A SURVEY ON VARIOUS TECHNIQUES TO BUILD EXTRACTIVE TEXT SUMMARIZER

Maitry Desai , Prof. Jatayu Baxi
M.tech scholar, Assistant Professor
Computer Engineering Department
Dharmsinh Desai University
Nadiad,india

***Abstract*:** With significant amount of textual information getting generated every day, text summarization has become essential part of any text analytic tool. Instead of having to go through the entire text, it is convenient to understand the essence of the text by its summary which covers important information from the original document. Summary of the text can be either extractive or abstractive based on the content of summary. This paper focuses more on extractive text summarization. In this paper we first give introduction of the text summarization problem and challenges associated with it. We focus on related work done in the field of extractive text summarization and prepare list of features which can be used to generate summary of the document. We analyze some of the popular techniques to build extractive text summarizer and their trade off. We also discuss recently used deep learning technique for text summarization. We discuss methods to evaluate accuracy of the automatically generated summary and discuss some popular applications where text summarization can be effectively used.

Keywords— Text summarization, Extractive summarization, Abstractive text summarizations

I.INTRODUCTION

Nowadays, there exist lots of data and this rapid growth of data is required to be processed, store and manage. Sometimes, it is difficult to find the exact information from large amount of data that's why we need text summarization. This process reduces the reading time because only a summary needs to be read instead of the entire document. The aim of automatic text summarization is to reduce the source text which preserves content and general meaning of the original text.

Text summarization is a process of extracting or collecting important information from original text and presents that information in the form of summary [12]. In simple words we can say that it is the process of shorting a text document which maintain significant information and general meaning of source text.

Nowadays online information magnification is very high; the task of searching the information for the user has become very tedious and time consuming as there are millions of result generated for one search query. Among these results which identifying the most relevant result for the user is more difficult task. That's why text summarization has become necessary. Instead of having to go through the entire text, it is convenient to understand the text fast and easily by relevant summary of it.

Summaries reduce reading time, it make the selection process easier when researching document. Summarization is to minimize the amount of information you have to go through, before you can understand the overall concepts described in the document. Business leaders, analysts and academic researchers need to comb through huge numbers of documents every day to keep ahead and a large portion of their time is spent just figuring out what document is relevant and what isn't. By creating comprehensive summaries, it's possible to quickly assess whether or not a document is worth reading.

Summary of the text can be generated by two techniques which are Extractive Technique and Abstractive Technique [13]. In abstractive text summarization, the concept of the original text is understood and it is retold in fewer words without changing the meaning of the original text. In Extractive summarization, the summary is generated by using most important sentences from the document exactly as they appear in it. The summary generated by the extractive technique contains those sentences which have highest scores among all the sentences means important sentences of the original document. The importance of sentences is decided on the basis of statistical and linguistic features of sentences in the document.

Extractive summarization can be done by following steps:-1) Pre-processing step and 2) Sentence scoring.

In Pre-Processing step , the original text is represented in structured form. It includes three sub-processes which are sentence boundary, stop-words removal, stemming. Preprocessing phase reduce size of the original text [1].

(a) In Sentence boundary, sentence in the original text is segmented with the help of punctuation mark like '.', '?', '!' etc.
(b) In stop-word removal, Common words with no meaning like "is", "am", "a", "the", "in" etc. are removed.
(c) Stemming -to get the stem or base of every word that emphasize its semantics.

Example of preprocessing step:
He did not try to navigate after the first bold flight, for the reaction had taken something out of his soul. He did take it positively.

Output of sentence segmentation process:
["He did not try to navigate after the first bold flight, for the reaction had taken something out of his soul".
"He did take it positively"]

Output of word segmentation process:
[“He”, “did”, “not”, “try”, “to”, “navigate”, “after”, “the”, “first”, “bold”, “flight”, “,”, “for”, “the”, “reaction”, “had”, “taken”, “something”, “out”, “of”, “his”, “soul”, “.”, “He”, “did”, “take”, “it”, “positively”, “.”]

Output of stop words removal process:
[“try”, “navigate”, “first”, “bold”, “flight”, “,”, “reaction”, “taken”, “something”, “soul”, “.”, “take”, “positively”, “.”]

