A Semantic Machine Learning approach to Automatic PPT generation

Krutika Ramchandra Phage[1] Supriya Sunil Rawade[2] Kajal Shamrao Thorat[3] Rohini V Agawane[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]KJCOEMR, Pune, India

Abstract— In today's world as the size and complexities of the documents are increasing it becomes very difficult to extract the important points from the multiple documents in a given stipulated time. Natural language Processing (NLP) is playing a vital role in this, where it extracts the important feature from each of the sentences and they are framed together to yield best semantic data. This research article mainly concentrates on yielding the semantic presentation slides for the given input of multiple documents of type pdf, text and MSword. Some methodologies are existed to concentrate on yielding automatic PowerPoint presentation files based on the given input of documents. Most of the research technique do suffer from the matter of precision of the obtained slides. So this research article throws some light on automatic power point presentation generation technique by extracting some NLP features and then by using Random forest technique model provides accurate desired slides.

Key words: NLP, Feature Extraction, Random Forest, PPT Generation

I. INTRODUCTION

PowerPoint was developed by a start-up in 1983 for the Macintosh Platform as an addition to the Graphical user interface. It was an immense success, such that it has the majority of the worlds market share for a presentation software.

The reason of PowerPoint popularity is attributed to the fact that it is very easy to present as a very powerful visual aid that can have a captivating impact. This can leave a lasting impression on the viewer and conveys the information very efficiently. Due to this fact, the Automatic Generation of a PowerPoint Presentation Generation is a valuable resource as in this fast-paced lifestyle, it is not possible to create a High-Quality presentation in a less time. Therefore, automatic PowerPoint Presentation Generation is very helpful for businesses and academicians to help get their point across efficiently and with maximum impact.

Feature extraction refers to the practice of isolating certain data elements from a collection with similar features. It is quite an essential aspect of Machine Learning that is to segregate useful data and useless data. this allows the Data Scientists to work a lot more efficiently with only relevant data that is being isolated. The features here refer to the important aspects related to the data that is being segregated. These points are the features that are essential for the particular problem at hand.

Feature extraction is the most basic and the core part of a machine learning interface that deals with a lot of data and has to understand the relevance of the data elements to solve the particular problem at hand. The algorithm is basically trained with the help of relevant data before actually applying the feature extraction on the real data. This helps the algorithm understand which features are essential for the application and which features are not useful at all for the application.

The data that is usually used in machine learning applications is large and is almost impossible to select the relevant data items that are required for that particular machine learning problem manually. As doing it manually would take very large man hours and would be extremely inefficient and uneconomical. This is why feature extraction is clearly one of the most essential parts of a Machine Learning Algorithm.

Random forest is one of the most powerful and widely used concepts for the purpose of prediction. It is one of the leading algorithms that are a part of the machine learning paradigm. Random forest is based on the concept of decision trees, which are basically a tree that is branched with all the possible outcomes or answers for a specific query. Random Forest is a Collection of these Decision trees together, hence the name, Random Forest. Random forests are quite versatile in their applications and can be used for implementing regression as well as classification problems.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. LITERATURE SURVEY

A.Alzand [1] Arabic language is one of the special and the unique language for pronunciation and also in writing. When two different letters of Arabic get together will change and change the meaning of the words. This paper proposed an alternative tool which translates the Arabic word to English word by using the formation of Arabic letters. In this system, Arabic words are given as input and then Arabic Natural Language Processing converts the Arabic words in English words the outputs shows the percentage of successfully translated words.

Li Zhao [2] Natural language is used to represent human action. A business rule is very hardto understand and describe them into the natural language.Thus in this paper, they have proposed the statistical machine learning method to understand the natural language later it should be converted into a language which machine can understand. Unified Modeling Language and Object Constraint Language are one of the powerful languages to understand the overview of the business modeling process.

Bill.Manaris [3] Generally the natural Language processing tools havebeen developed form the last four decades.Currently, this system is dependent on linguistic theories, cognitive models, and engineering approaches. Unrestricted NLP is very hard to understand and some domainis available for the restricted domain. Thus domain usually depends on operating system interface languages and on the database. Some of the NLP tools are Machine Translation Systems,Natural Language Interfaces,and User Interface Management Systems, Text Processing Understanding Systems.

