ISSN 2321 3361 © 2020 IJESC

Research Article Volume 10 Issue No.7

Abstractive Text Summarization Techniques

Supreetha. D[1], Rajeshwari. S. B[2], Jagadish. S Kallimani[3]
Student[1], Assistant Professor[2], Associate Professor[3]
Ramaiah Institute of Technology, Bengaluru, India

Abstract:
Text summarization is a technique in natural language processing for constructing a tiny and crisp version of the large textual document. Creating the summary physically is a tedious job. Automatic summarization using machine learning solves this problem by generating the short summary from the original document by picking vital sentences from the original document as is known as extractive summarization, alternative technique to this which overcomes the drawback and much nearer to what human communicate is abstractive summarization, which consumes semantic information from the original document and produces the summary with completely new sentences. Automatic text summarization is applied in different domains, example search engines applies in reviews of textual resources, news websites use it for condensing the information to the shorter sort of news. This paper focuses on discussing various semantic similarity measures for achieving abstractive summarization and evinces the lateral idea of much optimal approach for the semantic similarity.

Keywords: Text summarization, Extractive summarization, Abstractive summarization, Semantic similarity measure.

I. INTRODUCTION

Everything is based on data and computation in recent days. Data is being generated enormously and consumed largely equally, for instance when a vehicle is driven it gives rise to various data such distance travelled, speed, mileage etc. data is preserved and fetched via semiconductor or cellular systems. After internet came into picture, massive amount of data got generated. The internet is the source of data.It houses information related to numerous domains such as entertainment, health, politics, world, medicine, education, literacy, news, topology etc. is accessible on internet. Data involves various forms such as images, pictures, videos, mathematical, statistical, numerical, and textual data. Among these text data is the quite hard to interpret since it involves huge amount of characters, hence it needs a system to extract the vital parts of the information we need, Text summarization is one way to achieve this. Text summarization is popular topic for study and research from past few years. Many models have been explored and experimented on the datasets to compute crisp summaries. All are compared. Summarization can extractive, abstractive, multi document or single document, and generic or query based.Extraction text summarization is a method of creating summaries picking up sentence directly from the original document. Abstractive text summarization is a technique which creates a general summary and focuses on the important topics in the original document. Single document summarization method generates summaries of the text from a single document and multi document summarization technique produces summaries from multiple documents. There is also another summarization technique which involves queries. Query based method receives query from the user.Text summarization is widely used across many areas like generating summaries of prescription given by doctors in medicine field, same way large amount of news articles is summarized to reduce the time taken by the user to read and gain information from several topics in very short time. In this paper we have surveyed various techniques of abstractive text summarization techniques. Natural language processing technique for text summarization uses python libraries such as nltk, scapy, scikit learn, Count Vectorizer for processing text [1]. Abstractive text summarization techniques use the semantic information of the text to understand the meaning of the document and generate the summary using completely new sentences. This paper discusses on various semantic measuring methods.Semantic similarity measure is a notion of determining the similarity between the words based on their semantic meaning. These bits of text are associated with algorithms and computer simulators that imitates the human conduct. semantic similarity is a classic approach which is popularly applied in text mining and natural language processing applications, example information extraction, chatbots, text summarization, machine learning etc.The similarity between a couple of words is computed using semantic similarity technique in natural language processing, two words are considered to have semanticsimilarity if they both give the same meaning. Computation of similarity in the meaning between the text context. content that is compared may be in the mode of documents, sentences, phases or words. Semantic similarity has a vital part in many natural language processing tasks and in many associated areas such as text summarization, text classification, document clustering etc. Word2Vec is a model in natural language processing used to represent the words in terms of vectors. These vectors are then used for achieving several tasks in natural language processing. One of the tasks in text summarization is that the similarity between the words can be calculated by applying mathematical functions on these vectors. The couple of words considered for semantic similarity calculation are transformed into vectors, the rate of closeness is computed using cosine similarity. It is calculated by obtaining the angle between the vectors using cosine function. Semantic similarity measures can be categorized as knowledge based and corpus based. The knowledge-based extracts meaning from the sentence using details from semantic storage area. The corpus-based methods process the document and fetches important information from it, and these are used for calculating the similarity among the words, phrases or sentences.

