Hybrid Text Summarization: A Survey: Proceedings of SoCTA 2017

Chapter · January 2019

DOI: 10.1007/978-981-13-0589-4_7

Hybrid Text Summarization: A Survey

Mahira Kirmani, Nida Manzoor Hakak, Mudasir Mohd and Mohsin Mohd

Abstract Text summarization is the technique of shirking the original text document in such a way that its meaning is not altered. Summarization techniques have become important for information retrieval as large volumes of data are available on Internet and it is impossible for a human to extract relevant information from enormous amount of data in a time-bound situation. Thus, automatic text summarizer is a tool for reducing the information available on Internet by providing nonredundant and salient sentence extracted from a single or multiple text documents. Text summarization has two approaches: extractive and abstractive. Extractive approach generates the summary by selecting subsets of words, sentences, and phrases of text documents whereas abstractive approach understands the main idea of the document and then represents that idea in a natural language using natural language generation technique to create summaries. This paper represents a Survey of Automatic Hybrid Text Summarization.

Keywords Text summarization · Extractive summary · Abstractive · Features Machine learning

1 Introduction

As an enormous amount of data is available over the Internet, the summarization technique is becoming most popular tool for compressing the volume of data in a timely and efficient manner as it is very difficult for humans to summarize such large text documents manually.

M. Kirmani (✉) · N. Manzoor Hakak
Department of CSE, Maharishi Dayanand University, Rohtak, Haryana, India
e-mail: mahira.kirmani@yahoo.com

M. Mohd
Department of Computer Science, Kashmir University, Srinagar, India

M. Mohd
Department of CSE, Kurukshetra University, Kurukshetra, India

K. Ray et al. (eds.), *Soft Computing: Theories and Applications*,
Advances in Intelligent Systems and Computing 742,
https://doi.org/10.1007/978-981-13-0589-4_7

The aim of text summarizer is to compress large text documents into shorter text which includes importance sentences and retains its original meaning [1]. Researchers have put much effort in automatic summarization techniques and have tried different combinations of both statistical and semantic features along with machine learning approaches for generating summaries [2]. Summarization can be done on a single document or multiple documents. This survey tries to find out some of the most relevant techniques used for both single-document and multiple document summarizations.

Text summarization methods can be classified [3–6] into extractive and abstractive summarization. An abstractive summarization [7] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then finds the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information [8] of the original text document.

Extractive summaries [7] are formed by extracting stand out sentences from the text documents and then a generic method is used for scoring these sentences [6]. These sentences are scored based on statistical as well semantic features [9]. Statistically, each sentence is analyzed based on features such as sentence length, sentence to sentence cohesion, cue phrases, verb phrases, etc.

A summary can be employed in an indicative way or informative way [10]. Indicative waypoints to some important parts of the original document while informative way covers all the information of text that is relevant to a summary. Summary generated automatically by certain methods and procedures has following advantages

(1) Summary size can be controlled.
(2) Summary contents are deterministic.
(3) The link between a text element in the summary and its position in the original text can be established earlier.

2 Related Work

Automatic text summarization came into existence in 1950s with the work of Luhn [11]. In his work, he suggested that salient sentences [7] can be identified by the frequency of a particular word in that text document. Sentences containing words whose frequency is more in a particular document are considered as important and must be included in a summary. Other features for extracting salient sentences were highlighted by Edmundson and Wyllys [12]. These were the word frequency, count of title or heading words in a sentence, sentence position, and cue phrases in a document. Yong et al. [13] contributed to the work by developing an automatic text summarization system which integrates the learning ability by combining a statistical approach with a neural network and keywords extraction with unsupervised learning. Further, Alguliev et al. [14] have defined automatic summarization as an interdisciplinary research area of computer science that includes artificial Intelligence, data mining, statistics as well as psychology. Goldstein et al. [15] defined MMR Model which is popularly used to reduce redundancy of sentences in a summary. The MMR (Maximal Marginal Relevance) criteria "strive to reduce redundancy while maintaining query relevance in rearranging retrieved documents and in selecting relevant passages for text summarization". This technique gives better result for Multi-Document Summarization.

