A Survey on Extractive Text Summarization Techniques

Diksha Harbola
Department of Computer Engineering
Shankersinh Vaghela Bapu Institute of Technology, Gujarat, India

Abstract— Text summarization is a technique for creating a short but useful information from a text document. Automatic text summarization plays a vital role in producing a summarized information from a large corpus. An extractive summarization is a technique that is based on extracting the sentences from the documents for the summary generation. Various systems are based on feature extraction method. The main goal of all the systems is to obtain relevant data and summarize them as per the users' input. These systems differ in the main aspects of the features provided and the architecture used by them. This paper is consisted of the basics of multi-document summarization, then several approaches for extractive summarization and extractive methods, and then the various evaluation methods for the estimating the performance of the summarization techniques.
Key words: Text Summarization, Graph-Based, Cluster-Based, Tf-Idf, Sentence Extraction, Extractive Summarization, Summary Evaluation

I. INTRODUCTION

Various internet sites are used for fetching the information. Enormous increasing and easy availability of information on the World Wide Web have recently resulted in brushing up the classical linguistics problem - the condensation of information from text documents [3].

Text summarization is the technique that provide significant information from an original document. This technique provides the information that is been extracted as a summarized detail to the users. The objective of automatic text summarization is to condense the origin text into a precise version preserves its report content and global denotation [4].

The main benefit of the summarization technique is that it reduces the time consumption for reading the whole document or information. The user may know the content of the document just by reading the summary. In short, we can say that text summarization provides the abstract for the single document or multi-document.

The text summarization is usually classified as abstractive and extractive summarization. Abstractive summarization is a technique in which the system understand the original text and then itself creates the summary for the document. Extractive summarization technique involves the selection of the some important sentences or features from the one or more document(s) and then combining them to create the summary (may not be in order).

II. EXTRACTIVE SUMMARIZATION FEATURES

In text summarization technique, the key sentences are identified and extracted from the original text document and then are combine together to for a concise summary. Numerous feature discussed [11][12][13] used to include important sentences are:

A. Title word feature:

The sentences consist of words in title of the source document have more chances to be included in final summary as it depicts the theme of the document.

B. Cue Phrase method:

The sentences containing cue words (in conclusion, because, furthermore, therefore) are more likely to be in the final summary. Cue phrases are the words that would affect positively or negatively to the respective sentence weight.to indicate significance of the summary [13].

C. Sentence Location:

The sentences occurring at the starting and in the end of the document are more likely to be in the final summary. It is because of the intuition that most of the document have hierarchical structure with important information in starting and ending of the paragraph.

D. Sentence Length:

Sentence length is the important feature for identifying the key sentences. Neither the too short sentence nor the too long sentence are suitable for the summary. The normalized length of the sentence is calculated as the ratio between a numbers of words in the sentence to the number of words in the longest sentence in the document [4].

E. Term Frequency:

The mostly occurring words increases the score of the sentences. In addition, the sentences containing main keywords are more likely to occur in final summary. Widely used measure for term frequency is TF-IDF.

F. Sentence Similarity:

It indicate the similarity of the sentence with the title of the document or the similarity between two or more sentences.

G. Proper Nouns:

The sentences should have proper noun for the summarization.

H. Proximity:

The distance between text units, where entities occur is a determining factor for establishing relations between entities [13].

I. Paragraph Location:

Similar to the sentence location, the location of the paragraph is also important for the selection of the important sentence.

III. EXTRACTIVE SUMMARIZATION METHODS

Extractive summarization is classified to supervised learning and unsupervised learning methods. Most of the system are built on unsupervised learning methods.

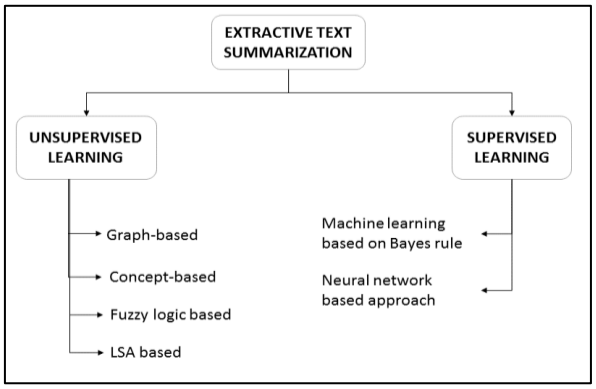

Fig. 1: extractive methods

A. *Supervised Learning Method*

The supervised learning techniques are based on the classification approach at the sentence level where the system learns using the examples to classify between summary and non-summary. The examples in the supervised learning consist of the pair of an input object and the desired output object. The main drawback of the supervised learning method is that it requires known human summary to label the sentences in the original training document enclosed with "summarized sentence" and "non-summarized sentence" and requires more labeled training data for classification [4].