II. RELATED WORKS

Ordonezet. al.,[2] proposed the higher dimensional space example to language detects the relevance among the words and gathers all of them in a semantic area as a matrix with values, these values expresses the degree of relationship between respective words in rows and column. The algorithm calculates the relevance between the neighboring words and assigns the weight to each element based on the distance between the words under comparison. The higher the distance between the words in the matrix, greater is the weight assigned for the value in the matrix.Selection of similar words from the sentences using co-occurrences imagines that identical meaning can be found at identical contexts. The techniques use 3 words for finding the co-occurrences. Similarity is obtained by employing values to the respective vector od words. The fundamental calculation used for sematic similarity are based on words location and similarity.Explicit analyses the semantic similarity between text is determined by representing texts in the form of vector and calculating the relativeness between the text by applying the cosine measure.

Silvaet. al.,[3] discussed that the similarity between the words based on distance is calculated by measuring the distance between the text. Value zero specifies the two words occur together. Different values are assigned if the words are at certain distance apart.Semantic similarity is calculated based on probability, Higher the frequency of occurrences of the words at the same time, larger the score assigned to the word. Another variation in this method is to obtain the sorted catalogue of neighboring words that belongs to the two words under semantic similarity calculation. the similarity is calculated based on the same co-occurrence of the neighboring words which belongs to the words under comparison. Semantic similarity based on normalized distance, which is determined by comparing the word with the group of keywords.Distance between two words are close when who words have similar meaning. Words which are together have distance zero and words which are detached have defined distance.

Jainet. al., [4] presented that abstractive text summarization can be mainly categorized into two techniques: Structured based technique and semantic based technique.

1. Structured based approaches

It chooses the critical parts from the source documents using deep learning techniques: template, lead and body phrases, extraction rule structure, tree and template based structure.

•Tree based summarization:In this method the information/text of the source text in represented in the form of dependency tree.the information for the summary generation is absorbed based on applying various algorithms. The disadvantage with this method is it does not have perfect model that has exact representation of information retrieval.

•Template based summarization is a technique in which the end user has an independence to create a template of what should be included in the summary. The template contains the POS tags such as adverb, verb, noun etc. and the end user can specify in which order the sentences has to be present in the summary. End user can give as many templates as needed to generate the summary. End user can decide whether the dates and named items should be encompass in the resultant summary or not. This finalizes the task of creating the template. The template-based method considers the end user's requirement while creating the summary.it builds the highly logical summary, the drawback of template-based method it is effortful to create a template to generate summary.

•Another technique for text summarization is ontology based. The technique is acquired for creating ontology utilizes data pre-processing, semantic information extraction and ontology generation. Ontology is a technique of picking key phrases and then selecting the sentences for creating the rules for summarization. The disadvantage with this approach is that only domain experts can construct the ontology for the field which is tedious task.

•Lead and body phrase method: This technique is based on the mechanism of stages (insert and replace) which hassimilar syntactic head chunk in the beginning and core sentences to overwrite the leading sentence. Text is represented in the form of lead, body and supplement structure. Content is selected based on the maximum text with same lead and body structure. Summary generation is performed by applying insertion and substitution technique. This method is beneficial for semantic based summary, the drawback with this approach is grammatical error while generating summary since it deals with the semantics of the sentences.

•Rule based method: In this technique, the textual documents are summarized by representing them as a group and set of details. The subject matter selection component picks the most important sentences among several sentences which are generated by information absorption rules to answer the group of details. lastly, pattern generation component is used to generate summary out of sentences. This technique is suitable for generating abstractive summary for news articles on similar incident. The abstractive summarization technique uses rule-based information extraction component, subject matter selection components, rules and several patterns for sentence creation. all abstraction technique deals with specific topic and subgroups. Many verbs and nouns sharing common meaning are identified and syntactic location of roles are also located for generating extraction rules for extraction rules for abstraction technique. The information extraction (IE) component identifies many sentences for every details of the group. The sentence for the summary is picked based on the output of the information extraction unit. This unit act as a summary for the document using creation pattern build for every abstraction method. This technique can generate summary for huge document, but the disadvantage with is technique is the pattern and rules are designed physically, which is hectic and takes lot of time.

