Auto Generation of Presentation Slides From Text

Prof. Mrs Minal Nerkar[1], Sandhya Budar[2], Durgadevi Jadhav[3], Snehal Malawadkar[4], Akshata Shinde[5]
[1, 2, 3, 4, 5] Department of Computer Engineering
[1, 2, 3, 4, 5] AISSMS's Institute Of Information Technology, Pune-411001, Savitribai Phule Pune University, India

Abstract- *Presentation slides are used on a large scale in the corporate world, I.T. industries,schools and college because they are efficient and easy to understand. In this paper we have come up with a system that will automatically generate presentation slides from text. The slides will mostly comprise of only the important points related to the topic. They are a powerful means of presenting a topic to the the audience , as the important points also known as bullet points are covered in the slides and can be explained by the presenter in depth. Tools available in the market focuss on the formatting of the content , but not with the content itself. This will eventually help in reducing a great amount of the presenter's time and efforts. The proposed system works on the NLP rules to classify data for the desired slides.*

Keywords- *Crisp values, Feature Extraction, Fuzzy Classificaion, Pre-processing*

I. INTRODUCTION

Presentation slides are used to present a topic or a new concept before the audience. They can be used to address a large audience.They are especially used when a new product is to be launched in the market. The essential features of the product are highlighted by using presentation slides. For preparing slides the user has to study the topic throughly and also invest a lot of his time and efforts. This problem is taken care of by our system. Our system is efficient because it generates presentation slides automatically when text is given as input to the system. Important keyphrases and points related to the topic are extracted and displayed on the slides.

II. LITERATURE SURVEY

Yue Hu and Xiaojun Wan [1] used SVR based sentence scoring model to assign important scores to each sentence and ILP model to generate structured slides from academic papers.

R. Jha, A. Abu-Jbara, and D. Radev [2] investigate the problem of automatic generation of scientific surveys starting from keywords which are provided by the user. The system takes a topic query as input and generates a survey of the topic. It selects a set of relevant documents and then selects relevant sentences from the documents to generate the survey. Content models namely Centroid , Lexrank , C-Lexrank are used.

A. Abu-Jbara and D. Radev [3] proposed an approach which focuses on coherence and readability aspects of the problem. It produces citation-based summaries in three stages: pre-processing, extraction and post processing. These summaries are better than several summarization systems.

M. Sravanthi, C. R. Chowdary, and P. S. Kumar [4] concentrate on generating slides from research papers. Latex documents which have rich structure and semantic information are given as input to the system. The documents are initially in XML format. The XML file is first parsed and then information in it is extracted . QueSTS Summarizer,a query specific extractive summarizer is used to generate slides. All graphical elements are placed at appropriate locations in the slides.

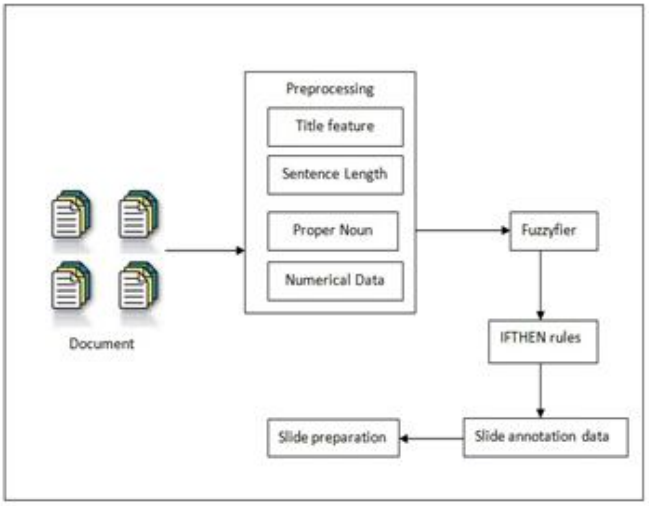
M.Y. Kan [5] present their work on SlideSeer , a customized digital library which comprises of an offline discovery, alignment and indexing system and an online web user interface. Three major system components of SlideSeer are discussed namely 1) resource discovery, 2) fine-grained alignment , 3) the user interface.

M. Utiyama and K. Hasida [6] in their paper discuss how to automatically generate slides shows. The system inputs documents which are annotated with the GDA tagset and XML tagset. These tagsets allow machines to automatically infer the semantic structure underlying the raw documents.