All rights reserved by www.ijsrd.com

E1Torunski[4] As there is tremendous growth in the multimedia sector this multimedia sector is moved towards the education system. In this paper, there is research made on the automaticgenerationofcontent navigation. This content navigation is based on the generation of multimedia like XML, PPT,and extraction of the algorithm. By using this we can reduce the man work of developing the multimedia. There has been a lotof research made on this topic but there is a lot of improvement to be made further.

P. Rani [5] By using the Artificial Intelligence and NLP process their effective applications are developed. Nowadays everything is developed smarter and smarter thus the home appliance is also progressing. There are much software working in the market for a solution and less workload for the people. This paper presents by using the voice modulation system we can automatically switch off or switch on the appliance. The user sends the command by speech to mobile then mobile will send the command to the appropriate devices.

C.Xiaoduo [6] In recent years the bilingual teaching in increasing and developing in Chinese Higher Education.Bilingual education is nothing but involving two languages in teaching one is the native language and the other one is the second language. To implement the bilingual language there is an important tool in CAI. PPT should be used in the Bilingual language to understand in an easy way. In this paper, there is an advantage to implement the this in bilingual teaching so this project should be insisted on.

Reshma. EU [7] Nowadays the usage of the database growing day by day. In industries and in school the relational database is used on a large scale. The people who as the knowledge of SQL the can extract the data fromthe relational database but the people who as no knowledge can't retrieve the data from the database. For these people, this project is solved by using NLP. For this Natural Language interface is used to convert the natural language in SQL.The language which is used are Malayalam, English, Hindi, Tamil etc. instead of the SQL query

S.Weigelt [8] There is a vast evaluation begun spoken the natural language there is ProNat tool which used the script –like programming. There are some devices is such as Apple Siri,google voice search and many more.A pronate tool as made the NLU processing very easily and it is one of the independent domain. The main goal of this paper is to extract the control structure and processes to be mapped on programming contracts.

A.BARNARD [9] As there is emergence in artificial intelligence such as robotic and it is increasing in nursing and health environments. There are also many issues in for natural processing language. One of the major problems in nursing is the communication problem and these challenges of communication are highly uncertain in the nursing practice environment. The nursing should be understood in term important on holism and further on physical, mental, social, and psychological of communicating with people.

A.Hermawan [10] In machine learning process the grammar induction is one of the learning processes of grammar corpora. To parse the sentence there is one of the function i.e. fitness function is used to count sentence which is parsed. There have introduced this grammar induction in an Indonesian language using a genetic algorithm. The result matches with the corpora grammar but the structure is developed manually. The accuracy rate of this paper can be improved so that the result may be better.

L.Lin [11] Usage of the Internet is growing day by day internet.The Internet is a source where an answerto every question is given. Some time the sensitive information is leaked from the internet or the fake information is leaked from the sites which can create a bad effect on society. So to protect this information the new method is used in this paper by combining natural language processing with data mining technology. Just to hamper this browsing information, filtering it is necessary. Thus the NLP technique is useful for monitoring the sensitive web information.

P.Gupta [12] Database is one of the important parts of the business and nowadays most of the small scale and as well as mid-sized businesses depend on the database. If any inexperienced persontries to access the data they should have the knowledge of DBMS languages. Thus the proposed paper introduces new technology Intelligent Querying System where the inexperienced person will ask query in his own natural language this system will parse this query understand this query and generate the required output as the user want.

N.Srinivasan[13] As there is massive growth in digital science most of the youngsters are moving towards the digital world but there are some youngsters are showing some interest and moving towards agriculture farming. Some youngsters and people had owned land for farming but there in need of proper guidance in the farming sector. The traditional method of framing should be transferred tothe upcoming generation. The framer those are using a smartphone can share their experience under the SME i.e. Subject Matter Expertise and then it can be shared with the youth by using NLP in form of mobile application.