2. Semantic based approaches

•Multimodal semantic model: In this technique, a semantic unit, which extracts the subject matter and correlation among the topic, is created to represent the topic (pictures and text data) of one or more document. The vital topic is given score based on some criteria and lastly the chosen topics are generated as a sentence to form summary. The multimodal semantic model technique has a framework to create the summary. The technique contains both images and text. There are three stage in the framework, in the first stage, a semantic model is built based on the knowledge representation of topics by ontology. In the next stage, information content(topics) are assigned credits based on the richness of information measure. This measure gives the relationship of topics based on the amount of relatedness of topic, appearance of topic in present document and number of relatedness to other topics. In the last step the vital concepts are represented as sentences. The main advantage of this technique is that it generates abstract summary with all important information covered from the entire document. The drawback with this framework is that the resultant summary is physically verified by humans which could have been automated process [5].

•Information item-based method: in this technique, the information for the summary are created from abstract representation of original text document, then using the sentences from the original document. The abstract representation of the data entity, which is the little instinct of the related information in the text. The technique consists of the framework which has three units: information item retrieval, sentence generation, sentence selection and summary creation. In the information item retrieval phase, syntactic analysis of input data is performed with parser and the verb's topic and object are picked. Therefore, the information item-based method is defined as the positioned topic-verb-object triple. In sentence generation phase, a sentence is created from information item retrieval based on the average document frequency value. Lastly, a summary creation phases produces the crisp, relevant information retaining the original meaning of the source document by covering entire document. The advantage with this approach is that it gives short, exact and more content full summary without repetitive summary. The technique also has several limitations such as many sentences and phrases cannot be considered for summary generation If It is difficult to express with new sentence and if the linguistic metric of the summary is less due to wrong parses.

•Semantic graph-based method: This technique creates a summary building the semantic graph called Rich Semantic Graph (RSG) on the source document, condensing the semantic graph, and creating the complete abstractive summary from the condensed semantic graph. This method consists of 3 parts. The input document which is semantically using RSG. RSG represents the ouns and verbs of the input document as nodes of the graph, edges represent the topological and semantic relation between them. The second part condenses the created semantic graph of the input document to more compressed graph using heuristic rules. The third part creates the abstractive summary from the condensed RSG.This part receives the semantically represented RSG and produces the summary [6].

The important sentences for generating sentences are picked by using the following technique:

$\text{Concept}_{\text{weight}}$ =Average weight of every concept.

$\text{Sentence}_{\text{weight}}$ =Average weight of all the concepts in a sentence.

The best part of this technique is that it generates crisp, logical and unique and grammatically correct sentence,relevant, grammatically legit sentences and less repetitive summary. The shortcoming of this approach is it is can only work on single document for generating abstractive summarization and cannot be applied to multiple document.

III. PROPOSED METHODOLOGY

The architecture of the proposed model is as shown in the above figure 1. First step is to collect the dataset. Once the CSV file of the dataset is downloaded from the Kaggle.com.In the proposed system, CSV files are considred. The dataset consists of some unwanted things,need to remove all the unwanted symbols, characters, etc. From the text that do not affect the objective of our problem. The noise by using different preprocessing techniques. After removing the unwanted things using nltk, re and bs4 python libraries, final preprocessed data is obtained. The next step is the model building where we build the encoder-decoder network in LSTM, a three stacked LSTM for the encoder, decoder and attention mechanism layer.