M. Utiyama and K. Hasida [7] in their paper discuss how to automatically generate slides shows. The system inputs documents which are annotated with the GDA tagset and XML tagset. These tagsets allow machines to automatically infer the semantic structure underlying the raw documents.

III. PROPOSED SYSTEM

Many existing systems are yielding low semantics. Proposed system works on NLP rules to classify the data for the desired slides. When text is given as input to the system it goes through the following stages.



3.1 Reading

Text is given as input to the system .It can be in pdf or doc format.The text is read in string format.

3.2. Preprocessing

The input is preprocessed by the following methods.

3.2.1. Remove Stopwords

Stopwords are those words, which when removed will not alter the desired meaning of the sentence. Hence stopwords are removed in order to increase the processing speed.

3.2.2. Stemming

In stemming process a word is brought to its base form.By doing this overhead is reduced and accuracy is increased.

3.2.3.Tokenization

In this process words are trimmed , spaces are removed , tokens are generated and are then put in an array.

3.3. Feature Extraction

In this stage important features are extracted by the following means.

3.3.1. Title Feature

Words which occur in the sentences and also in the title are called as Title words. The occurance of these words increases the score. This is determined by counting the number of matches between the words in the sentence and the words in the title. Thus a score will be generated by using the following formula:

Scoref1 (Si) = No.Title words in Si / No.Words in Title

3.3.2.Sentence Length

The length of the sentence is calculated by the number of words occuring in the sentence. Normalized length of the sentence is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document. Sentence length is calculated by using following formula.

Score f2 (Si) = No.Words occurring in Si / No.Words occurring in longest sentence

3.3.3. Term Weight

It is the number of times a particular word has occurred in a sentence.If the term weight is high then that term is considered to be important. tf_isf (Term frequency,Inverse sentence frequency) method is applied to calculate the score of the term.

Score f3 (Si) = Sum of TF-ISF in Si / Max(Sum of TF-ISF)

3.3.4. Numerical Data

Numerical data is the number of numerical data in sentence. Sentence that contains numerical data is important and it is most probably included in the document summary.

Score f4 (Si) = No. Numerical data in Si / Length (Si)

3.3.5. Proper Noun

Number of proper nouns in the sentence is calculated. Usually the sentence that contains more proper nouns is an important one and it is most probably included in the document summary.

Score f5 (Si) = No. Proper nouns in Si / Length (Si)

3.4. Fuzzy Classification

Classification is done into 5 types:
Very low-0(VL)
Low(L)
Medium(M)

High(H)
Very high-1(VH)

Generated scores of the sentences are checked according to the above classification.A score is termed as Very low if it has a score 0 and is termed as Very high if it has a score 1.Hence if the score is 0, the sentence is less important and if the score is 1 then the sentence is important. Thus importance of a sentence can be obtained.

3.5. If-Then Rules

IF (NoWordInTitle is VH) and (SentenceLength is H) and (TermFreq is VH) and (SentencePosition is H) and (SentenceSimilarity is VH) and (NoProperNoun is M) and (NoThematicWord is VH) and (NumbericalData is M) THEN (Sentence is important)

3.6. SLIDE GENERATION

PPT slides are generated at the end.

IV. ALGORITHM

4.1. Algorithm for Preprocessing

Step 0: Start
Step 1: Get contents of Query
Step 2: split in Words
Step 3: Remove Special Symbols
Step 4: Identify Stopwords
Step 5: Remove Stopwords
Step 6: Identify Stemming Substring
Step 7: Replace Substring to desire String
Step 8: Concatenate Strings
Step 9: Preprocessed String
Step 10: Stop

4.2. Algorithm to find stop words

Step 0: Start Step
Step 1: Read string
Step 2: Divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its frequency
Step 6: Add frequency in List Called L
Step 7: end of for
Step 8: return L
Step 9: stop

4.3. Algorithm to find noun

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its occurrence in Dictionary
Step 6: if present then return true
Step 7: else return false
Step 8: stop

V. ADVANTAGES

1. Well structured slides are generated.
2. Presenter's time and efforts are saved to a great extent.
3. Slides will include important keyphrases and sentences related to them.

