IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 33-35

Project Feature Extractor

Krina Rathod[1], Rohan Parab[2], Yash Chaukekar[3], Prachi More[4], Yogita Shelar[5]
[1,2,3,4]*(Information Technology, Atharva College of Engineering/ Mumbai University, India)*
[5]*(Assistant Professor, Information Technology, Atharva College of Engineering/ Mumbai University, India)*

Abstract : *Feature Extraction is a facet of Text Mining that is gaining exponential popularity due to the information explosion on the Internet. It is imperative at this point to explore this domain so as to develop systems that make it easier to understand the vast data by summarizing it. Text Summarization can be performed by abstraction and extraction. Our proposed system makes use of the extraction approach. Summarization by extraction is the process of identifying important features of the document as well as extracting the sentences on the basis of their score or weight. In our proposed system, we consider research papers as input, perform different preprocessing techniques, then use six features to calculate the weight of each sentence to get a short summary as output.*

Keywords: *Artificial Intelligence, Data Mining, Feature Extraction, Text Summarization, Text Mining*

I. Introduction

The rampant growth in the availability and complexity of information on the internet has led to the emphasis put on the improvement and development of mechanisms and tools that can be used to find relevant information from the plethora of sources available with us. The academic curriculum these days require students to do thorough research in order to undertake any assign

ment or project. It is no wonder that students and even general users spend a considerable amount of time on this research. The proposed system aims to aid users in their endeavor to find germane material for their research.

Text Summarization is a timely tool that can facilitate interpretation of large volumes of data by providing a concise version of the data without modifying the meaning of the data. It is an important tool that saves a lot of time. In the proposed system, we attempt to build a tool that summarizes the input document by extraction. Automatic text summarization using the extraction method includes selection of the phrases or sentences that have the highest score on the predetermined set of features. Another method for summarization is abstraction in which linguistic mechanisms are used to examine the input data. One of the main challenges of text summarization is to identify the important phrases from the source data and eloquently providing a summary without changing the meaning of the source data.

In the proposed system, the features based on which the importance of the phrases or sentences is determined are sentence length, sentence location, term frequency, number of words occurring in title and number of proper nouns. The incredibility of the proposed system is that it is a rather dynamic one that can summarize any research paper the user shall want to interpret.

II. Proposed System

2.1. Input Data and Preprocessing

Research papers, scientific papers, newspaper articles or any well-structured document can be used as input for the proposed system. In a text document, the text portion of the research paper fetched from the URL is saved. The system is flexible in a way that data from any .pdf, .doc, .docx or even .txt file format can be extracted and saved as plain text.

2.2. Preprocessing

The input document provided to the system is in plain text format. Before we can begin summarizing the document, the input plain text must be processed such that it is compatible for extraction. This involves running a few algorithms to ensure that the input text is clean and ready to be extracted. This is called as preprocessing. The preprocessing phase of the proposed system comprises of three predominant processes, namely: Stop word removal, Stemming and Part of Speech Tagging (POS tagging).

2.2.1. Stop word removal

It is an important step which involves removal of insignificant and irrelevant characters such as symbols, words such as “a’, “an”, “or”, “the”, special characters such as “^”, “/”, “$”, etc. This filtration enables the system to

understand the important phrases of the document and ignore the ones that do not add value to the extraction process.

2.2.2. Stemming

Stemming is a process of boiling down the words into its root form. This is a type of filtration that helps us to obtain the base form of the word by extracting the prefix and suffix of the concerned word. For example, "studying" is stemmed down to "study" by removing the 'ing' from the word.

2.2.3. Part of Speech Tagging

POS tagging is an integral part of Natural Language Processing (NLP) and Natural Language Understanding (NLU) and it is also called as grammatical tagging that helps assign part of speech to each the word/token such as noun, adjective, etc. This step is crucial because an extraction feature out of the total five that we have employed is contingent upon the result of this process.