Mohd et al. [16] proposed another method for automatic text summarization. In this method, the NLP features and machine learning techniques are used. The process of summarization is divided into three stages, preprocessing, assigning ranks, and postprocessing. Preprocessing involves segmentation, synonym removal, initial ambiguity removal, stop words removal, POS tagging, and word stemming [17]. The second step is divided into two stages, first words are ranked according to different features and then the sentences are ranked by using the features like sentence length, summation of TF-DF, existence of noun and verb phrases, and summation of power factor. The last step of this method is postprocessing which includes four steps for the generation of the final summary

(a) *Sentence extraction*
Every sentence is ranked in this step. The sentence with the highest rank is selected to be included in the summary.

(b) *Dealing with connecting words*
Sentences that contain words like however, although, but, etc., signifies that the meaning of these sentences is incomplete without previous sentences.

(c) *Removing additional information*
When sentences contain words like typically, moreover, additionally, etc. it means the author is giving some extra information which is not important for the summary.

(d) *Using WordNet*
WordNet is used for getting the Synsets of a particular word and Lesk algorithm [18] is employed for word sense disambiguation to get synonym of the word to be replaced. By using above two novel structure generators, a final summary is obtained which is partially abstract.

Neto et al. [10] proposed a method for automatic text summarization using a machine learning approach. The method is divided into three steps. Preprocessing, processing, and generation. In preprocessing step, the original documents are represented in a structured way that helps in generating quality summaries. This also reduces the dimensionality of the representation space; it includes stop word elimination, case folding and stemming. Each sentence is represented in terms of vector and the similarity between sentences is determined by cosine similarity measure. The method uses the concept of reference summary that evaluates the performance of automatic summary objectively using classical and recall measures. Second step that is processing step involves two key points of the method, the first being a set of features used for extracting important sentence and the second is to define a framework for the trainable summarizer and classifier. This method also employs two classical algorithms “Naive Bayes” and “C4.5”. In the last step, final summary is derived from the summary structure.

Cheng [19] proposed a data-driven approach of text summarization based on neural network [13] and continuous sentence features. It is composed of a hierarchical document encoder and an attention-based extractor modeled by recurrent neural networks [20]. Document Encoder peruses the sequence of sentence in continuous space representation. The decoder generates the target sequence from the continuous space representation of original document. A general frame is built for extracting words and sentences. The proposed method employs a Document Reader, Convolutional Sentence Encoder, Recurrent Document Encoder, Sentence Extractor, and Word Extractor. The purpose of document reader is to provide meaningful representation of document from its sentences. Convolutional Sentence Encoder uses a convolutional neural network that is used to obtain representation vectors at sentence level. Several Features are used to compute a list of features that match the dimensionality of a sentence under each kernel width. All these sentences vectors are summed to get the final sentence representation. Recurrent document encoder employs a recurrent neural network that composes of sequence of sentence vectors into document vectors. The RNN used has an LSTM memory activation unit for overcoming vanish gradient problem when training long sequences. Sentence extractor is another recurrent neural network that performs a sequence labeling task for extraction of meaningful sentences. It is used not only to label the sentence that are important to summary but also extracts mutually redundant sentences. Word extractor, instead of labeling the sentence, directly provides the next word that is relevant for the summary. Two datasets were used for training sentence and word-based summarization models. These datasets were taken from DailyMail News. Models were then evaluated on DUC-2002 single-document summarization task.

Sutskever [21] describes a method that uses machine learning model which is extremely significant as it attempts to model high-level abstractions in data. This method uses a multilayered Long Short-Term Memory (LSTM). One LSTM is used to read the input sequence, one time step, at a time, to get a vector representation and then to use another LSTM to extract the output sequence from that vector. It was found that deep LSTM performance was better than shallow LSTM, so the method uses LSTM with four layers. The second LSTM is essentially a recurrent neural network language model except that it is conditioned on the input sequence. In this method, WMT’14 English to French dataset is used. It was discovered that LSTM learned better when the source sentences were reversed.

Yadav et al. [22] proposed an approach for single text document summarization. The approach is extraction based that depends on combination of the statistical and semantic features. Text summarization procedure in this method consists of three steps. In first step, important sentences are extracted by using sentence score, according to linear combination of different features. Some of the features used are aggregate cosine similarity, sentence position, sentiment of sentence, centroid score [23]; [24], and TF-IDF. For finding sentiment score of sentences, the first sentiment of each entity is determined and then the sum of all these entities give the sentiment score of each sentence. Sentence sentiment in this method is either positive or negative. The second step uses an algorithm for redundancy removal. The sentence with the highest rank is added to summary and then the next sentence, whose similarity is less than predefined threshold and length of summary < L, is included in the summary. Evaluation of the summary in this method is done using MEAD, MICROSOFT, and OPINOSIS (Table 1).