III. PROPOSED METHODOLOGY

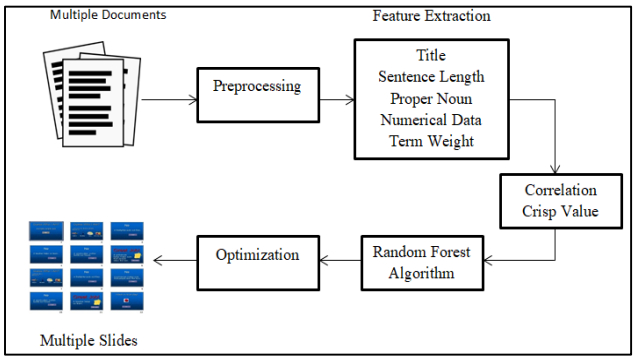

Figure 1: System Overview

The proposed methodology of automatic PPT generation is depicted using Fig 1 for multiple documents. And this process of summarization is explained in details with below-mentioned steps:

Step1: Data Collection - This is the first step of the proposed model wherea folder is been given as an input. Once this folder is accepted by the system then the system is read all the pdf, txt and doc files. The pdf and doc file are read using itext and apache poi API respectively, whereas text files don't need any kind of external API.

All the contents of pdf, text and doc files are concatenated into a single string to feed this to the next step of preprocessing.

All rights reserved by www.ijsrd.com

Step2: Preprocessing - This is the step where the redundant data is being shredded off to make the data light weight so that complexity can be reduced. So to achieve this proposed model uses four different steps as explained below: *Tokenization*: This is a process where each of the sentences of concatenated string from the previous step is split on regex string ". ". This yields a list of all the sentences of all the documents thatare fed to the system. Each of these sentences are handled for the process of stemming and the stopword removing by splitting the sentence into individual words.

Special Symbol Removing: Here all the special symbol like ! @ # ,] $.-_are identified and nullified using the empty character.

Stopword Removing: Stop words are the conjunction words like and, of, the, are, is which hardly plays any important role in the semantics of the sentence. So shredding of these stopwords simply makes the data lightweight less complex.

Stemming: Here in this step a word is bring back to its base form by replacing its postfix content with the meaningful phrases. For example *going* becomes *go*, *Writing* becomes *Write* .

Step 3 : Feature Extraction: This step uses the processed data from the last section to extract the important features which plays a vital role in finding the facts of the text which is explained below.

*Title Feature :*This feature is very important that shows the frequency of availability of the words of first sentence. The title feature for the each sentences can be calculated in the below show equation

$$T_F = \frac{\text{No. of simimlar words b/w title sentence and sentence}}{\text{The toal of number words in title sentence}}$$

*Sentence Length :*This feature actually represents the sematic sentence which eventually contains the more narrative words in it. This means any sentence which is having higher words means they are having more narrative property. So this feature is very important and this can be estimated with the below mentioned equation.

$$S_L = \frac{\text{Total number of words in a sentence}}{\text{Number of words in a longest sentence}}$$

*Proper Noun :*This feature plays an important role in identifying the scenario based power point slide generation. As we know that proper nouns are always revealing the names of some person or place. So to evaluate this proposed model uses the oxford Dictionary which contains around 1,50,000 words for all the letters of the English Language.

So this step compares each of the words with the respective bags to get the count of the proper noun. And this can be evalauted using the following equation.

$$P_N = \frac{\text{Total number of Proper Noun in a sentence}}{\text{Length of sentence}}$$

Numerical Data : As we know that numerical data is always revealing the statistical information about the document so they are very important. And these data can be evaluated using the following equation.

$$N_D = \frac{\text{Total number of numeriacal data in sentence}}{\text{Length of sentence}}$$

Term Weight : Term weight is a peculiar feature which represents the most repeated words in a document. To estimate this fist all preprocessed words are added in a list to remove all duplicates. And then for each of these unique words, their frequencies are being estimated to store in a double dimension list.

Then this list is sorted in descending order to get the top 10 repeated words in the string. And they are called as frequent set, Term weight is evaluated using the following equation.

$$T_W = \frac{\text{Total number of matched words with frequent Set}}{\text{Length of sentence}}$$

Step 4:Crisp values and Random Forest classification - Once all the features from above mentioned steps are estimated , then they are framed in a double dimension list with the normalized values after the decimal point to call them the crisp list. This crisp list is subjected to random forest classification model, Where the trees and subtrees are formed based on the evolved factors of the features.

