IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Text Summarization Using Ranking Algorithm

Aruna Kumara B[1*], Smitha N S[2], Yashaswini Patil[3], Shilpa P[4], Sufiya[5]

[1,2,3,4,5]School of C & IT, REVA UNIVERSITY, Bengaluru, India

[*]*Corresponding Author: arunakumara.b@reva.edu.in Tel.: 0-8951755795*

DOI: https://doi.org/10.26438/ijcse/v7si14.266269 | Available online at: www.ijcseonline.org

***Abstract* —** The rapid growth of the online information and textual resources has made the text summarization more favourite domain to emphasise the importance and intention of textual information. Manual summarization of large source documents is arduous. Text summarization is automatic text summarization which shortens and condenses the original text document withou t any loss of original content in an efficient way. In recent years text summarization is one of the most favourite research domai ns in natural language processing and could attracted more attention of NLP researchers. The intact relationship exists between text mining and text Summarization. In this work, topic of text mining and text summarization considered in the beginning. Th ere after a model has been designed on some of the summarization approaches and essential parameters for exerptting predomi nant sentences, found the main steps of the summarizing process, and the most significant extraction criteria are presented.

***Keywords*—** Text summarization, manual summarization, summary, text ranking**.**

I. INTRODUCTION

Text summarization is the method of developing small, preci se, and eloquent summary of a larger text document. Raved e t al. (2002) define a summary as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". This simple d efinition catches three essential aspects that characterize rese arch on automatic summarization: Summaries can be created from a one or more documents, summaries must conserve es sential information, Summaries need to small in size.

Automatic text summarization methods are mostly required t o address the rapidly increasing amount of text data present o nline to help explore related information and to absorb relat ed information rapidly.

Present methods try to correlate and match the chunks of the summary with the chunks of summaries produced by humans and measure the similarity of the chunks in summery generat ed compared to the human produced summary. One approach is to take the sentence as the chunk text unit in the calculatio n procedure, but the challenge is the sentences consists indivi dual meaning which will not be used by human as reference. Selecting the correct chunk size and comparing it with appr opriate one is a crucial challenge. The essence of the proble m is to excerpt the relatable units which express the informati ve contents of a text.

The ranking of key phrases is carried out. That represents the important concepts the given text and ranking based summar y is introduced. In the evaluation process, the evaluator consi der the key phrases as the matching unit. The main motive o f this to count the matches of the generated summary with res pect to the reference summary. The Dataset are present into t hree modules, a) Feature Extraction module that breaks the te xt into words and extracts their lemma forms and the associat ed lexical and syntactic features, ii) Sentence Ranking that ex tracts important key phrases in their lemma forms and the ev aluator that scoring the summary based on counting the matc hed key phrases and ranks them, iii) Redundancy Reducing o ccur between the peer summary and one or more reference su mmaries. The remaining of this paper is organized as follows : Section II reviews the previous works; Section III the propo sed Systems development research methodology; Section 4 d iscusses the performance evaluation; and section 5 is the con clusion.

II. RELATED WORK

In this section, we will discuss certain other research studies that have been conducted on Text summarization LUHN's work on text summarization showed that frequency of words in sentences has more importance and relevance in the final result. The methods proposed by Luhn are still effective even after 50 years old. He also proposed removal of stop words, stemming .The words are given a hierarchy and each word's significance is described by its index. This will then calculate the number of time that particular word occurs in the sentence and then it is ranked according to that [1]. JING

says from his work that removal of irrelevant phrases like prepositional phrases, clauses, to infinitives, gerunds from sentences was of prime importance as they don't have any significance in the summarization process [2]. BAXENDALE in his study on over 200 paragraphs found that, in over 85% of those paragraphs the topic of the para would appear in the first sentences itself. And in 7% of the paragraphs the topic would appear in the last sentence. By this he came to a conclusion, that most of the times the topic appears in either the first or last sentence of the paragraph [3]. FANG CHEN ET Al in their work observed 3 features. The Sentence location feature meant that most of the times the beginning and the end of the sentences would contain the useful matter. The second one is the paragraph location feature which is same as the sentence location feature. The third feature is the sentence length feature where the sentences that are too long or too short are not featured in the summary. The threshold for the number of words can be preset [4]. EDMUNDSON typical structure that produces extract. He used the word frequency and word position feature. He also gave us two new features, cue words and skeleton. The sentences were scored basing upon these features which were then extracted for summarization [5].