3 Extractive Text Summarization Features

(1) *Sentence Length*: Length of a sentence is considered to be important for making a decision regarding which sentence is to be included in the summary and which not.

(2) *Sentence Position*: Importance of a sentence is also determined by its position as suggested by the Edmundson.

(3) *Cue Phrases*: It is one of the important features that help us in extracting useful information from the text document. Edmundson used this feature in 1968. The sentences that begin with phrases such as “in particular”, “significantly”, “surely”, “the best”, “impossible”, etc. can provide useful information to us.

(4) *Noun and Verb phrases*: Sentences containing noun and verb phrases are considered important. These sentences are included in summary as they contain valuable information.

(5) *Similarity to Title*: Similarity of document title is computed by querying the title of document against all sentences of a document. Similarity between title and document is measured by cosine similarity. A sentence that contains title words is considered important for the summary.

(6) *Upper Case*: The sentences of the input text containing words starting with the capital letter are considered important for the summary. This feature is not applicable for Hindi text summarization.

(7) *Centroid feature*: Similarity between set of words in a cluster is determined and based on its value importance of sentence is determined.

(8) *Sentence to Sentence cohesion*: This feature allows us to identify sentences that are important for a summary producer. First, the similarity between each sentence “S” and every other sentence “s” of a document is computed and then these similarity values are added up to generate the final value of this feature for the sentences.

(9) *Similarity to Keywords*: Sentences containing keywords are considered important. Similarity between keywords and sentences is obtained by using a query against all sentences of a document [25].

(10) *Sentence to Centroid Cohesion*: This feature is obtained for a sentence as follows: first, we compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then we compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence.

Table 1 Comparison of work done by researchers

Researchers	Input	Methods	Datasets	Results
Padma Priya G. et al.	Multiple documents	RBM and Deep Learning Algorithm	Multiple documents from each of the different domains are collected and processed	The maximum recall values marked for the existing approach is 0.72, while for the proposed approach it Comes around 0.62
Jianpeng Cheng et al.	Single document	Data-driven approach based on neural networks and continuous sentence features	Two Datasets created from DailyMail News, Models Extracted on DOC 2002	The NN-SE outperforms the LEAD and LREG baselines with a significant margin, while performing slightly better than the ILP model
Chendra Shekhar Yadav et al.	Single document	Combination of Statistical and Semantic technique	Self Designed (Taken from various newspapers)	Summary length is nearly 27% but high precision, and F-score w.r.t. MEAD, reference summary and high recall w.r.t Microsoft-generated summary
Arlay Barrera et al.	Single Document	POS, Name entity recognition, Text Rank Word Extraction, Sense Learner	Two datasets used DUC 2002 and Scientific Magazine article set	Their system outperforms MEAD and Text Rank sentence extraction in all experiments and is consistently higher than the baseline. Its ROUGE scores are also statistically higher than the baseline for the scientific magazine article set

(continued)

Table 1 (continued)

Researchers	Input	Methods	Datasets	Results
Joel Larocca Neto et al.	Multiple documents	Two machine learning algorithms used: Naïve Bayes and C4.5	TIPSTER Document base Data used	The values of precision and recall for all the methods are significantly higher with the rate of 20% with the compression rate of 10%

(a) Mean TF-ISF
It is a feature that computes the importance of a word in whole document. The importance increases with the number of times a word appears in a sentence (TF) but is balanced by the frequency of the term in the document.

(b) *Power Factor*
Sentences which contain important information are written in Capital, Bold or Italics. This feature helps in extraction of such important sentences.

(c) Biased words
Biased words are those words which are previously defined and may contain substantial information. If a word appearing in a sentence is from biased word list, then the sentence is important and should be included in the summary

(11) *Sentiment features*
Sentiments play a vital role in extracting important sentences from a document. Thus, sentences with emotional content are important to an author and should be included in the summary. Several emotion classes are used such as (positive, negative, fear, joy, surprise, hate, disgust) as seed words which are used to identify emotions in text.

4 Comparison of Different Types of Summarization

See Table 2.

