JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 18, 2020

TAXONOMY ON DOCUMENT SUMMARIZATION METHODS, TOOLS AND OPPORTUNITY

Sowmya M S[1], Dinesh R[2]

Research Scholar, Jain University[1], Visiting Professor, School of Engineering and Technology, Jain University[2]

mssowmya.sbmjce@gmail.com[1], dr.dineshr@gmail.com[2]

Received: 28 May 2020 Revised and Accepted: 06 July 2020

ABSTRACT: In the current age, internet is formed of incredible information, it is very much needed to bring a better way to gather the information faster and proficient. For a human being it is not an easy job to manually write the summary for a large text document. Internet as ample text materials available in it. So, it is difficult to search for relevant document from N document in different sources in internet. In order to solve these problems, the there is need to automate text summarization. Text summarization is the procedure to find the most important meaningful data in set of documents and to reduce them into a brief version preserving its overall denotations. Here, we highlight the different document summarization methods existing in literature, the tools that does this summarization and the scope that shows us the limitation and possibilities in enhancing the accuracy of the results of those methods.

KEYWORDS: Extractive summarization, Abstractive summarization, Natural language processing.

I. INTRODUCTION

A summary may be a short script that's made of one or many transcripts, that carries significant data within the original document, and in small proportion. the target of automatic text summarization is to offer the input text as a short gist with meaning. This lessens the understanding time. Approaches of Text Summarization are classified as technique of extraction and abstraction [1,2]. The extractive method comprises of choosing key sentences, subsections, etc. from the input document and consolidating them into short form. Abstractive technique includes the semantically thoughtful sort of the concepts in input file and then expressing those as human natural tongue

Inductive and informative are the twofold of text summarization procedures. Inductive method denotes the key idea of the text [2]. The output length of this type of summarization will be up to 0.1 percent to given input text [7]. Informative method provides short-term information about the given input document. The output summary length will be up to 0.3 percent of the given input.

II. SUMMARIZATION PROCESS

The main stages of summarizing documents include topic selection, Interpretation, and summary generation [3].

- Topic selection is about finding the key information in the input document. Some forms of these are Word frequency, Cue Phrases, Position are mainly used techniques and most are dependent on the phrases position [9].
- Interpretation is to form the abstract summaries. Here, dissimilar topics are attached to generate a universal content.
- Generation of the summary is possible by applying method of text generation in the system.

III. EXTRACTIVE METHOD

The extractive process includes Pre-Processing and Processing. Pre-Processing tries to represent the input text in a structured form. It has steps like stemming, stop-word filtering, sentence boundary identification [4,8]. The role of a dot in the end of the sentence is used to give a structure in sentence boundary. In stop-word filtering, the most common words are removed removing context. In stemming, it tries to find the stem of the word by removing suffices like ed, ing, ed, es, etc. The processing step tries to find the features that matter for sentences, then using the weighted learning method, the features are given weights. The feature-weight equation is used to find the final score of sentences. The ultimate summary will be a collection of top-ranked sentences. Summaries

ISSN- 2394-5125 VOL 7, ISSUE 18, 2020

are assessed by measures either intrinsic or extrinsic. Intrinsic methods summary is a human judgement based and extrinsic methods are performance-based like that of information retrieval-oriented task.

IV. HISTORY OF TEXT SUMMARIZATION

Extractive summarizers performed on sentence scoring on the input document. And it works on techniques like statistical methods or linguistic approaches [5]. Mostly the sentence weighting method, frequency-based method, cue method, standard keyword, Location or Title approaches.

The extractive summarization method is more used, because abstractive methods need NLP implementation in order to generate a summary with statements that are not in the document but without a change in meaning. it's sort of a summary said by human. One among the extractive methods generates a summary called sentence scoring. It does it by assigning a [2] numerical value to a sentence and later based on its compression rate the sentences are selected for a summary generation. The compression rate may be a factor that states the ratio between the input text and length of the summary. The compression rate increases making the summary to be larger. The decrease in compression rate makes the summary be short, but it lacks information. As per literature, the standard summary must have the compression rate of 5-30% [10].

