VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Programme	:	**M.Tech Software Engineering**	Semester	:	**Fall2021**
Course	:	**Natural Language Processing**	Code	:	**SWE1017**
Faculty	:	**Dr. Tulasi Prasad Sarkar**	Slot	:	**G1**

NLP FINAL REVIEW DOCUMENT

PROJECT TITLE

TEXT SUMMARIZATION USING TEXT RANKING

TEAM MEMBERS

Setty Ruthvik-18MIS1048

Supriya.T-18MIS1064

ABSTRACT

Text Summarization is one of those operations of Natural Language Processing (NLP) which is bound to have a huge impact on our lives. The demand for automatic textbook summarization systems is spiking these days thanks to the vacuity of large quantities of textual data. In this design we take a dataset with the interview papers of sportspersons and epitomize the big papers into small paragraphs.

The algorithm we use is Text Ranking. We use Extractive Summarization, this relies on rooting several corridor, similar as expressions and rulings, from a piece of textbook and mound them together to produce a summary. Thus, relating the right sentence for summarization is of utmost significance in an extractive system

LITERATURE SURVEY

TECHNIQUES USED FOR TEXT SUMMARIZATION

Text summarization is generally divided into abstractive and extractive. The short description about each approach is discuss in following section:

Abstractive Summarization Approach

Summarizations by abstractive technique are commonly classify into two categories: Structured based approach and Semantic based approach

Structured Based Approach:

Structured based approach encodes mainly main information from the text through cognitive schemes such as templates, extraction rules and other structure such as tree, ontology, lead and body phrase arrangement

ABSTARCTIVE TEXT SUMMARIZATIONMETHODS: USING STRUCTURED BASED APPROACHS

Methods	Description	Advantages	Limitation	Author & Year
Tree Based Method(T BM)	-It use a dependency tree to represent the text of a document. -It uses either a language generator or an algorithm for generation of summary.	- It walks on units of the given document read and easy to summary.	- It lack a complete model which would include an abstract demonstration for content selection.	Barzilay and McKeown (1999, 2005) et al.
Template Based Method	-It uses a template to represent a whole document. Linguistic pattern or extraction rules are matched to classify text snippets that will be mapped into template slots	-It generates summary is highly coherent because it relies on relevant information identified by IE system	Requires designing of templates and generalization of template is to difficult	Harabagiu and Lacatusu (2002)
Ontology Based Method	-Use ontology (knowledge base) to improve the process of summarization. -It exploit fuzzy ontology to handle uncertain statistics that simple domain ontology cannot	-sketch relation or context is easy due to ontology -Handles uncertainty at realistic amount	-This approach is limited to Chinese news only. -Creating Rule based method for handling uncertainty is a difficult task.	Lee and Jian (2005) , Meghana viswanath(2006), et al.
Lead and Body Phrase Method	- This method is based on the operations of phrases	-It is good for semantically appropriate revisions for	-Parsing errors degrade sentential completeness	Tanaka and Kinoshita (2009) .

	(insertion and substitution) that have same syntactic head chunk in the lead and body sentences in order to rewrite the lead sentence.	revising a lead sentence.	such as grammaticality and repetition. -It focuses on rewriting techniques, and lacks a complete model which would include an abstract representation for content selection	
Rule Based Method	-Documents to be summarized are represented in terms of categories and a list of aspects.	-It has a potential for Creating Summaries with greater information density than current state of art.	-The drawback of this methodology is that all the rules and pattern are manually written, which is tedious & Time consuming.	Genest and Lapalme (2012)[2]

Semantic Based Approach

In Semantic based approach, semantic illustration of file is used to feed into natural language generation (NLG) system. This technique focus on identify noun phrase and verb phrase by processing linguistic data.