5 Techniques Used in Automatic Summarization

(1) Term Frequency-Inverse Document Frequency (TF-IDF) Method
Term frequency (TF) measures how frequently a term occurs in a document

Table 2 Types of summarization

Types of summarization	Description
Single document	Summary extracted from a single document
Multiple document	Multiple documents are used for extracting a summary
Extractive	Important sentences are determined from the document
Abstractive	Develops main idea of document and expresses that idea in natural language
Indicative	Only some important parts of the text are covered that gives a main idea to the user
Informative	All relevant information of text is covered
Generic	Summary are generated containing main topics of original document
Query based	Generates summaries that contain sentences that are relevant to the given queries
Domain dependent	Knowledge of domain and text structure is needed
Domain independent	Does not require prior knowledge of text structure and can accept any type of text

$$\left\{\left[\mathrm{Tf(t)} = (\text{No of times term t appears in a document}) \big/ (\text{total no of terms in a document})\right]\right\}$$

Inverse document frequency (IDF) measures how important a term is, while computing TF, all terms are considered equally important. However, it is necessary to weigh down the frequency of those terms that are not so important for the document.

$$\left\{\mathrm{IDF(t)} = \mathrm{loge}(\text{total no of documents}) \big/ (\text{no of documents with term t in it})\right\}$$

The method is used to identify the importance of sentences. Based on the score of TF-IDF, sentences that get highest score are included in summary.

(2) Cluster-Based Method
Normally, if documents are written for different topics, they are divided into sections either implicitly or explicitly to generate a significant summary. This aspect is known as clustering. The overall score of a sentence is calculated as the weighted sum of three factors: similarity of the sentence to the theme of a particular cluster, location of the sentence in the document, and similarity of the sentence to the first sentence in the document to which it belongs.

(3) Machine Learning Method
Machine learning approach takes training dataset as input and uses their extractive summaries for generating the summary of the original text document. The process of summarization is modeled as classification problem. The classification is obtained by application of a trainable machine learning algorithm on the document. Therefore, the summarization task in machine learning method can be envisaged as a "two" class classification problem. If sentence belongs to the extractive reference summary then these sentences are marked as correct and

if not then they are marked as incorrect. Training data is used to learn the pattern (classification) of sentences which is based on classification problem that results in the generation of summaries. Summarization based on the application of trainable machine learning algorithm [2, 23] uses set of features directly extracted by the original text.

(4) Text Summarization with Neural Networks
Using neural networks [13] for text summarization has become a popular tool for generating automatic summaries. Neural networks represent the human brain structure and these artificial structures are trained to learn types of sentences which are important for summary. Neural network consists of multiple layers and single path traversal, from front to back. The goal of the neural network is to solve problems in the same way as the human brain would. Neural networks are based on real numbers, with the value of the core and of the axon typically being a representation between 0.0 and 1. Neural networks are fed with some training data which includes information about which sentence should be included and which should not be included, human experts are employed for this work. The neural networks learn from the pattern of the sentences in this training data.

(5) *Query-based extractive text summarization*
Here, sentences are extracted by using a query based approach. Features like similarity to sentence, similarity to keywords, and similarity to title use the query based extraction are used for identifying important sentences in the text document. In this method, the title or keywords of the document is used as query against all sentences. Sentences which contain query words are ranked and based on this ranking sentences are included in the summary. Cosine similarity measure is used to compute similarity between document title or keyword and sentences.

6 Conclusion

Text summarization has become a main source of interest for researchers from past few years due to gigantic amount of information available on Internet. It is impossible for humans to manually summarize such a large amount of information and get precise and meaningful summaries in less time. Hence, automatic text summarizer is needed to get the job done in less time. This survey paper has presented all the features and methods used by researchers over years for extracting a proper and meaningful summary. Text summarization is classified into three categories Extractive, Abstractive, and Hybrid summarization. All the three categories have been used for generating summaries. Some of the machine learning approaches combined with certain features have also been covered in this paper. Features can be statistical and semantic. In future, we can combine deep learning, artificial neural network along with semantic feature which can provide us quality summaries that will be less redundant, processing time will be less and such summaries will be more human oriented.