These feature values are arranged in the tree to traverse them in pre-order traversal technique to classify the sentences based on the feature values of all the category. This can be shown in the below mentioned algorithm 1.

Algorithm 1: Tree Formation
// Input : Crisp Value Set C_{SET}
// Output : Tree T
Function : treeFormation(C_{SET}*)*
Step 0: Start
Step 1: T=Ø
Step 2: $RN_{SET} \rightarrow$ *CRISP value*
[RN_{SET} *: Root Node]*
Step 3: for i=0 to size of C_{SET}
Step 4: Temp $_{SET} \rightarrow$ $C_{SET\,[i]}$
Step 5: CRISP Value $C_V \rightarrow Temp_{SET[1]}$
Step 6: IF $C_V < R_N$
Step 7: CREATE NODE N
Step 8: ADD N as LEFT CHILD of T
Step 9: ELSE
Step 10: ADD N as RIGHT CHILD of T
Step 11: End for
Step 12: return T
Step 13: Stop

Once this tree is created, then it is traversed in preorder to accumulate the features lied on the levels of the tree to form clusters. Once these clusters of the sentences are formed, then according to the requirement of the slides of the power point presentation. These sentences are categorized under titles like Abstract, introduction, motivation, objective, summary and keywords. These clusters of sentences are finally used to create the powerpoint presentation using the Apache POI API.

IV. RESULT AND DISCUSSION

The proposed model of automatic Power point presentation creation system using machine learning is deployed using Java technologies. For this very purpose proposed model uses windows based machine with Core i5 as the intelligence and 6GB of primary memory. The Model uses the Netbeans as the IDE and MySQL as the database storage Server.

The developed system takes mainly three formats of the input like pdf, text and word files. The model takes a

All rights reserved by www.ijsrd.com

folder containing the mixture of these files. And yields the Powerpoint presentation of slides containing some indices like Abstract, Motivation, Problem Statement, Objectives, introduction , Summary and Key words.

The Quality of the Powerpoint presentations can be perfectly measured by the users only. To do this proposed model uses Mean Reciprocal Ratio (MRR) as the measuring parameter. Where a rank is being given by the end user for the obtained power point slides and these ranks are called as the Reciprocal ratios (RR).

The RR is called as Reciprocal Rank , the values of RR are 1,1/2,1/3,1/4,1/5,0. For example If a sentence is fit on its correct position and the user is highly satisfied with that then that sentence may get a rank of 1. On the other hand, if a correct sentence appears on the second rank, then it is one over two, so the score will be 0.5, etc. Like this sentences are assigned ranks for all the top five levels and then for all these ranks a inverse value is being estimated as the model of MRR. If none of the top five responses contained a correct rank, then the score is considered as zero.

$$MRR = \frac{\sum_{i=1}^{N} 1/Rank_i}{N}$$

The mean reciprocal rank (MRR) is the average score over all assigned ranks.

Where,

N - Number of Trails

i- Trail number

Rank - Rank provided by the user

On conducting of experiment on MRR for set of 10 Trails each of the experiment set produces the result as recorded in the table 1 and figure 2.

The plot in figure clearly indicates that the average MRR of the proposed model is about 0.8086 for the automatic PPT generation model using machine learning technique. The obtained MRR is pretty high and it is a good sign of the system as accuracy is concerned.