V. METHODS OF EXTRACTIVE SUMMARIZATION

A. Term Frequency-Inverse Document Frequency (TF-IDF)

To know the importance of a word in a input document using numerical statistics this technique is vital [6-10]. In the input document, based on the frequency of a word, a proportional increase within the TF-IDF cost is predicted. The key techniques to this are inverse sentence frequency and weighted term-frequency. Sentence-frequency is the count of the of sentences within the document that comprise of term t. Summary is made up of the scores generated by highest-scoring sentences and similarity [2] called sentence vectors.

The target words are usually nouns. the idea is that the sentence which is relatively important will have more specific words in that sentence. Here, each sentence is as document and the frequency of words in this document is its document frequency(df), and that is how it is different from TF. The TF/IDF score is calculated as follows:

$$TF/IDF(w) = DN\left(\frac{log(1+tf)}{log(df)}\right)$$

where DN is the number of documents.

B. **Cluster based**

Triplets include verbs, subjects, objects related to sentence S_i. This method captures the semantic nature of an input document; it expresses the natural language by triplets. These triplets are clustered using similar information [11]. So, related to computed clusters the sequence of sentences forms the summary.

C. **Graph theory based**

Here, the node represents a sentence [12]. When there is common word shared by two sentences, there will be an edge between two nodes, it will have the threshold that less than the similarity. This has two outcomes: one is sub-graphs that are not connected due to unique topics in documents. Second is that, it considers cardinality, high cardinality nodes help in identifying the important sentences in the document. So, summary will retain the higher preference.

D. **Machine Learning**

This method models the summarization as a classification problem by using a training dataset as a reference [13]. Features will division the sentences into non-summary sentences and summary sentences [11]. Using training data, the statistical classification probabilities are learnt in Bayes rules:

ISSN- 2394-5125 VOL 7, ISSUE 18, 2020

$$P(s \in <S \mid F1, F2, ..., FN) = P(F1, F2, ..., FN \mid s \in S) * P(s \in S) / P(F1, F2, ..., FN)$$

here, P (s∈< S | F1, F2, ..., FN) [4] is the sentence 's' probability, chosen to form the summary 'S' by satisfying features F1, F2...FN.

E. Text summarization with neural networks

Here, the list of sentences is made from the document. [2] A vector form [f1, f2,..f7], with features like paragraph, title, location, sentence location, first sentence, sentence length, count of thematic words and count of title words make a sentence [14].

Initially [4,12] in neural networks, the summary includes the types of sentences as per the training given. It also considers the summary with needed features. Feature fusion phase includes the features that are essential in the mainstream of sentences by determining relationships and trends. It will remove the rare features step and breaking up the influences of communal features.

E. Text summarization with neural networks

Here, the input fuzzy system is a sentence length, similarity to tittle, the similarity to the keyword, etc. the knowledge base will have the rule as per the need [6] of summarization. A zero to one is a value got for sentence S_i and based on sentence features & existing rules it is part of output. This value in the output determines the degree of the importance of the sentence in the ultimate summary.

VI. METHODS OF ABSTRACTIVE SUMMARIZATION [15-17]

It is important to advance the topic coverage by focusing more towards semantics of the words to achieve readability of automatic summaries and its correctness [1]. This is needed to experiment with re-phrasing of the input sentences in a human-like fashion.

A shift from extractive to abstractive method of summarization will solve the problem. Opposing to extractive methods, abstractive technique's summary output will be more readable and it takes care of even grammar.

Encouraged by accomplishment in machine conversion, abstractive summaries are formed from a set of techniques that are deep-learning based [1]. Depending on their focus, the approaches are semantic-based ones and structure based.

A. Abstractive summarization structure-based approaches

Structure based methods principal is to use the emotional feature schemas and the preceding information like substitute structures like trees, ontologies, lead and body, graphs to encode the most vital data, patterns, and extraction rules[15-18].

i. Tree-based methods

In this group, the important knowledge is based on a dependence tree that is the text of a document. The algorithms of content selection vary meaningfully to different algorithms from theme intersection having application in the outline-based content choice. The language generator or degree algorithm helps in the generation of the outline.

ii. Template-based methods

Here, extraction rules and linguistic pattern are coordinated to identify the text snippets that is mapped to the guided slots [1] of an entire document making use of guide. For outline contents, these snippets function

are the indications. GISTEXTER, is a summarization scheme that aims to identify the topic-related information within the given document, it interprets this to entries in database and inserts sentences to unexpected summaries.