References

1. Dalal, V., Malik, L.G.: Proceedings of Emerging Trends in Engineering and Technology (ICETET), 2013 6th International Conference on Emerging Trends in Engineering and Technology
2. Yadav, C.S., Sharan, A.: Hybrid approach for single text document summarization using statistical and sentiment features, Jawaharlal Nehru University, Delhi, India. Int. J. Inf. Retr. Res. (IJIRR) **5**(4), 46–70 (2015)
3. Das, D., Martins, A.F.: A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU, vol. 4, pp. 192–195 (2007)
4. Amini, M.R., Usunier, N., Gallinari, P.: Automatic text summarization based on word-clusters and ranking algorithms. In: Proceedings of the 25th ACM SIGIR, pp. 105–112 (2002)
5. Patil, V., Krishnamoorthy, M., Oke, P., Kiruthika, M.: A statistical approach for document summarization. Department of Computer Engineering Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai, Maharashtra, India. Int. J. Adv. Comput. Technol. (IJACT). ISSN 2319-7900
6. Gupta, V.: A survey of text summarization extractive techniques. J. Emerg. Technol. Web Intell. **2**(3), 258–268 (2010)
7. Ren, F.: Automatic abstracting important sentences. Int. J. Inf. Technol. Decis. Making **4**(1), 141–152 (2005)
8. Alguliev, R.M., Aliguliyev, R.M.: Effective summarization method of text documents. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), pp. 1–8 (2005)
9. Gupta, V., Lehal, G.S.: A survey of text summarization extractive techniques. J. Emerg. Technol. Web Intell. **2**(3) (2010)
10. Neto, J.L., Freitas, A.A., Kaestner, C.A.A.: Automatic text summarization using a machine learning approach. In: Advances in Artificial Intelligence: Lecture Notes in Computer Science, vol. 2507, pp. 205–215. Springer, Berlin, Heidelberg (2002)
11. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**, 159–165 (1958); Baxendale, P.B.: Machine-made index for technical literature: an experiment. IBM J. Res. Dev. **2**, 354–361 (1958)
12. Edmundson, H., Wyllys, R.: Automatic abstracting and indexing—survey and recommendations. Commun. ACM **4**(5), 226–234 (1961)
13. Yong, S.P., Abidin, A.I.Z., Chen, Y.Y.: A neural based text summarization system. Int. J. Eng. Trends Technol. (IJETT) (2005). ISSN 2231-5381
14. Alguliev, R.M., Aliguliyev, R.M., Hajirahimova, M.S., Mehdiyev, C.A.: MCMR: maximum coverage and minimum redundant text summarization model. Expert Syst. Appl. **38**(12), 14514–14522 (2011). https://doi.org/10.1016/j.eswa.2011.05.033
15. Goldstein, J., Mittal, V., Carbonell, J., Callan, J.: Creating and evaluating multidocument sentence extract summaries. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, pp. 165–172. ACM (2000). https://doi.org/10.1145/354756.354815
16. Mohd, M., Shah, M.B., Bhat, S.A., Kawa, U.B., Khanday, H.A., Wani, A.H., Wani, M.A., Hashmy, R.: Sumdoc a unified approach for automatic text summarization. In: Fifth International Conference on Soft Computing for Problem Solving, SocProS 2015, At Indian Institute of Technology Roorkee, vol. 1. Springer
17. Porter, M.F.: An algorithm for suffix stripping. Program **14**, 130–137 (1980). Reprinted in: Sparck-Jones, K.
18. Banerjee, S.: Adapting the Lesk algorithm for word sense disambiguation to WordNet. In: Proceeding of the Third International Conference on Linguistic and Intelligent Text Processing, pp. 136–145
19. Cheng, J.: Neural summarization by extracting sentences and words. In: ACL2016 Conference Paper. arXiv:1603.07252v3 [cs.CL]. Accessed 1 July 2016

20. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Proceedings NIPS'15 of the 28th International Conference of Neural Information Processing System, pp. 1171–1179 (2015)
21. Sutskever, I.: Sequence to sequence learning with neural networks. In: Proceeding NIPS; 14 of the 27th International Conference on Neural Information Proceeding Systems, pp. 3104–3112
22. Yadav, C.S., Sharan, A., Kumar, R., Biswas, P.: A new approach for single text document summarization. In: Second International Conference on Computer and Communication Technologies, pp. 401–411
23. Fattah, M.A., Ren, F.: Automatic text summarization. Proc. World Acad. Sci. Eng. Technol. **27**, 192–195 (2008). ISSN 13076884
24. Radev, D.R., Jing, H., Stys, M., Tam, D.: Centroid-based summarization of multiple documents. Inf. Process. Manage. **40**(6), 919–938 (2004)
25. Tonelli, S., Planta, M.: Matching documents and summaries uses key-concepts. In: Proceedings of the French Text Mining and Evaluation Workshop, pp. 1–6 (2011)