1) *Machine learning based on Bayes rule*

A machine learning method can be objectified by having the collection of the documents and their respective referential summary that are fed as the input for training. Sentences of each document are represented in the form of the vectors of the feature extracted from the text [14].

The sentences are classified as "summary" and "not-summary" based on the features extracted. Naïve Bayes classifier is used for learning from the document. For the given sentence s, the probability of it to be chosen for summary is given as [15],

$$P(s \in S | F_1, F_2, \dots, F_n) = \frac{\prod_{i=1}^{n} P(F_i | s \in S).P(s \in S)}{\prod_{i=1}^{n} P(F_i)}$$

Where, F1, F2,…,Fn are the sentence features for the classification and S is the summary to be generated.

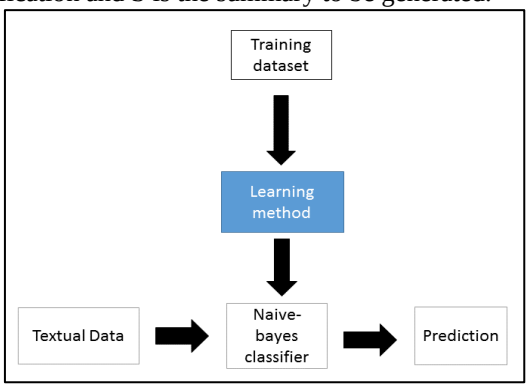

Fig. 2: machine learning approach based on naïve bayes

2) *Neural Network Based Approach*

Artificial neural networks are used for creating the summary of the arbitrary length articles. The network is trained according to the style of the human reader and to which sentences the human reader deems to be important in a paragraph [17]. Input of the neural networks can be either real vector or binary vectors.

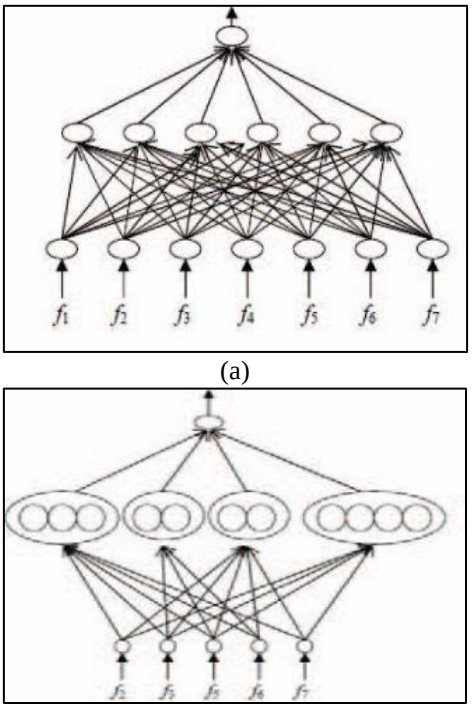

Fig. 3: [17] Neural network after training (a) and after pruning (b)

In the approach proposed in [16], SummaRuNNer outperforms or matches the state-of-art models for the summarization. In addition, it allows interpretable visualization of its decision. Also, includes training mechanism using abstractive summaries for end-to-end training of their extractive model.

B. *Unsupervised Learning Method*

Unsupervised learning technique do not require any human input for deciding the valuable features of the document. Unsupervised summaries provide a higher level of automation compared to supervised model and are more suitable for processing Big Data [4].

1) *Graph-Based Method*

Graph based methods are broadly used for document summarization as graphs can represent the document structure more accurately. TextRank [18] based on the Google's PageRank(Brain and Page, 1998) is a graph based method that consider the local vertex-specific information based on global information recursively drawn from the entire graph. Main idea of this ranking model is 'voting' or 'recommendation' by the vertex to another vertex while linking to it. The votes cast for the vertex determines the score associated with that vertex and the score of the vertices casting vote for it. For considering text as a graph following are the steps involved:

1) Identify the texts units that define the task more appropriately and add them to the vertices of the graph .
2) Identify the relation between the text units and use that relation to draw edges between the vertices.
3) Repeat the ranking algorithm until convergence takes place
4) Sorting vertices based on their last score and then use these sorted vertices for selection purpose

2) Concept-Based Method

Concept based methods extracts the deep concepts from the text from external knowledge base such as Wikipedia [19]. In the proposed method [20], the summarized text are provided for supervision to learn the summary contents distributions from the set of documents. Summary sentences in this method are identified through sLDA.