Output of stemming process:
[“tri”, “navig”, “first”, “bold”, “flight”, “,”, “reaction”, “taken”, “someth”, “soul”, “.”, “take”, “posit”, “.”]

In sentence scoring stage, the various features are decided and calculated to score the sentence. Sentence having higher score are added in summary.

II. LITERATURE SURVEY

This section gives the overview of the research work carried out related to the text summarization. This overview mainly focuses on the extractive text summarization.

Vishal Gupta and Gurpreet Singh Lehal, “A Survey of Text Summarization Extractive techniques”. In this paper they describe various extractive text summarization techniques which consist of two steps pre-processing and processing. Pre-processing step is further divide into sentence segmentation, stop word removal and stemming. In processing step, various features are decided and calculated to score the sentences. Sentences having higher score are included in final summary. Problems with extractive summary are also discussed in this paper [1].

Surajit Karmakar, Tanvi Lad, Hiten Chothani, “A Review Paper on Extractive Techniques of Text Summarization”. In this paper they describe the methods of extractive text summarization[2].

Saranyamol C S ,Sindhu L, “A Survey on Automatic Text Summarization”. In this paper, they describe the various techniques used in automatic text summarization which are extractive text summarization and abstractive text summarization. They also discuss each methods of extractive and abstractive text summarization in brief. They also define the Evaluation parameters such as Precision and Recall[3].

Saiyed Saziyabegum , Priti S. Sajja, “ Literature Review on Extractive Text Summarization Approaches”. In this paper author described types of text summarization. They also give the brief description of various features used to perform extractive summarization and its methods[4].

Neelima Bhatia , Arunima Jaiswal, “Automatic Text Summarization and it’s Methods-A Review”. This paper gives brief description about three types of text summarization which are extractive vs abstractive ,single vs multiple document , Generic vs quey based. It also gives methods of extractive text summarization in brief[5].

Aditya Jain, Divij Bhatia, Manish K Thakur, “Extractive Text Summarization using Word Vector Embedding”. They have proposed an approach to extract a good set of features followed by neural network for extractive summarization. They take Word to vector embedding as a new feature. They use first 284 documents of DUC2002 dataset. They tested the performance of model against some online text summarizers. Autosummarizer got ROUGE-1= 0.33651, ROUGE-2=0.11738, ROUGE-L=0.24874 and their proposed model got ROUGE-1=0.38249, ROUGE-2=0.2256, ROUGE-L=0.27486[6].

Hans Christian1, Mikhael Pramodana Agus, Derwin Suhartono, “Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)”. This research paper explains the use of the algorithm of TF-IDF in an automatic text summarization program[7].

Nabil ALAMI, Mohammed MEKNASSI, Said ALAOUI OUATIK and NourEddine ENNAHNAHI, “Arabic Text Summarization based on Graph Theory” In this paper they have presented a graph-based approach to Arabic text summarization that uses a graph model and ranking algorithm. They compare their model with different summarization system. LexRank got recall=0.646, precision=0.746 and their model got recall=0.718, precision=0.790[8].

Sankar K, Sobha L, “An Approach to Text Summarization”. They propose an efficient text summarization technique that involves two basic operations which are Finding coherent chunks in the document and Ranking the text in the individual coherent chunks and picking the sentences that rank above a given threshold. They use graph theoretic ranking model for text ranking approach. They compare ROUGE score of 567 news articles of DUC 2002 using the proposed algorithm without and with the coherence chunker. Without coherence chunker module got ROUGE-1=0.5103, ROUGE-L=0.4863 and with coherence chunker module got ROUGE-1=0.5312, ROUGE-L= 0.4978[9].

Yogesh Kumar Meena,Dinesh Gopalani , “Feature priority based sentence filtering method for extractive automatic text summarization”. They proposed a feature priority based filtering method for summarization. They use DUC2002 dataset. They have started their summary with the first sentence and ended with last sentence. To select intermediate sentence they have used 3 features,TF-ISF ,named entity presence , proper noun presence[10].