iii. Lead and body phrase method

This method of summarization to broadcast news by Tanaka has analysis of syntax of the sentence [1]. By the inspiration of technique of sentence fusion, the method finds the similar phrases in this method continued by adding and assigning of phrases to bring summary via sentence revision. Operations are parsing syntactically, finding search combinations triggers, alignment of phrases. As last step, add, assign or both are done to get a modified sentence.

iv. Rule-based methods

This method the documents that are inputted in terms of features list and classes. A sentence is generated in this method using more than one patterns, heuristics of content selection and module of extraction. Common nouns and verbs are found for the extraction rule generation and many equivalent rules are found and given to the generation module of summary. For the outline sentences generation, patterns generations are used. The method is good for summary but it takes much time when the inputs are given manually.

v. Graph-based methods

This method novelty is in the rule that each node is a word entity that indicated the sentence structure having directional edges. Opinosis is the system that uses this method. It is a outline that produces dense summaries of abstractive type, it is of tremendous opinions that are redundant. It recurrently finds the opinosis graph for encoding the sub-graphs as a sentence of validness and phrases of candidate summary is generated by this model to identify the valid paths. Redundant paths are removed by the use of measure of Jaccard for the short generation of summary and all paths are graded scores in descending order.

vi. Ontology-based methods

Any domain be, it has structure and can be illustrated in ontology. This method has step to reduce sentences by the method of reformulation. It can be by making use of NLP or linguistic methods. Fuzzy ontology is a method by Lee, news summary generation of Chinese is its main use with input that are information of uncertain. Membership degrees are generated by the phase of fuzzy inference for every fuzzy concept and agent does the news summary generation related to fuzzy ontology.

vii. Semantic-based approaches

These methods services based on morphology representation of text(s) to input to a system of NLP. It mainly focuses on the finding of verb and noun phrase concept. In Multimodal semantic model, the text and image in the document are represented by the relation forms and concepts. A semantic model is initially constructed using facts illustration based on objects. Nodes are thoughts and links are relationships. The selected thoughts are made as sentence to form summary. SimpleNLG is one such example which provides control on phrases and joined.

viii. Information item-based methods

Here, the abstract is produced from nonconcrete illustration of the input file. The focus is to find the entities of text, its features, and base characteristics of them. This method first does information item fetch, using statistical analysis subject object and verbs are made. It has to combine structure to give text only then full sentence can be generated. Next, the sentences are ranked using frequency score of an average document. The highest rank sentences are used to give abstract with a correct plan. It gives short and redundancy less summary, but it ignores the most important information and less in quality in terms of linguistic concepts.

ix. Semantic Text Representation Model

Here, the focus is to examine the inputted text for semantics than syntax. Semantic role labeling [1] is one such method that excerpt base argument structure and the sentences are taken from the document with positions. SENNA labeler helps in assigning the number to the position. The scores of similar semantics of graph creates matrix of similarity. An algorithm of updated graph will find the structure predicate, similar semantic and relationship set of the document. Using Maximal Marginal Relevance, the redundancy of summary is reduced.

x. Semantic Graph Model

Here, a graph called rich semantic graph (RSG) is created to generate the summary. It works in three steps, the document inputted is kept in the RSG form where the nodes are nouns and verbs of the document and relationships are edges. The graph generated makes use of rules of heuristics to reduce the original graph. From this an outline abstract is created. This method gives unique and Grammarly correct sentences in summary but it works for only one document.

Abstractive summarization demonstrations are less steady unlike extractive approaches, it is good for human-like summaries but still needs improvisation in accuracy as per applications.

VII. EXISTING TEXT SUMMARIZER TOOLS [15-18]

A. Text Compactor

Here, user have to set the percentage of text to keep in the summary. User can choose within the range of 1-100%. If not satisfied with the result, we can change the percentage and try again. The limitation of this is that it won't do any study on contents, it just takes N% given data from beginning and displays it as summary.

B. SMMRY

This summarization tool has all that user need for a faultless summary. It is flexible to use with more features and updated settings. SMMRY permits user to breif the text with upload and URL input too along with direct text input. The limitation of this is that, it can generate only seven sentence summary that may not include the connectivity in statements that are shown in results. It will make the summary meaning less.

C. Autosummarizer

The limitation of this is that, it can generate only six sentence summary that will not consider the relation in statements to retain the meaning.