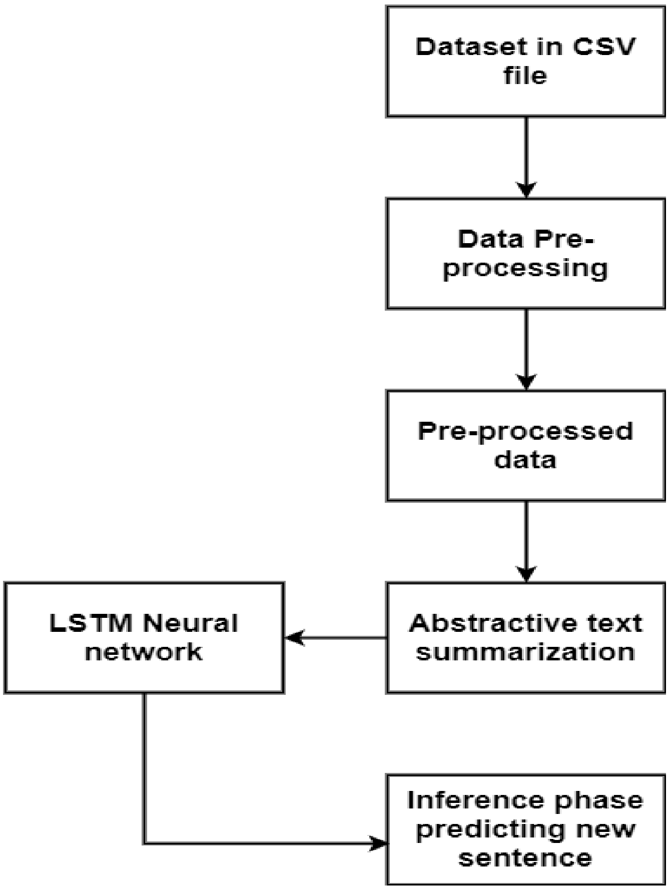

Figure.1. Proposed system architecture

Table 1 shows the comparison of different techniques, the advantages and disadvantagesas shown. The below figure 2 shows the graph of estimation for text summarisation. X-axis represents the Time in seconds and Y-axis represents the number of characters in dataset.

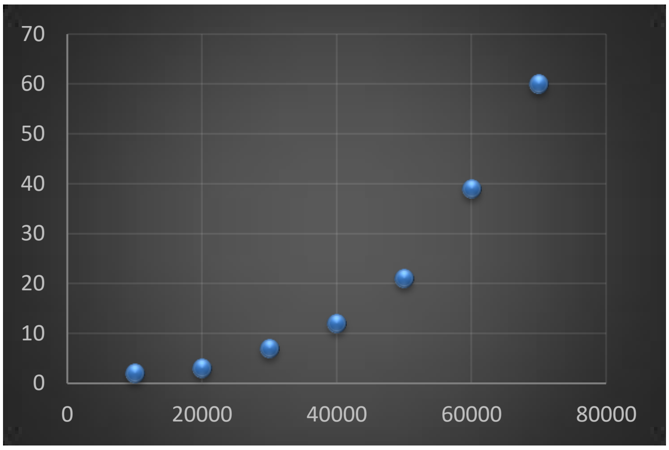

Figure.2. Graph of text summarisation

IV. CONCLUSION

Natural language processing has various applications and automatic text summarization is one of the popular and great techniques. There are generally two approaches to achieve text summarization such as extractive summarization and abstractive summarization.Researchin the field of automatic text summarization is an ancient challenge and the focus is shifting from extractive summarization to abstractive summarization. Abstractive summary technique generates relevant, exact, content full and less repetitive summary. Since the abstractive summarization technique focuses on generating summary which is nearer to the human intelligence it is a challenging field. Therefore, this study exercises the techniques for abstractive summarization along with procs and cons of various approaches. All the techniques are discussed and compared. This survey provides the way to get insight on the abstractive summarization.