EXTRACTIVE TEXT SUMMARIZATION TECHNIQUES USING SEMANTIC BASED APPROACH

Methods	Description	Advantages	Limitation	Author & Year
Multimodal semantic model	A semantic model, which captures concepts and relationship among concepts, is	-An important advantage of this structure is that it produces abstract summary,	- The limitation of this structure is that it is automatically evaluated by humans.	Greenbacker (2011)

	built to represent the contents of multimodal documents	whose coverage is excellent because it includes salient textual and graphical content from the entire document		
Information Item Based Method	-The contents of summary are generated from Abstract representation of source documents, rather than from sentences of Source documents. - The abstract Representation is Information Item, which is the smallest element of Coherent information in a Text	-The major strength of this approach is that it produces short, coherent, information rich and less redundant summary	-It rejected due to the difficulty of creating meaningful and grammatical sentences from them. - Linguistic quality of summaries is very low due to incorrect parses	Genest and Lapalme (2011)
Semantic Graph Based Method	-This method is used to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the Original document, reducing the Generated semantic graph.	- It produces concise, coherent and less redundant And grammatically Correct sentences	This method is limited to single Document abstractive summarization	Moawad & Aref (2012) et al.

B. Extractive Summarization Techniques

An extractive summarization technique consists of selecting main sentences, paragraphs etc. from the original file and concatenating them into shorter form. The importance of sentence is determined based on statistical and linguistic features of sentences

Methods	Description	Author & Year
Term Frequency Inverse Document Frequency Method	-Sentence frequency is defined as the number of sentences in the document that contain that term. -Then this sentence vectors are scored by similarity to the query and the highest scoring sentences are picked to be part of the summary	M.Fachrurrozi, Novi Yusliani, and Rizky Utami Yoanita, (2013) et al.
Graph Theoretic Approach	-Graph theoretic representation of passages provides a method of identification of themes. -After the general pre-processing steps, specifically, stemming and stop word removal; sentences in the documents are represent as nodes in an undirected Graph	Rada Mihalcea, Niraj Kumar et al.
Text summarization With the Neural Networks	This technique involves training the neural networks to learn the types of sentences that must be integrated in the summary. -It uses 3- layered Feed Forward neural network	Khosrow Kaikhan(2004), Sarda A.T. and Kulkarni A.R.(2015).
Automatic TS based on fuzzy logic	-This method considers each characteristic of a text such as similarity to title, sentence length and similarity to key word etc. as the input of the fuzzy system.	Ladda Suanmali, Naomie Salim, and Mohammed Salem Binwahlan (2009) et al.
Query Based Extractive Text Summarization	In query based text summarization system, the sentences in a given document are scored based on the frequency counts of terms. -It uses Vector Space Model	Ibrahim Imam, Nihal Nounou, Alaa Hamouda et al.

The existed works on this topic and how they solved:
For text summarization they used many methods like tree , template , ontology , lead and body phrase , and rule based methods in structured based approach but they lacks in different ways like lack in model for content selection , designing and generalization of template is difficult and some are time consuming etc same for semantic based approach. In extractive summarization technique they solved by using methods like graph theoretic approach , with neural networks ,term-frequency inverse document method, automatic TS based on fuzzy logic, query based here also they lack like in TF-IDF which may be slow for large vocabularies. , measures are fundamentally limited in GTA , for NN in text summarization requires a lot of computational power, and some doesn't have the correct sense to summarize so these are all the methods used and solved.

The proposed method to solve the problem:
We use text ranking algorithm for text extraction, in this number of common words measure the sentences similarity , it gives the most informative document or summary also used in order to find the most relevant sentences in text and also to find most relevant keywords

