect location of words in the sentence will cause the illogical meaning. 98 testers have performed the evaluation of this model. By inserting 1,137 natural language statements to the model, the results showed that the models were constructive in retrieving and processing data accurately with very high values of precision, recall, and F-score which were all higher than 0.9. The error in the outputs are produced by only 3.2% of all statements or18 statements.
Reshma E U ventured to brought in the basic concept of Natural Language Interfaces to Databases (NLIDB) and different frameworks of NLIDBs. Retrieving information from database by using natural language is an easier way. The interface capable for translating the natural language query given by the user into an equivalent one in database query language is

the natural language interface. The reason for developing a natural language interface to database is that the computers can't understand the natural language so they need an interface. Therefore NLIDBs were made to convert natural language to SQL query and to get the corresponding result from the database. When a natural language query is given by the user is sent to NLIDB then it will first automatically understands the natural language both syntactically and semantically and then the intermediate natural language is converted into a query that the database management System accepts and produce result from the database to produce results accordingly.

5. "Based on TF-IDF Algorithm".

Zhang Chi elucidated a way that enhances the weight of headlines by explaining a new keyword extraction algorithm which is a combination of the TextRank and TF-IDF algorithm. For the investigation object English news text has been taken from
this record for keyword extraction method.

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Zhang Chi started counting word frequency and inverse document frequency, by constructing a word graph model for extracting keyword from the text and considering the load of the positioning of headlines by connecting TF-IDF and the Text Rank algorithm. The performance of keyword extraction can be increased effectively combining the TF-IDF and Text Rank algorithm and also by picking the correct title weight presented by the outcomes of the experiment. The Outcomes shows that in performance criterion and extraction event the common algorithm significantly fall behind the integration of the Text Rank and the TF-IDF algorithm.

Caizhi Liu proposed the weighting factor E(t), which reflects the rate of inter-class diffusion, the rate of intra-class diffusion, and the rate of association between feature words and categories by upgrading the TF-IDF algorithm. According to Caizhi Liu based on the deep learning tool Word2veca, a vector depiction of feature words is presented in this paper, and with the help of upgraded TF-IDF algorithm the calculation of the weightage of the feature words is done. The upgraded TF-IDF algorithm has a surpassing classification accuracy compared with the old TF-IDF algorithm demonstrated by the experimental outcomes.

FUTURE SCOPE

The upcoming work on which we are targeting is to calculate the working tendency of the application, so if any problem occurs, we can resolve it as soon as possible and show that our proposed system is better than previous applications. Also, we are planning to add features to the system including an additional set of dictionary which could help in finding the possible statements which have the same meaning as the answer more precisely. The wider the vocabulary the better the system works. It would also make the grading system more accurate and grading would be more exact including decimals.

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 facilitates easy retrieval of information from CSV files to the people with the help of natural language . In making of text summarization keywords plays a vital role, there are many keywords extractions algorithm. The author in this paper applied the most general keyword extraction algorithm i.e. TF-IDF, Text Rank 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 brief and precise answer to the user. We have also studied the similarity of text by extracting keywords and tokenizing it in order to reduce the duplicity in the sentence.

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".