INFORMATION EXTRACTION FOR NOTES GENERATION

Jaideep .D. Shinde, Shamli .V. Mali, Ajinkya .S. Mohite, Viraj .R. Kamble

NBNSSOE, Pune

ABSTRACT

For a specific topic we find many reference books and textbooks to refer and the content is much more than one needs, therefore at the time of revision or quick study we need the summary and short version of the whole content for revision, especially at the time of examinations.

Therefore we aim to develop a semi-automated technique to generate notes from English text documents like Reference Books and Text books. The technique discussed is considered to be a pioneering attempt in the field of NLP (Natural Language Processing).This technique has a wide scope in the educational domain. The technique when implemented as an application can be used by both faculty members and students.

Key words *NLP, Segmentation ,Parsing, Ontology*

1.INTRODUCTION:

Here we talk about condensing a content report into progressively reasonable archive with less substance which covers a large portion of the vital focuses and gives a reasonable thought regarding the entire record, similar to a synopsis.

For the most part, in the vast majority of the course readings, just 20-22% of the words contain the data you have to comprehend or require while overhauling the theme. They are known as watchwords. Furthermore, the staying 80% contains near no fundamental data in that capacity, which comprises of pronouns, connectives like "of", "has", "for", and so forth. The main motivation behind these words is to connect the catchphrases together to frame significant and reasonable sentences. They are valuable for first time perusing, yet for amendments, simply the watchwords can take the necessary steps effectively, so overhauling only a little measure of content is much useful as perusing the entire content once more.

[1]There are many techniques involved such as information extractor which combines certain NLP methods like chunking, segmentation, summarization etc, with certain special linguistic features of the text such as the ontology of words, semantic links, noun phrases found sentence centrality etc. The process of the technique comprises of extracting text, creating an ontology, identifying important phrases for bullets and generating brief summary accordingly.[1]

We achieve this by using various methodologies and tools like Parsing, ontology creation, segmentation etc which are discussed further. For a given text document like a reference book we find too much information on a topic which is more than one needs or can handle, to tackle this problem we can make notes and study a topic, here we try to accomplish the generation of notes automatically with abstractive approach. This application can be used to get a gist of a given document and can be used for quick study like revision and omit the information from a document which is not much useful.

To achieve higher level of accuracy in [3] document classification more informative features of documents are taken into account. For this purpose, for instance, weight is assigned to HTML tags, which affects the efficiency of information retrieval; these are defined using genetic algorithms. Documents classification takes place at the level of

separate words, but not like classical works, the relevance of each word here is defined in relation to their informative features, which are the occurrences of a word in the title, emphasis is given on words by means of italic, bold fonts or its underlining and position of a word on the page. A DIG (Document Index Graph) algorithm is based on graph theory and phrases and their weights are taken into account for making suggestions in work [3].

In this approach for text segmentation,[5] we use an efficient linear text segmentation algorithm (called TSHAC).It considers both computational complexity and segmentation accuracy. The process of TSHAC has 4 steps. First, Preprocessing of long text; tokenization, stopwords are removed, and stemming are conducted to construct the vocabulary of the text. Text is then represented as vector after text preprocessing, each of which represents a sentence within the text. A part of sentence similarities are then computed to construct the sentence-similarity matrix. Finally, a the optimal topic boundaries are identified by the proposed algorithm.[5]

The method for extraction of meaningful sentences from the text to summary based on definition of score of relevance for each of the sentences and called "sentence by sentence" is suggested in this section.

The technique to be used is domain-independent which makes it unique from other techniques. From the evaluation measures applied to the technique, we can say that the technique helps in semi- automated generation of notes. The performance totally depends on the ontology provided as input. The more accurate the ontology creation, the more accurate the output will be.

2. METHODOLOGY:

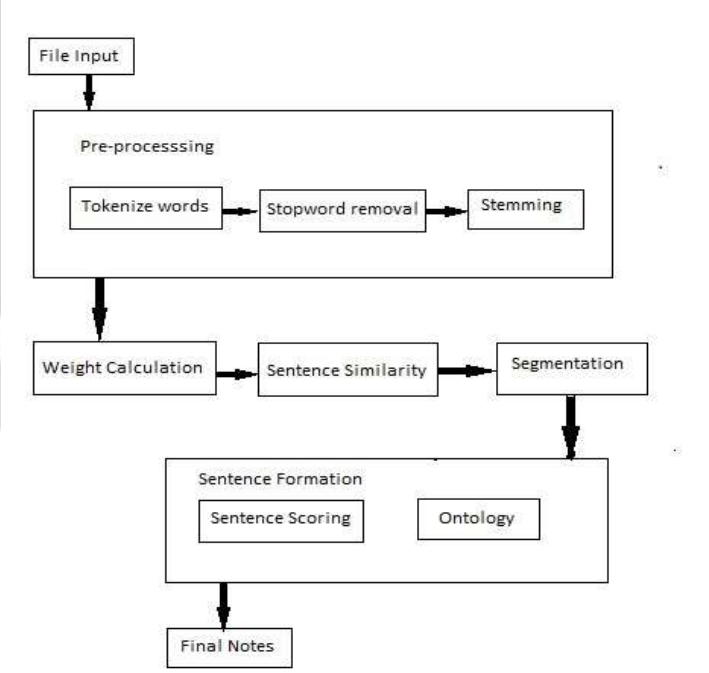

Figure 1. Architecture

- File Input

Input the file that is to be read and analyzed.

- Pre-processing

In pre-processing module there three main activities that take place namely (i) Word Tokenization (ii) Stopword Removal (iii) Stemming. In Word Tokenization we separate the input documents into separate unique individual words and sentences which are formerly known as tokens. For this we have used the NLTK.tokenize library.