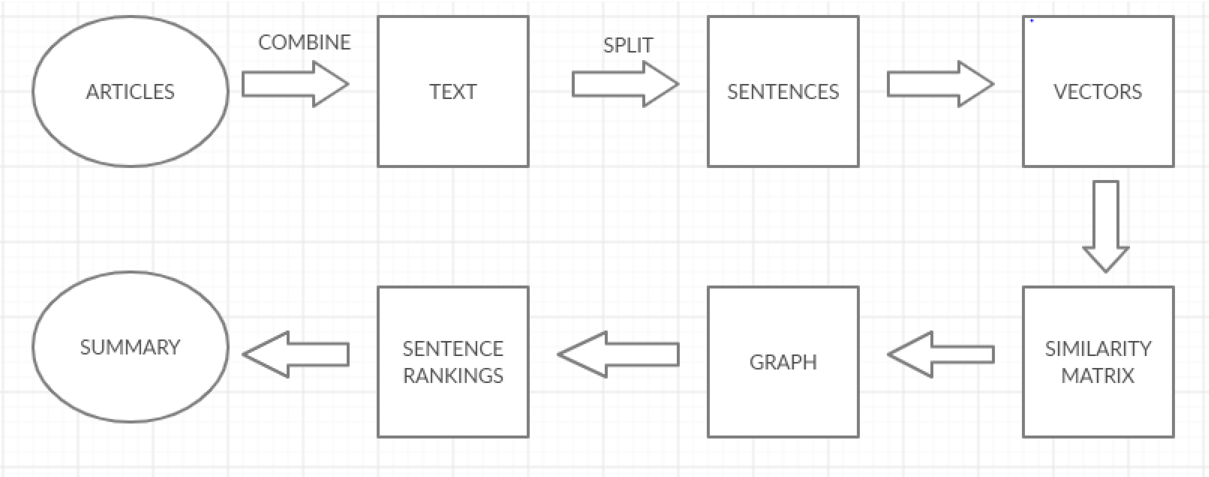
ALGORITHM

Text Rank Algorithm

Let's understand the Text Rank algorithm, now that we have a grasp on PageRank. we have listed the similarity between these 2 algorithms below:

- In place of web pages, we make use of sentences
- Similarity among any 2 sentences are used as an equal to the web page transition probability
- The similarity scores are store up in a square matrix, similar to the matrix M use for PageRank algorithm

Text Rank is an extractive and unsupervised text summarization method. Let's come across at the flow of the Text Rank algorithm that we will be following:



- The initial step would be to concatenate all the content contained in the articles
- Then split the text into the individual sentences in second step
- After that step, we will find vector representation (word embeddings) for every sentence
- Similarities between sentence vectors are then calculate and those are stored in a matrix
- The similarity matrix is converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation
- Finally, definite number of top-ranked sentences form the final summary

CONCLUSION

Automatic Text Summarization is a hot topic in research .Text Summarization is one of those application of NLP which is having clear to have a vast impact on our lives. With rising digital media and ever growing publishing – who has the point to go through entire articles / documents / books to choose whether they are useful or not.so here we are using text ranking algorithm which give the finest summarization and also gives us proficient prediction.

RESULT

The text summarization of the article was done efficiently by the text ranking algorithm. This algorithm finds plays an important role in summarization and is used in various application. And as well as the LSTM model also able to perform and evaluate efficiently.

FUTURE WORK

Coming to future work, we will explore the abstractive text summarization technique. In addition, we also look into the following summarization tasks: Problem specific-Multiple domain text summarization, Single-document summarization. Algorithm-specific: Text summarization using Reinforcement Learning.

REFERENCES

[1] Dazhi Yang_ and Allan N. Zhang Singapore Institute of Manufacturing Technology "Title of the paper Performing literature review using text mining, Part III: Summarizing articles using Text Rank".

[2] Ali Toofanzadeh Mozhdehi, Mohamad Abdolahi and Shohreh Rad Rahimi title " Overview of extractive text summarization" .

[3] Wengen Li and Jiabao Zhao School of management and engineering title "Text Rank algorithm by exploiting Wikipedia for short text keywords extraction."

[4] Sonya Rapinta Manalu, Willy School of Computer Science title "Stop Words in Review Summarization Using Text Rank ".

[5]Blog Vidhya Analytics https://www.analyticsvidhya.com/blog/2018/11/introduct ion-text-summarization-textrank-python/