Experiment No	Slide Heading	MRR
1	Abstract	0.733
2	Introduction	0.88
3	Keywords	0.94
4	Summary	0.8
5	Objective	0.69

Table 1:MRR for Different experiments

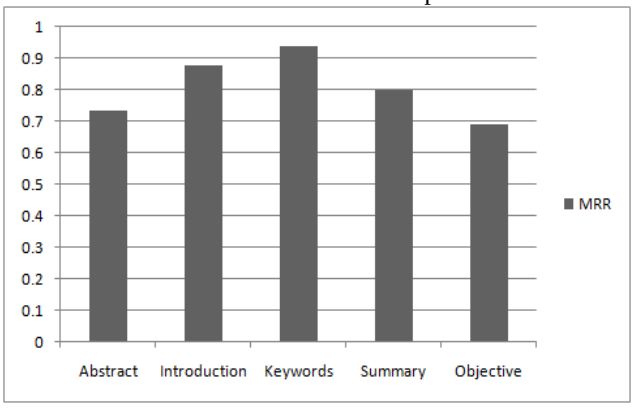

Fig. 2: MRR for Different heading of the slides

V. CONCLUSION

The proposed model of automatic Powerpoint presentation creation is enriched with the natural language processing and Feature extraction. The proposed model takes input of three formats of files like pdf, text and word files and generates the Powerpoint presentation with the indices like Abstract, Objective, Introduction, Motivation , summary and keywords.

The model uses the Random forest classification technique to classify the feature selected sentences. The accuracy of the system is measured using MRR and it yields a good accuracy about 0.8086.

This system can enhance to generate the Powerpoint presentations using online data through an interactive web crawler and this can be designed to work as ready-made API.

REFERENCES

[1] Abdulrahman Ahmed Alzand and Rosziati Ibrahim, "Diacritics of Arabic Natural Language Processing (ANLP) and its Quality Assessment", Proceedings of the 2015 InternationalConference on Industrial Engineering and Operations ManagementDubai, United Arab Emirates (UAE), March 3 – 5, 2015.

[2] Li Zhao and Feng Li, "Statistical Machine Learning in Natural Language Understanding: Object Constraint LanguageTranslator for Business Process", IEEE International Symposium on Knowledge Acquisition and Modeling Workshop, 2008.

[3] Bill Z. Manaris, "Natural Language Processing Tools and Environments:The Field in Perspective", Proceedings Sixth International Conference on Tools with Artificial Intelligence. TAI 1994.

[4] Wu Linjing and Liu Qingtang, "Research on Automatic Generation of MultimediaCourseware ContentNavigation", IEEE International Conference on Granular Computing, 2008.

[5] Mrs. Paul Jasmin Rani1, Jason Bakthakumar,Praveen Kumaar andSanthosh Kumar,"voice-controlled home automationSystem using natural language processing(nlp) and internet ofthings (IoT)", Third International Conference on Science Technology Engineering &Management (ICONSTEM), 2017.

[6] Xiaoduo, "Application of PowerPoint in Bilingual Teaching of Managerial Classesin Chinese Local Higher Education Institutions", International Conference on EducationTechnology and Computer, 2009.

[7] Reshma E U and Remya P C, "a review of different approaches in naturallanguage interfaces to databases", Proceedings of the International Conference on Intelligent Sustainable Systems(ICISS 2017).

[8] Sebastian Weigelt and Walter F. Tichy. "PosterProNat: An Agent-based System Design forProgramming in Spoken Natural Language", IEEE/ACM 37th IEEE International Conference on Software Engineering, 2015.

[9] Alan BARNARD, "The Nursing Profession: Implications for Al and NaturalLanguage Processing", Journal of the American Medical Informatics Association, July 2017.

All rights reserved by *www.ijsrd.com*

[10] Arya Tandy Hermawan, Gunawan and Joan Santoso, "Natural Language GrammarInduction of Indonesian LanguageCorpora Using Genetic Algorithm", International Conference on Asian Language Processing, 2011.
[11] Liu Lin, Fan Xiaozhong and Zhao Xunping, "Research on Web Monitoring System Basedon Natural Language Processing", International Conference on Natural Language Processing and Knowledge Engineering, 2003.
[12] Prashant Gupta, Aman Goswami and Kashinath Sartape, "IQS- Intelligent Querying System using NaturalLanguage Processing", International Conference on Electronics, Communication and Aerospace TechnologyICECA 2017.
[13] Dr. N. Srinivasan and Anandaraj Selvaraj, "Mobile Based Data Retrieval using RDF andNLP in an Efficient Approach", Third International Conference on Science TechnologyEngineering & Management (ICONSTEM), 2017

All rights reserved by *www.ijsrd.com*