Table.1. Comparison of text summarization techniques

Techniques/methods	Description	Advantages	disadvantages
Template based summarization	The end user has an independence to create a template of what should be included in the summary	Considers the end user's requirement while creating the summary. it builds the highly logical summary	It is effortful to create a template to generate summary.
Ontology based	The technique is acquired for creating ontology	Utilizes data pre-processing, semantic information extraction and ontology generation	Only domain experts can construct the ontology for the field which is tedious task
Lead and body phrase method	Based on the mechanism of stages(insert and replace) which has similar syntactic chunk in the beginning and core sentences to overite the leading sentence	Beneficial for sematic based summary	Drawback with this approach is grammatical error while generating summar since it deals with the semantics of the sentences
Rule based method	The textual documents are summarized by representing them as a group and set of details	This technique can generate summary for huge document	The pattern and rules are designed physically,which is hectic and takes lot of time
Multimodal semantic model	In this technique, a semantic unit, which extracts the subject matter and correlation among the topic,is created to represent the topic(picturesand text data) of one or more document	It generates abstract Summary with all important information covered from the entire document.	The resultant summary is physically verified by humans which could have been automated process
Information item-based method	The information for the summary is created from the abstract representation of the original text document,then using the sentences from the original document	It gives short, exact andmore content full summary without repetative summary	As many sentences and phrases cannot be considered for the summary generation if it is difficult to express with new sentence and if the linguistic metric of the summary is less due to wrong parses
Semantic graph-based method	Creates a summary building the sematic graph called rich semantic graph (RSG)on the source document,condensing the sematic graph,and creating the complete abstractive summary from the condensed semantic graph	It generates crisp, logical, unique and grammatically correct sentences, relevant, grammatically legit sentences and less repetative summary	The shortcoming of this approach is it is can only work on single document for generating abstractive summarization and cannot be applied to multiple document

V. REFERENCES

[1]. Rahul, SurabhiAdhikari and Monika, "NLP based machine learning approaches for text summarization",2020.

[2]. C. Ordonez, Y. Zhang and S. L. Johnsson,"Scalable machine learning computing a data summarization matrix with a parallel array DBMS," *Distrib. Parallel Databases*, pp. 329–350, Vol. 37, 2019.

[3]. G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Rissand H. O. Cabral, "Automatic text document summarization based on machine learning," pp. 191–194, 2015.

[4]. A. Jain, D. Bhatia and M. K. Thakur,"Extractive Text Summarization Using Word Vector Embedding," *Proceedings of International Conference on Machine Learning and Data Science*, pp. 51– 55, Vol. 2018, 2018.

[5]. Atif Khan and Naomaisalim, “A review on abstractive summarization methods”, 2014.

[6]. J. N. Madhuri and R. Ganesh Kumar, “Extractive Text Summarization Using Sentence Ranking,” *International Conference on Data Science Communication*, pp. 1–3, 2019.

[7]. Sreejith C, Sruthimol M P and P C Reghuraj, “Box Item Generation from News Articles Based Paragraph Ranking using Vector Space Model”, *International Journal of Scientific Research in Computer Science Applications and Management Studies*, Vol. 3,2014.

[8]. M. Moradi, G. Dorffnerand M. Samwald,“Deep contextualized embeddings for quantifying the informative content in biomedical text summarization,” *Comput. Methods Programs Biomed.*,pp. 105117, Vol. 184, 2020.

[9]. B. Mutlu, E. A. Sezerand M. A. Akcayol,“Multi-document extractive text summarization: A comparative assessment on features,” *Knowledge-Based Syst.*, pp. 104848, Vol. 183, 2019.

[10]. M. Afsharizadeh, H. Ebrahimpour-Komlehand A. Bagheri, “Query-oriented text summarization using sentence extraction technique,” *Fourth Int. Conf. Web Res.*, pp. 128–132, 2018.

[11]. M. Mauro, L. Canini, S. Benini, N. Adami, A. Signoroniand R. Leonardi,“A freeWeb API for single and multi-document summarization,” *ACM Int. Conf. Proceeding Ser.*, Vol. Part F1301, 2017.

[12]. T. Jo,“K nearest neighbor for text summarization using feature similarity,” *Proceedings of Int. Conf. Commun. Control. Comput. Electron. Eng.*, pp. 1–5, 2017.