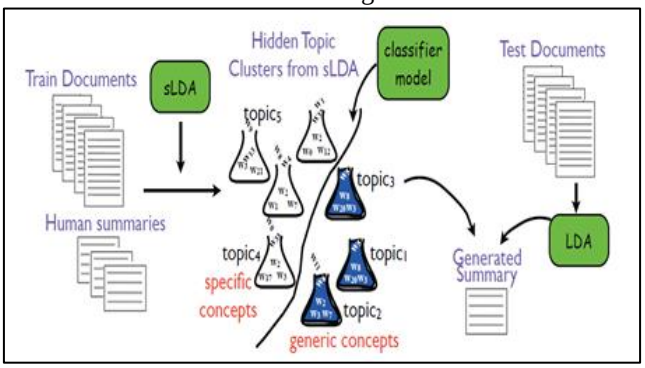

Fig. 4: Framework of the extractive summary generation process using concept-based classification model [20].

3) Fuzzy Logic Based Method

The fuzzy logic approach mainly contains four components: defuzzifier, fuzzifier, fuzzy knowledge base and inference engine [4]. In the proposed method [21], the fuzzy logic was use for sentence extraction to improve the quality of the summary created by the statistic method. Methods to extract the meaningful sentences are text summarization based on general statistic method (GSM) and fuzzy logic method.. Fuzzy logic techniques provide decision-support and expert systems with effective reasoning capabilities. The system consists of following steps [21] :

1) Load the source document into the system
2) In preprocessing step, first each sentences are extracted, then input document is separated into different words, then stopwords are removed and at last word stemming is performed;
3) Each sentence is associated with vector of extraction features, whose values are derived from the content of the sentence;
4) To obtain the sentence score, the features are calculated based on general statistic method (GSM) and fuzzy logic method.
5) A set of highest score sentences are extracted as document summary based on the compression rate.

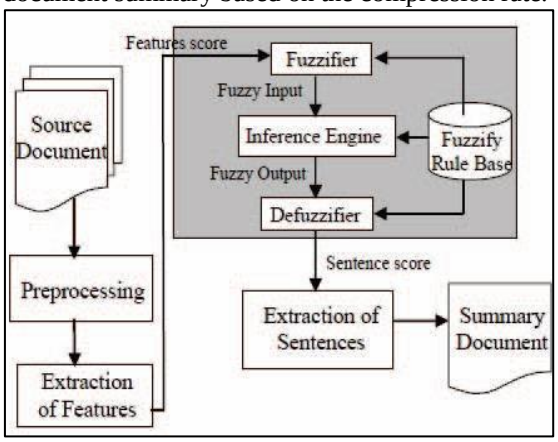

Fig. 5: Overall architecture of text summarization based on fuzzy logic approach proposed in [21]

4) Latent Semantic Analysis based

Latent Semantic Analysis (LSA) is a method, which excerpt hidden semantic structures of sentences and words that are popularly used in text summarization task [4]. LSA takes the text of the input document and extract the sentences with frequently occurring words or the same words from different sentences. The relations between the words in different sentences is depicted by the Singular Value Decomposition [23], that has the capability to reduce the noise to improve the accuracy.

Josef Steinberger [16](2004) listed out the two main disadvantages of LSA:

1) The number of dimensions to be used must be same as to the number of the sentences needed for summary
2) Sentences with high index value will not be chosen although it is suitable for the summary.

The method focuses on the modification in the existing SVD-based summarization by recalculating SVD of a term by sentences matrix.

IV. Evaluation Method

The human specialists usually check the quality of the text by assigning the value from the predefined scale to each summary [5]. The evaluation methods are classified into intrinsic and extrinsic method.