VI. LIMITATIONS

1. Slides will contain only text.
2. Images and tables will not be included in slides.

VII. RESULTS

To show the effectiveness of the proposed system some experiments are conducted on java based windows machine using Netbeans as IDE. To measure the performance of the system we set the bench mark by considering the system with number PPT's.

For evaluation the system is required to submit a ranked list of five opinions to a generated PPT. Each PPT received a score equal to the inverse of the rank at which the first correct opinion was found. That is called the Reciprocal Rank (RR), that the values of RR are 1,1/2,1/3,1/4,1/5,0. E.g., if a correct rank appears on the second rank, then it is one over two, so the score will be 0.5, etc. If none of the top five responses contained a correct rank, then the score was zero. The mean reciprocal rank (MRR) is the average score over all generated PPT.

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

Where, Ranki is the rank of the first correct occurrence in the top five ranks for generated PPT i; N is the

number of test PPT asked for the opinion; If for a PPT i, the correct rank is not in the top five responses then it is taken to be zero.We performed an experiment to evaluate the rank retrieval using the MRR metric up to the top three responses, defined as follows. The result is shown in Table 1.As the result, the proposed method achieved average MRR of 0.608 for ranks of different generated PPT's.

Table 1. Result of the Experiment

SR No	MRR
1	0.65
2	0.5
3	0.5
4	0.7
5	0.69
	MEAN=0.608

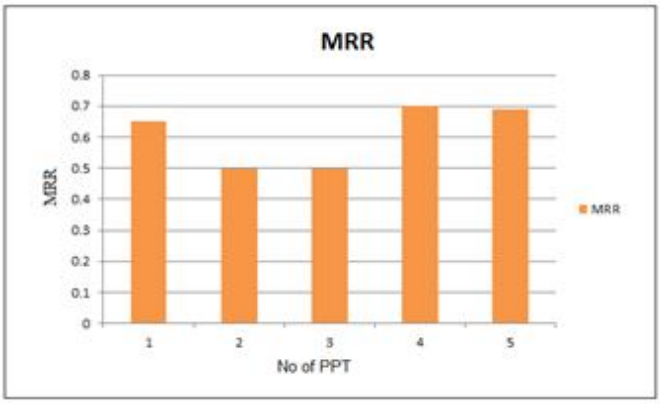


Figure 2. MRR of the Different PPT

In the Fig. 2, we observe that the tendency of average MRR is 0.608 for the generated PPT's table 1. So this is actually a better performance in our very first attempt of PPT generation using NLP protocols and Fuzzy Logic.

VIII. CONCLUSION

It is a tedious job to create presentation slides.Thus our system will save a huge amount of the user's time and efforts.Presentation slides are generated in an efficient and quicker way after using the above methods.

IX. FUTURE SCOPE

The system can be enhanced to work on all cross platforms.

REFERENCES

[1] Yue Hu and Xiaojun Wan,"PPSGen: Learning-Based Presentation Slides Generation for Academic Papers", ieee Transactions on knowledge and data engineering, vol. 27, no. 4, april 2015.

[2] R. Jha, A. Abu-Jbara, and D. Radev, "A system for summarizing scientific topics starting from keywords," ACM Comput. Surv,vol. 40, no. 3, p. 8, 2013.

[3] A. Abu-Jbara and D. Radev, "Coherent citation-based summarization of scientific papers," in Proc. 49th Annu. Meeting Assoc.Comput.Linguistics: Human Lang. Technol.-Volume 1, 2011,pp. 500–509.

[4] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "SlidesGen:Automatic generation of presentation slides for a technical paper using summarization," in Proc. 22nd Int. Flairs Conf.,2009, pp. 284–289.

[5] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "QueSTS: A query specific text summarization approach, in Proc. 21st Int.Flairs Conf., 2008, pp. 219–224.

[6] M.Y. Kan, "SlideSeer: A digital library of aligned document and presentation pairs," in Proc.7th acm/ieee-cs Joint Conf. Digit.Libraries, Jun. 2006, pp. 81–90.

[7] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centralityas salience in text summarization," J. Artif. Intell. Res., vol. 22,no. 1, pp. 457 479, 2004.

[8] A. Nenkova and R. J. Passonneau, "Evaluating content selection in summarization: The pyramid method," in hlt-naacl, vol. 4,pp. 145–152, May 2004.