2.3. Implementation of Fuzzy logic:

After the preprocessing phase, text summarization plays a pivotal role. In this proposed paper, the system obeys five set of features to summarize the document in the most meaningful way. They include:

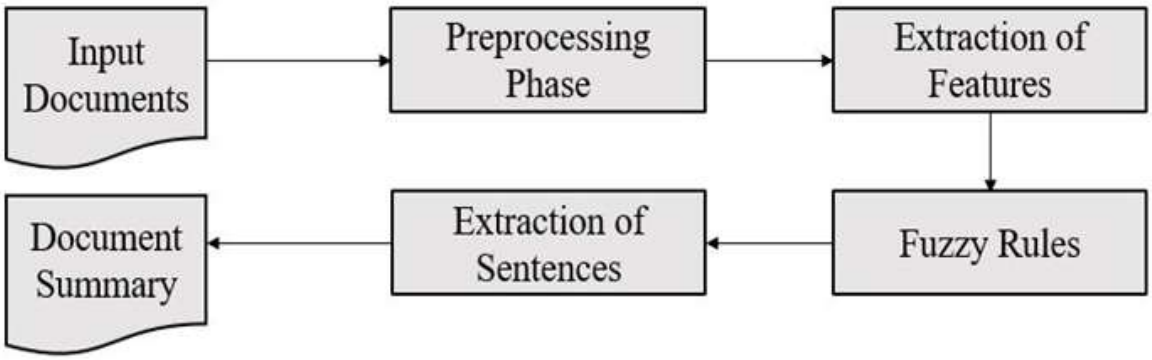

Figure 1: Block Diagram

2.3.1. Title feature

Title feature extracts the similarity between the title sentence and every other sentence in the document by counting the number of matches as shown in equation (1).

$$S\_F1(S) = \frac{\text{No of Title words in sentence S}}{\text{No of words in Title}} \quad \ldots\ldots\ldots (1)$$

2.3.2. Sentence length

Sentence length we remove the short sentences from the input document as shown in equation (2).

$$S\_F2(S) = \frac{\text{Length of sentence S}}{\text{Length of longest sentence in document}} \quad \ldots\ldots\ldots (2)$$

2.3.3. Term weight

In term weight, we calculate the frequency of a term used in the document as shown in equation (3).

$$S\_F3(S) = \sum_{i=1}^{k} (TFi) \quad \ldots\ldots\ldots (3)$$

2.3.4 Sentence position

Sentence position helps us determine the importance of the sentence. This is done on the basis of its position in the document/paragraph as shown in equation (4).

$$S_{F5(S)} = \text{1st sentence} = \frac{5}{5}; \text{2nd sentence} = \frac{4}{5}; \text{3rd sentence} = \frac{3}{5};$$
$$\text{4th sentence} = \frac{2}{5}; \text{5th sentence} = \frac{1}{5}; \text{other sentence} = \frac{0}{5}; \quad \ldots\ldots\ldots (4)$$

2.3.5 Proper noun

Proper noun feature decides the number of proper nouns in the sentence. This count further helps us to decide how likely it is to be included in the summary as shown in equation (5).

$$S\_F5(S) = \frac{\text{Number of proper Nouns in Sentence S}}{\text{Sentence Length (S)}} \quad \ldots\ldots\ldots (5)$$

III. Conclusion

In the proposed system, we have used Natural Language Processing (NLP) to successfully generate text summary of the provided document. The system was tested on 10 research papers. The proposed system reduces the time spent by the user in reading the whole paper or article. The paper or article in which there are a greater number of images are difficult to process during pre-processing phase. The proposed system can be used to summarize research papers, news articles, or any well-structured document. The accuracy of the summary can be increased by using deep learning and artificial intelligence.

Acknowledgements

We are extremely grateful to our guide Prof. for his continuous motivation and guidance. We would also like to thank the faculty of Information Technology Department who greatly assisted our research and for their valuable help.

References

[1]. Kaustubh Patil and Pavel Brazdil, "SUMGRAPH: Text Summarization Using Centrality In The Pathfinder Network",International Journal on Computer Science and Information Systems, vol.2, no.1, 2007, PP. 18-32.

[2]. Khushboo Thakkar and Urmila Shrawankar, "Test Model for Text Categorization and Text Summarization," Web Intelligence and Intelligent Agent Technology, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 4, Apr 2011.

[3]. Leonhard Hennig, Winfried Umbrath and Robert Wetzker, "An Ontology-based Approach to Text Summarization," Web Intelligence and Intelligent Agent Technology, Vol.3, 2008, PP. 291- 294.

[4]. Rajesh Shardanand Prasad and Uday Kulkarni, "Implementation and valuation of Evolutionary Connectionist Approaches to Automated Text Summarization", Journal of Computer Science, vol. 6, no. 11, 2010, PP.1366-1376.

[5]. W.M. Darling, and F. Song, "Probabilistic document modeling for syntax removal in text summarization". Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (CL' 11), ACM Press, Stroudsburg, PA.2011, PP. 642-647.