What is Automatic Text Summarization?

Automatic text summarization, or just text summarization, is the process of creating a brief and comprehensible interpretation of a longer document.

Text summarization is the process of refining the important information from a sources to produce an abstract adaptation for a specific users and works.

Advances in Text Summarization : Human beings are good at understanding the raw or given information then analyse it and refine according to the needs without the loss of real meaning. As such, the target of automatically creating summaries of text is to create the resulting summaries as efficient as the summaries written by human beings. The motive of automatic text summarization is to implement the techniques which imitate the technique of summarization from human beings.

Innovations in Text Summarization Developing the summaries with the phrases, lines catching the gist of real document will not suffice if the summaries are not as fluent as standalone document.

Different approaches for Text Summarization

There are two main approaches to summarizing text documents; they are:

1. Extractive methods
2. Abstractive methods.

The different aspects of text summarization can be broadly classified depending on its input type, purpose (generic, domain specific, or query-based) and output type (extractive or abstractive).

Extractive text summarization involves the choice of phrases and sentences from the source document to develop the new summary. Techniques include ranking the relevance of phrases in order to select only those most relevant to the meaning of the source.

Abstractive text summarization involves generating completely new phrases and sentences to catch the meaning of the source document. It is a more challenging approach which is finally used by human beings. Classical methods operate by selecting and shrinking the information from the source document.

III. METHODOLOGY

We use "Systems development research methodology" from the information system research field as our research methodology.

Working Concept:

- In the previous techniques we used ranking based on the methodology used.
- The word level and sentence level features are used in text summarization literature.
- In the present work, we use different kind of documents as datasets and summarize them in an
- Efficient manner

The following steps were followed to explore automatic text summarization:

Step 1: Choose and clean datasets
Step 2: Build the extractive summarization model
Step 3: Build the abstractive summarization model
Step 4: Test and compare models on different datasets
Step 5: Tune the abstractive summarization model
Step 6: Build an end to end automatic summarized application

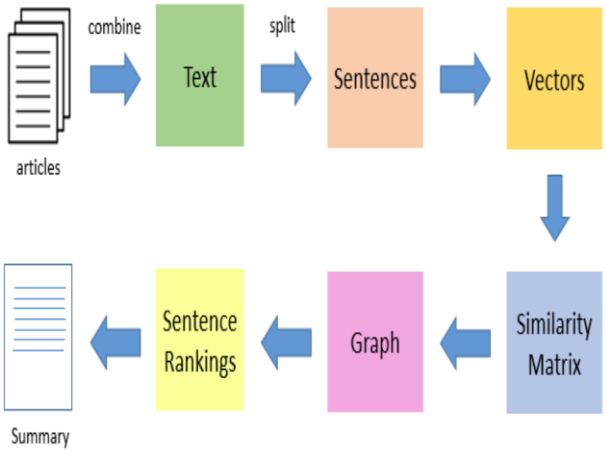

Fig 1: Block diagram of Text summarization

Choose and clean datasets
This section introduces the basic information about each data set we used, precisely the contents of the dataset, and the reason for using the dataset.

Datasets Information
We work on 2 datasets. The issues dataset and amazon reviews dataset. There are 1lakh+ cases in this data frame, which is the largest dataset we worked on. For each case, we filtered the dataset to only keep the unique question id, the question title, the question body, and the answer body. Then we cleaned the filtered dataset by removing chunks of code, non-English articles and short articles. The reason we chose to work with these Datasets is because it contains technical issues similar to that of the KB Dataset. However, the reviews Dataset is supposedly cleaner than the KB Dataset, and by running our models in a cleaner dataset, we could first focus on designing our model to set a benchmark.

Data Cleaning
The datasets we worked were very noisy containing snippets of code, invalid characters, and unreadable sentences. For an efficient training, our models needed datasets with no missing value and no noisy words. Based on this guideline, we followed these basic steps to clean our datasets.
- Read data file and make a data frame
- Check for missing values.
- Detect and remove the code part in all texts.
- Detect and remove the unknown words with special symbols in all texts.