D. Summary Generator

This tool generates summary by covering few paragraphs in the given input contents. It highlights only definitions and may not consider working, algorithms in its summary.

E. Resoomer

A fantastic paraphrasing and summarizing tool that comprehends numerous tongues, including many foreign languages, limited to Spanish, German, Italian, Engligh and French.

But the contribution must be solitary texts of argumentation. It is only for this purpose and don't consider any special cases.

F. Summarizer and Simplify

These both are flexible feature to summarize the document. This simple Chrome extension will provide user with a summary within a couple of clicks. Install the add-on, open the article or select the piece of text user want to summarize and click the button "Summarize". But summary accuracy is still not as required.

G. Split Brain summary tool

This tool summarizes articles for lot many languages. Thirty-nine language choice are available and many sentences can be made out of these. It will also produce the difference in summaries based on ration given. 5% to 80% is the allowed paraphrasing density. URL input works here too. But the summary generated is just a percentage split than the meaning gist.

F. Summarize Bot

It is combination of many algorithms to give the accurate summary as per the current need. It works for most language compared to other tools discussed. It makes use of AI, block chain, NLP and machine learning as the main core technologies. This suffers from the lack of user friendliness in application.

VIII. OBSERVATIONS

Extractive techniques are more used due to implementation complexity of abstraction techniques. There is hardly any ready application which gives accurate summary for the given documents. Few of the tools that are discussed are all not even 90% accurate in its results. Most are formed from extractive methods. We will choose the best method and work on improving the acuracy in the results of the existing methods in our next paper.

IX. CONCLUSION

Text summarization is a need in the current research emerging trends. This is due to overloaded information in these fields, especially when connected to internet applications or business. Extractive summarization selects relevant sentences, words from the input document mostly based on the similarity or frequency of sentences and words as per the method chosen. The abstractive method is like human written summary, so it quite challenging compared to extractive technique. Lot of algorithms and tool are available to help in generating summary as discussed, but still there is a scope to improving the accuracy of the summary generated.

X. REFERENCES

[1] Debabrata Khargharia, Nomi Baruah, " APPLICATIONS OF TEXT SUMMARIZATION", International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Volume 9, No. 3, May-June 2018, pp 76-79

[2] Personalized Multimedia Web Summarizer for Tourist , Xiao Wu et. al. , Key Laboratory of Intelligent Information Processing Institute of Computing Technology, CAS, Beijing, China , April 21-25,2008

[3] Improving Legal Document Summarization Using Graphical Models. https://www.researchgate.net/publication/220809892_Improving_Legal_Document_Summarization_Using_Graphical_Models [accessed Dec 16 2017].

[4] Rada Mihalcea and Hakan Ceylan. Explorations in Automatic Book Summarization. Department of Computer Science. University of North Texas. January 2014

[5] Summarization from Medical Documents: A Survey. Available from: https://www.researchgate.net/publication/220103096_Summarization_from_Medical_Documents_A_Survey [accessed Dec 16 2017].

[6] https://doi.org/10.1016/j.jbi.2014.06.009 [accessed Dec 16 2017]

[7] Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah, "A Review on Automatic Text Summarization Approaches", Journal of Computer Science, science publications, 2016, 12 (4): 178.190 DOI: 10.3844/jcssp.2016. pp 178.190

[8] DEEPALI K. GAIKWAD and C. NAMRATA MAHENDER, "A review paper on Text summarization", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016, ISSN (Print) 2319 5940, pp 154-160

[9] Ahmed El-Refaiy, "Review of recent techniques for extractive text summarization", Journal of Theoretical and Applied Information Technology · December 2018 Vol.96. No 23, pp 7739-7759 .

[10] Vishal Gupta , Gurpreet Singh Lehal ,"A Survey of Text Summarization Extractive Techniques", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010, pp 258-268

[11] https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning- 3d27ccf18a9f

[12] https://machinelearningmastery.com/gentle-introduction-text-summarization/

[13] https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/

[14] https://pkghosh.wordpress.com/2019/06/27/six-unsupervised-extractive-text-summarization-techniques-side-by-side/

[15] https://heartbeat.fritz.ai/extractive-text-summarization-using-neural-networks-5845804c7701

[16] https://medium.com/sciforce/towards-automatic-summarization-part-2-abstractive-methods-c424386a65ea

[17] https://ivypanda.com/online-text-summarizer

[18] https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-revisited/