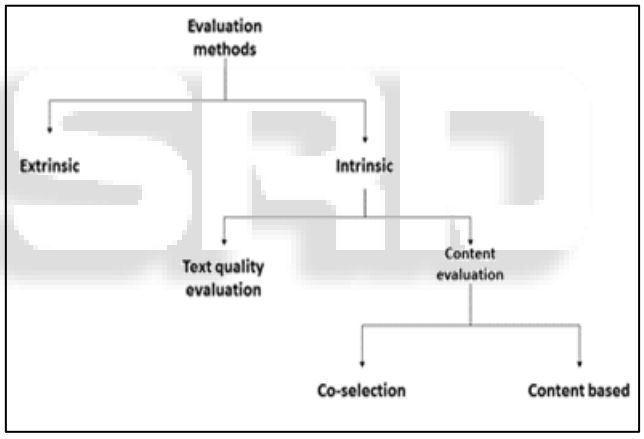

Fig. 6: Evaluation measures

Intrinsic method compares the summary with the ideal or human summary whereas the extrinsic evaluate the summary through some task-based performance, such as, information retrieval.

A. Text Quality Measure:

The text quality have various aspects:-

1) Grammatical: the summary must be free from any incorrect words, wrong grammar or pronunciation errors, or any item that is not a text.
2) Non-redundancy: there must be no redundant text in the summary
3) Reference clarity: in the summary, the noun and pronoun should be clearly referred.
4) Coherence and structure: the summary should be in good format and the sentences should be understandable.
5) This cannot be done automatically, so for this, annotators must assign some marks for summary i.e. form A- very good to E-very poor at DUC2005.

B. Co-selection Measures:

1) Precision, Recall, f-score:

These are the main evaluation measures of the co-selection measures.

Precision is the ratio of the total number of sentences occurring in system and ideal summary to the number of sentences in the system summary.

Recall is the ratio of the number of sentences occurring in system and ideal summary to the number of the sentences in the ideal summary.

f-score is the combination of the recall and precision.

2) Relative Utility:

In this, the model summary represents all sentences of the input document with confidence values for their inclusion in the summary [5].

C. Content based Measures:

Co-selection measure matches only the sentences that are similar in both summary i.e. system and ideal summary. But there may be some sentences that are differently written but have same meaning. Also, the summary written by different annotator may be different and may not share same identical sentences. For this content based measures are used.

1) Cosine Similarity: It is represented as[8]:

$$\cos(X,Y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i {x_i}^2} \cdot \sqrt{\sum_i {y_i}^2}}$$

Where X and Y are the representations of the system summary and its reference document based on the vector model.

2) Unit overlap:

It can be represented as[7]:

$$overlap(X,Y) = \frac{||X \cap Y||}{||X|| + ||Y|| - ||X \cap Y||}$$

Where X and Y are representations based on sets of words or lemmas. ||X|| is the size of set X.

3) Longest Common Subsequence:

It can be represented as:

$$lcs(X,Y) = \frac{len(X) + len(Y) - edit_{di}(X,Y)}{2}$$

where X and Y are representations based on sequences of words or lemmas, lcs(X, Y) is the length of the longest common subsequence between X and Y , len(X) and len(Y) are the length of the string X and string Y respectively, and $edit_{di}(X,Y)$is the edit distance of X and Y [6].

4) Rouge(N-gram co-occurrence Statistics):

Rouge-n score of the system summary is computed as[9]:

$$rouge - n = \frac{\sum_{C \in RSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{C \in RSS} \sum_{gram_n \in C} Count(gram_n)}$$

Where RSS is the referential summary set, $\text{Count}_{\text{match}}$ is the maximum number of n-grams co-occurring in a candidate summary and a reference summary and $\text{Count}_{\text{match}}$ is the number of n-grams in the reference summary

D. Task-Based Measures:

Instead of analysing the sentences in summary, task-based measures the prospect of using summaries for certain task [5]. There are various task-based evaluation measures, some of them are listed as below:

1) Document Categorization:

Document summarization refers to the task of creating document representatives that are smaller in size but retain various characteristics of the original document [10].

2) Information Retrieval:

Information Retrieval systems rank and present documents based on measuring relevance to the user query. But not all the information retrieved are really useful to the user, and it always takes a lot of time to read and select before the user get what he wants[10].

V. Conclusion

The automatic text summarization is playing a vital role in various fields such as biomedical, emails, blogs and customers' reviews. This review paper contains the various methods of the extractive summarization as most of the summarization techniques are based on extractive methods. Extractive summarization provides the information according to the users' input that describes the original document in a small but to the point sentences that may not be in order. This paper contain the comparison of various extractive methods that are used for the summarization. Many extractive methods have evolved but it is difficult to mention which method creates the more concise summary with the high performance. Using the natural processing language (NPL), provides the more precise summary in terms of the semantics.