Techniques Used
Frequency-Driven Approaches: The two most common techniques are used to determine the more relevance words to the topic namely Word Probability (WP) and Term Frequency -Inverse Document Frequency (TFIDF). The WP is used Frequency of words as indicators of importance is word probability. Text Rank is an extractive and unsupervised text summarization technique.
- In the first step we link all the text present in the articles in a chain.
- We divide the text into separate sentences
- The next step contains the searching of vector representation (word embeddings) for individual sentence
- Similarities between sentence vectors are then calculated and stored in a matrix
- The similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation
- The top ranked sentences form the summary in the last step.

Extractive text summarization: Extractive text summarization involves the choice of phrases and sentences from the source document to develop the new summary. Techniques include ranking the relevance of phrases in order to select only those most relevant to the meaning of the source. We use Sentence Rank algorithm here.

Abstractive text summarization: Abstractive text summarization involves generating completely new phrases and sentences to catch the meaning of the source document. It is a more challenging approach which is finally used by human beings. Classical methods operate by selecting and shrinking the information from the source document. We use Text Rank algorithm here. Final results prove that Sentence rank algorithm is more efficient and accurate.

IV. RESULTS AND DISCUSSION

In the application we upload the particular document. Then the preprocessing, sentence scoring, sentence ranking is performed. The size of summary is provided in the code which can be modified according to the need of the user. The final result will contain the lines which have highest score. Those lines will be given as summary.

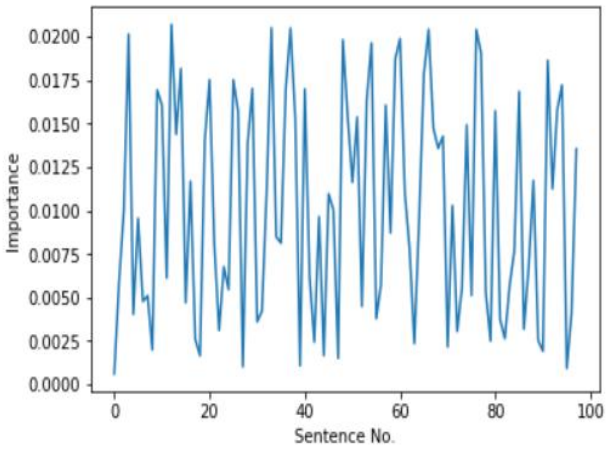

Fig. 4: Sample graph - Sentence extraction from a text article

V. CONCLUSION AND FUTURE SCOPE

Summaries are written or developed to lessen the reading time. The summaries make the process of researching documents easier. It collect data consisting of a combination of various attributes and then use it as inputs to various machine learning algorithms. These machine learning algorithms work on some selected features of the data and compare the performances. Automatic summarization improves the effectiveness of indexing and provides unbiased summaries compared to human being. Personalized summaries give personalized information used in interrogative systems. Using automatic or semi-automatic summarization systems enables commercial abstract services to expand the number of texts they are able to process. Further this work can be enhanced into a single line summary by feeding texts and getting summary in single line etc.

REFERENCES

[1] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modelling. arXiv preprint arXiv:1412.3555. (2014).

[2] Das, D., & Martins, A. F. A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU, 4, 192-195. (2007).

[3] Facial feature extraction Using Hierarchical MAX(HMAX) Method Akshaya Pisal ; Ravindra Sor ; K. S. Kinage.2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA).(2017).

[4] Graves, A., Mohamed, A.-r., & Hinton, G. Speech recognition with deep recurrent neural networks. Paper presented at the Acoustics, speech and signal processing(ices), 2013 ieee international conference on. (2013).

[5] Hinton, G. E., & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. science,313(5786), 504-507. (2006).

[6] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., &Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. (2012).

[7] Feature extraction for co-occurrence-based cosine similarity score of text documents Ammar Ismael Kadhim ; Yu-N Cheah ; Nurul Hashimah Ahamed ; Lubab A. Salman 2014 IEEE Student Conference on Research and Development.(2014).

[8] LeCun, Y., Bengio, Y., & Hinton, G. Deep learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539. (2015).

[9]Sentence Ranking with the Semantic Link Network in Scientific Paper Jiao Tian ; Mengyun Cao ; Jin Liu ; Xiaoping Sun ; Hai Zhuge 2015 11th International Conference on Semantics, Knowledge and Grids (SKG).(2015).

[10] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Paper presented at the Text summarization branches out: Proceedings of the ACL-04 workshop.