Nikita Desai, Prachi Shah, “automatic text summarization using supervised machine learning technique for Hindi language”.They present an approach to the design an automatic text summarizer for Hindi text that generates a summary by extracting sentences. They use SVM rank tool[11].

III. VARIOUS FEATURES TO SCORE A SENTENCE

The text document is represented by set, D= {S1, S2, - -, Sk} where, Si represent a sentence contained in the document D. The document is subjected to feature extraction. In this section we discuss some features such as Title word, Sentence length, Sentence position, numerical data, Term weight, sentence similarity, existence of Thematic words and proper Nouns etc.

1. **Sentence length Feature:**

This feature provides less weightage to short sentence because short sentences are relatively less important than the longer sentences in the text. Short sentences such as datelines and author names in news articles are less important so, they are not included in summary [14].

2. **Sentence position Feature:**

Usually first and last sentence of first and last paragraph of text document are more informative because sentence in beginning shows the theme of document and last sentence concludes the document [1]. So sentence at these position are included in summary.

3. **Numerical data Feature:**

The sentence containing numerical data like 1, 2, 3 and i, ii, iii.. etc. are consider as important and included in summary [15].

4. **Cue-phrase Feature:**

Phrases such as “in summary”, “It concludes”, “finally”, “In conclusion” etc. are define as cue phrases. These phrases indicate the resulting information about the text. So, sentence start with any of the cue phrase then that sentence consider as important sentences and include in summary [15].

5. **Sentence-to-sentence cohesion Feature:**

This feature use cosine similarity of each sentence to every other sentence. Finding value of this feature for sentence, add up those similarity values [4]. This process repeated for all the sentences. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

6. **Title word similarity Feature:**

Sentence which containing title word or similar to title words are included in summary because title of document define the overall theme of the document [15].

7. **Proper noun:**

Proper noun is name of person, place, thing etc. Sentence containing proper nouns are having more chance to include in summary [4].

8. **Font Feature:**

Sentence which containing words in bold, italic or underlined, upper case or having font size greater than rest of text are consider as important and should be included in summary [1].

9. **Biased Word Feature:**

First define the biased word list. Biased word list is previously defined and contain domain specific words. If sentence containing word from biased word list then that sentence may include in summary [4].

10. **Occurrence of Non – Essential information:**

Some words like “because”, “moreover”, “additionally”, etc. are indicator of non essential information. If Sentence start with this type of word then that sentence may not be included in summary because those sentence usually indicates non-essential information [6].

11. **Mean TF-ISF:**

This feature obtains from the basic feature TF-IDF. For text summarization, this feature termed as Term Frequency-Inverse Sentence Frequency (TF-ISF) where document in TFIDF is similar to sentence in the summarization. The TF-ISF for a word in sentence is computed using equation given below [6].

TF-ISF i,j =TF * ISF.

Where, TF(t)=number of time term t appears in a sentence/total number of term in sentence.

ISF(t)=loge(total number of sentences /number of Sentences with term t in it).

TF-ISF i,j is denotes the TF-ISF of jth word of the ith sentence.

The TF-ISF for a sentence is the mean of the TF-ISF score of the words present in the sentence.

IV.TECHNIQUES USED FOR EXTRACTIVE TEXT SUMMARIZATION

There are several techniques used to generate extractive summary.

1. Sentence feature based technique

In this technique use TF-IDF (Term Frequency-Inverse Document Frequency) feature.

This method use the TF/IDF score of sentence for generate the final summary.

TF-IDF is a numeric statistic which reflects on how important a word is to a document in a collection or corpus.

In this method, first document is preprocessed. Preprocessing step includes the sentence boundary, stop-word removal, stemming. Then calculate TF-IDF score of words with the help of following formula.

TF-IDF =TF * IDF.

TF(t)=number of time term t appears in a document/total number of term in document.

IDF(t)=loge(total number of documents /number of documents with term t in it).

In other words, TF-IDF assigns to term in document that is a

(1) Highest when term occurs many times within a small number of documents.
(2) Lower when the term occurs fewer times in a document, or occurs in many documents.
(3) Lowest when the term occurs in virtually all documents [2].
The importance value of a sentence is a sum of the value of every word in the sentence. Every sentence in the document is sorted in descending order. Then select the sentence with the highest TF-IDF value depending upon compression rate [7].