In the Stopword Removal phase we segregate the stopwords from the tokens using the NLTK.corpus library. And in the stemming phase we bring down the words to their root phase. Eg: using will be stemmed as use.

- Weight Calculation

Where termFreqij represents the occurrences of word i in sentence j; tokenContj represents the total number of words. In sentence j; docCount represents the number of documents from a corpus; docFreqi represents the total number of documents contain word i.

$$w_{ij} = \frac{\sqrt{termFreq_{ij}}}{\sqrt{tokenCount_j}} (\log(\frac{docCount}{docFreq_i + 1}) + 1)'$$

Using the above we calculate the weights of each individual words and store the value in the database.

- Sentence Similarity

In this module we compare every sentence with its consecutive three sentences. We calculate the similarity function of the sentences using the Euclidian distance formula. We have used the SKlearn library for this.

- Segmentation

In the segmentation module we club the similar sentences together within an array. This is done by using the value of the similarity function of the sentences.

- Sentence Formation

In sentence formation module we score the sentences based on the value of similarity function and the weights. The sentences above the threshold value are then selected. And then using ontology we make the sentence meaningful by connecting the sentences. This whole process makes the sentences extractive.

- Notes

This is the final output of the system. it is the summary of the input document.

3.LITERATURE REVIEW

Here we discussed the literature review of existing techniques:

K.Gokul Prasad, Harish Mathivanan,Madan Jayaprakasam,T.V.Geetha [1] they propose a semi automated technique that generates slide presentation from a text document which is given as an input. A Model is proposed which contains 13 modules after implementing which we get a slide presentation using the bullet points extracted from the input document.

Alguliev, R.M, Aliguliyev, R.M, [2] propose a text summarization method that creates text summary by assigning relevant score to each sentence and extracting sentences from the original documents. While summarization this

method considers weight of each sentence in the document. The relevance score of a sentence is calculated through its comparison with all the other sentences and with the document title by cosine measure.

Ji-Wei Wu, Judy C.R. Tseng, Wen-Nung Tsai [3] propose an efficient linear text segmentation algorithm based on hierarchical agglomerative clustering. The proposed linear text segmentation algorithm is implemented without any auxiliary knowledge base, parameter setting, and user involvement. This technique provides comparable segmentation accuracy with several well know linear text segmentation algorithms.

Sr. No	**Paper Name**	**Author**	**Method Proposed**	**Limitations**
1.	Document Summarization and Information Extraction for Generation of Presentation Slides	K.Gokul Prasad, Harish Mathivanan, Madan Jayaprakasam, T.V.Geetha	A semi automated technique that generates slide presentation from a text document which is given as an input.	This proposed system uses text tiling algorithm for text segmentation which is not the optimum way for text segmentation.
2.	Effective summarization method of text documents	Alguliev, R.M, Aliguliyev, R.M	A text summarization method that creates text summary by assigning relevant score to each sentence and extracting sentences from the original documents.	The process of important sentence identification could be made more efficient and fast.
3.	An Efficient Linear Text Segmentation Algorithm Using Hierarchical Agglomerative Clustering	Ji-Wei Wu, Judy C.R. Tseng, Wen-Nung Tsai	The proposed linear text segmentation algorithm is implemented without any auxiliary knowledge base, parameter setting, and user involvement. This uses an efficient linear text segmentation algorithm using Hierarchical Agglomerative Clustering.	The algorithm can be applied in a much better way so that it can tackle more real life problems during difficult situations and perform efficiently.
4.	Creation of Ontology in Education Domain	Ayesha Ameen, Khaleel Ur Rahman Khan, B.Padmaja Rani	In this paper they have created course ontology based on the various courses offered by the university and added appropriate properties and restrictions. They have proposed various properties based on which ontologies can be prepared and how these properties can be chosen	A generic approach can be proposed to ontology creation. Using the semantics and the linguistics of the data.
5.	Text Summarization using Sentence Scoring Method	T. Sri Rama Raju, Bhargav Allarp	This paper discusses the simple and easy extractive technique of text summarization. It also proposes various methods for sentence scoring.	The paper proposes a linguistic approach for sentence scoring. Higher accuracy can be obtained by taking into consideration the context of the document.

I. COMPARATIVE ANANLYSIS

4.CONCLUSION:

The proposed technique is domain-independent, so we can extract notes from any document without any auxiliary knowledge base which makes it unique from other techniques. From all the evaluation measures applied to the technique, we conclude that the technique generates notes using a semi automated notes generation system. The performance totally dependent on the ontology provided as input. The more accurate the ontology, the more accurate the output or the generated notes will be.

5.REFERENCES:

1. K.Gokul Prasad, Harish Mathivanan,Madan Jayaprakasam,T.V.Geetha, "Document Summarization and Information Extraction for Generation of Presentation Slides ",***2009 International Conference on Advances in Recent Technologies in Communication and Computing***

2. Shwetambari Kharabe, C., "An efficient study on various biometric methods", *International Journal of Civil Engineering and Technology, 2018.*

3. Alguliev, R.M, Aliguliyev, R.M, "Effective summarization method of text documents", ***Web Intelligence The 2005 EEE/WIC/ACM International Conference, pp.264 – 271, Sept. 2005***

4. Shwetambari Kharabe, C, "Using Adaptive Thresholding Extraction - Robust ROI Localization Based Finger Vein Authentication", Journal of Advanced Research in Dynamical and Control Systems, 2018.

5. Ji-Wei Wu, Judy C.R. Tseng, Wen-Nung Tsai" An Efficient Linear Text Segmentation Algorithm Using Hierarchical Agglomerative Clustering",**Feb.2011**

6. Shwetambari Kharabe, C, "Survey on finger-vein segmentation and authentication", International Journal of Engineering and Technology (UAE), 2018.