References

[1] Surajit Karmakar, Tanvi Lad, Hiten Chothani,A Review Paper on Extractive Techniques of Text Summarization, International Research Journal of Computer Science (IRJCS), Issue 1, Volume 2, January 2015, ISSN: 2393-9842

[2] Deepali K. Gaikwad1 and C. Namrata Mahender2 ,A Review Paper on Text Summarization, IJARCCE, Vol. 5, Issue 3, March 2016, ISSN 2278-1021

[3] Jezek K., Steinberger J., Automatic Text Summarization, in Snasel, V. (ed.) Znalosti 2008, pp 1-12. FIIT STU Brarislava, UstavInformatiky a softveroveho inzinierstva (2008) ISBN 978-80-227-2827-0

[4] N.Moratanch ,S.Chitrakala, A Survey on Extractive Text Summarization, communication and signal Processing (ICCCSP), IEEE, June 2017, ISBN 978-1-5090-3716-2

[5] Josef Steinberger, Karel Je˘zek, EVALUATION MEASURES FOR TEXT SUMMARIZATION, Computing and Informatics, Vol. 28, 2009, 1001–1026, V 2009-Mar-2

[6] Radev, D.—Teufel, S.—Saggion, H.—Lam, W.—Blitzer, J.—Qi, H.— Celebi, A.—Liu, D.—Drabek, E.: Evaluation Challenges in Large-Scale Document Summarization. In Proceeding of the 41st meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003.

[7] Saggion, H.—Radev, D.—Teufel, S.—Lam, W.—Strassel, S.: Developing Infrastructure for the Evaluation of Single and Multi-Document Summarization Systems in a Cross-Lingual Environment. In Proceedings of LREC, Las Palmas, Spain, 2002.

[8] Salton, G.: Automatic Text Processing. Addison-Wesley Publishing Company,1988.

[9] Lin, Ch.—Hovy, E.: Automatic Evaluation of Summaries Using n-Gram Co-Occurrence Statistics. In Proceedings of HLT-NAACL, Edmonton, Canada, 2003.
[10] Farshad Kiyoumarsi, Evaluation Of Automatic Text Summarizations Based On Human Summaries, 2nd GLOBAL CONFERENCE on LINGUISTICS and FOREIGN LANGUAGE TEACHING, LINELT-2014, Dubai – United Arab Emirates, December 11 – 13, 2014, Procedia - Social and Behavioral Sciences 192 (2015) 83 – 91
[11] F. Kiyomarsi and F. R. Esfahani, "Optimizing persian text summarization based on fuzzy logic approach," in 2011 International Conference on Intelligent Building and Management, 2011.
[12] F. Chen, K. Han, and G. Chen, "An approach to sentence-selectionbased text summarization," in TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, vol. 1. IEEE, 2002, pp. 489-493.
[13] Khan Atif, Salim Naomie, "A review on abstractive summarization Methods", Journal of Theoretical and Applied Information Technology, 2014, Vol. 59 No. 1.
[14] J. L. Neto, A. A. Freitas, and C. A. Kaestner, "Automatic text summarization using a machine learning approach," in Advances in Artificial Intelligence. Springer, 2002, pp. 205-215.
[15] Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah, "A Review on Automatic Text Summarization Approaches", in Journal of Computer Science, 2016
[16] Ramesh Nallapati, Feifei Zhai∗ , Bowen Zhou,"SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents", November 2016
[17] K. Kaikhah, "Automatic text summarization with neural networks," 2004.
[18] R. Mihalcea and P. Tarau, 'Textrank: Bringing order into texts, " in Proceedings of EMNLP, vol. 4, Barcelona, Spain, 2004.
[19] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh, "Text summarization using wikipedia," Information Processing & Management, vol. 50, no. 3, pp. 443-461, 2014.
[20] Asli Celikyilmaz and Dilek Hakkani-Tur, "concept-based classification for multi-document summarization", Acoustics, Speech and Signal Processing (ICASSP), IEEE, July 2011
[21] Ladda Suanmali , Naomie Salim and Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009
[22] J. Steinberger and K. Jezek, "Using latent semantic analysis in text summarization and summary evaluation," in Proc. ISIM '04, 2004, pp. 93–100.
[23] M. G. Ozsoy, F. N. Alpaslan, and 1. Cicekli, "Text summarization using latent semantic analysis," Journal of Information Science, vol. 37, no. 4, pp. 405-417, 2011.