2. Graph Theoretic Approach

This method is used to determine the theme of the paragraph. After applying preprocessing steps on the document it represented as undirected graph. Sentence of the original document are represented as node of the graph. If any of the sentences share common words, then the nodes in the graph are connected with common edges. The edges with the high cardinality are important sentence. They may be included in final summary [15].

3. Machine Learning method

This method takes summarization problem as classification problem. Sentences are classified as summary sentences and non-summary sentence based on the features that they process. It uses the set of training documents and their extractive summaries in order to generate the summary of the input text. Supervised learning and unsupervised learning algorithms are used in this method. Naïve Bayes, SVM (Support Vector Machine) etc. are the supervised learning algorithms and K-Means, DBSCAN etc. are unsupervised algorithms [11].

4. Fuzzy Approach

This method considers features such as sentence length, title word similarity, sentence location etc. as the input of fuzzy system.
This approach consists of following stages:
1. Preprocessing
2. Feature Extraction
3. Fuzzy logic scoring
4. Sentence selection

In preprocessing, sentence segmentation, stop-word removal, stemming are performed. In Feature extraction, some features are implemented and give fuzzy system as input. After feature extraction each sentence is associated with feature vector. The score for each sentence are derived using fuzzy logic method. The fuzzy logic method uses the fuzzy rules and any membership function. The input membership function fuzzifies each score into three values that is LOW, MEDIUM, HIGH. The important sentences are extracted using IF-THEN rules according to feature criteria. Then apply this fuzzy rule to determine whether sentence is important, average or unimportant [16].
The fuzzy logic system consists of three components: fuzzifier, inference engine, defuzzifier. Crisp inputs are translated into linguistic values using fuzzifier. the inference engine refers to the rule base containing fuzzy IFTHEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier [1].

5. Semantic based approach

This approach uses the LSA method. Latent Semantic Analysis is an algebraic-statistical method that extracts hidden semantic structures of words and sentences. It is an unsupervised approach that does not need any training or external knowledge. LSA uses the context of the input document and extracts information such as which words are used together and which common words are seen in different sentences. Singular Value Decomposition, an algebraic method, is used to find out the interrelations between sentences and words. SVD applied to document word matrices, groups documents that are semantically related to each other, even when they do not share common words. SVD has the capability of noise reduction, which helps to improve accuracy. This approach requires three steps that are Input matrix creation, perform SVD (Singular Value Decomposition) and sentence selection [17]. Advantage of LSA vectors over word vectors is that semantic relations as represented in human brain are captured automatically by LSA vectors while word vectors require explicit methods to originate those semantic relations [15].

6. Cluster based approach

If documents are written for different topics, they are divided into sections either implicitly or explicitly to generate a significant summary. This is known as clustering [15]. Summaries should address different “themes” appearing in the documents. Some summarizers incorporate this aspect through clustering. If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary. Sentence selection is based on cluster Ci. Another factor for selection is location of sentence Li. The last factor that increases the score of a sentence is its similarity to the first sentence in the document to which it belongs (Fi). The overall score (Si) of a sentence i is a weighted sum of the above three factors: Si =W1 *Ci + W2 *Fi+W3 *Li where Si is the score of sentence Ci, Fi and Li and w1, w2 andw3 are the weights for linear combination of the three scores [1].

7. Deep learning approach

This approach is broken down into main three phases: Feature extraction, feature enhancement and summary generation based on values of those features. It can be very difficult to construct high-level, abstract features from raw data. So deep learning is use in second phase to build complex features out of simpler features extracted in the first phase. First pre process the document then feature extraction is done. After that feature enhancement is done using any deep learning algorithms then select most important sentences [18].

V. Evolution measures for the text summarization

1. **Precision:**

Precision (P) is computed as number of sentences occurring in both candidate and reference summaries divided by the number of sentences in the candidate summary [8].

2. **Recall:**

Recall (R) is the number of matched sentences in both candidate and reference summaries divided by the number of sentences in the reference summary [8].

3. **F-score:**

F-score is combination of both precision and recall. It is a harmonic Average of the precision and recall.
F=2*P*R/ P+R [19].

4. **ROUGE:**

ROUGE stand for Recall-Oriented Understudy for Gisting Evalution. It includes measures to auto-matically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. ROUGE-N, ROUGE-S and ROUGE-L are different ROUGE measures [20]. ROUGE-N measures unigram, bigram, trigram and higher order n-gram overlap. ROUGE-L measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. In ROUGE-S, S means Skip-gram occurrence. skip-bigram measures the overlap of word pairs that can have a maximum of two gaps in between words. For Example, ROUGE-1 refers to overlap of unigrams between the system summary and reference summary. ROUGE-2 refers to overlap of bigrams between the system and reference summary.

VI. Conclusion

With lots of information getting generated every day, text summarization is becoming necessary. Text summarization is growing as sub branch of NLP (Natural Language Processing). Text summarization is process of shorting an original text document. It also reduces the reading time. Reviews, previews, summary of news articles, summary of the medical report are example of the text summarization. There are two basic approaches for the text summarization, first one is extractive and second one is abstractive. This survey paper discusses the extractive methods of text summarization. An extractive summary is a selection of important sentences from the original text that briefly describes the original text.

References

[1] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, August 2010.

[2] Surajit Karmakar, Tanvi Lad, Hiten Chothani, "A Review Paper on Extractive Techniques of Text Summarization", International Research Journal of Computer Science (IRJCS), January 2015.

[3] Saranyamol C S,Sindhu L, "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies,2014.

[4] Saiyed Saziyabegum, Priti S. Sajja, " Literature Review on Extractive text summarization Approach", International Journal of Computer Applications, December 2016.

[5] Neelima bhatia, Arunima jaiswal, "Automatic text summarization and it's methods-A Review",IEEE,2016.

[6] Aditya Jain, Divij Bhatia, Manish K Thakur, "Extractive Text Summarization using Word Vector Embedding", International Conference on Machine learning and Data Science, 2017.

[7] Hans Christian; Mikhael Pramodana Agus; Derwin Suhartono, "Single Document Automatic Text Summarization using Term Frequency- Inverse Document Frequency (TF-IDF)", December 2016.

[8] Nabil ALAMI, Mohammed MEKNASSI, Said ALAOUI OUATIK and NourEddine ENNAHNAHI, "Arabic Text Summarization based on Graph Theory", IEEE, 2015.

[9] Sankar K, Sobha L, "An Approach to Text Summarization", Third International Cross Lingual Information Access Workshop, 2009

[10] Yogesh kumar Meena, Dinesh gopalani, "Feature Priority Based Sentence Filtering Method for Extractive Automatic Text Summarization", International Conference on intelligent computing, communication & convergence, 2015.

[11] Desai Nikita, and Prachi Shah. "Automatic Text Summarization Using Supervised Machine Learning Technique for Hindi Langauge." International Journal of Research in Engineering & Technology (2016).

[12] Deepali K. Gaikwad, C.Namrata Mahender, "A Review paper on text summarization", International Journal of Advanced Research in Computer and Communication Engineering, March 2016.

[13] M S Patil, M S Bewoor , S H Patil, "Survey on Extractive Text Summarization Approaches", 2014.

[14] Mr.S.A.Babar, Prof.S.A.Thorat, "Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis", International Journal of Innovative Research in Advanced Engineering, May 2014

[15] Richa Sharma, Prachi Sharma, “A Survey on Extractive Text Summarization”International Journal of Advanced Research in Computer Science and Software Engineering, April 2016.

[16] Mr.s.a.babar, Ms.p.d.patil, “Fuzzy Approach for document summarization”, journal of information, knowledge and research in computer engineering.

[17] Ozsoy, Makbule Gulcin, Ferda Nur Alpaslan, and Ilyas Cicekli. "Text summarization using latent semantic analysis." *Journal of Information Science* 37.4 (2011).

[18] Verma, Sukriti, and Vagisha Nidhi. "Extractive Summarization using Deep Learning."(2017).

[19] https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c.

[20] https://rxnlp.com/how-rouge-works-for-evaluation-of-summarizationtasks/#.XAFne_kzbIV.