International Journal of Technical Innovation in
Morden Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2584
Volume 3, Issue 4, April-2017

Automatic Text Summarization of Single Document : A Survey

Bijal Patel[1], Prof. Tejal Patel [2]

PG Scholar[1], Assistant Professor[2]

Department Of IT, G.H. Patel Collage of Engineering and Technology,
V.V. Nagar, Anand-388120

ABSTRACT

The main objective of a text summarization system is to identify the most important information from the given text and present it to the end users. In the fast-moving world, it's difficult to read all the text-content. Hence, the need for text summarization is being in the spotlight. Automatic text summarization is a technique which compresses large text to a shorter text which includes the important information. In this paper, different text summarization methods are introduced in a single document. The advantages and drawbacks of the summarized methods have also been discussed.

Keywords
Automatic Text Summarization, Text mining, Single Document summarization.

I. INTRODUCTION

Automatic Text summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document and present it to the end users. Text summarization finds the importance because of its variety of applications like summaries of newspaper articles, book, News, resume, books, music, and magazine, stories on the same topic, event, scientific paper, weather forecast, stock market, plays, film and speech. The output of the summary can be of two types: Extractive summaries and abstractive summaries [1].

Automatic Text summarization has two types: Extractive summaries and Abstractive summaries. Extractive summaries are selecting a subset of existing words, phrases, or sentences in the original text to form the summary. Abstractive summaries build an internal semantic representation and to create a summary that is closer to what a human might generate [2].

Summaries can also be of two types: generic or query-focused. Topic-focused or user-focused summaries are the other names for query-focused summaries. Such a summary includes the query related content, whereas a general sense of the information presented in the document is provided in a generic summary [2].

There are three kinds of summaries on the basis of language: multilingual, Monolingual and Cross-lingual summaries. The monolingual summarization system is when the language of the source and target document is same. When source document is in a number of languages like English, Hindi, Punjabi and summary is also generated in these languages, then it is termed as a multi-lingual summarization system. If the source document is in English and the summary generated is in Hindi or any other language other than English, then it is known as a cross-lingual summarization system [2].

II. LITERARTURE SURVEY

Researchers have been working on text summarization within the Natural Language Processing (NLP) and Information Retrieval (IR) to create a better and more efficient summary.

Babara and Patil[1] focus on the Fuzzy logic Extraction approach for text summarization and the semantic approach of text summarization using Latent Semantic Analysis. In that proposed method improves the quality of summary by the latent semantic analysis into the sentence feature extracted fuzzy logic system to extract the semantic relations between concepts in the original text.

In [2], Extractive text summarization has been presented by applying score to score the sentences using different text features. They have used Wikipedia articles as input. In this paper, Author also described the preprocessing steps of text summarization.

In [3], Author presented a taxonomy of text summarization methods and also described brief introduction of methods of single document and multiple document and stated the current state of art.

In [4], The number of sentences of proposed summary would be equal to the number of paragraphs present in a given document. This is by using successive threshold. Author gives satisfying results when compared with commercial online Summarizer and Microsoft Summarizer uses a successive threshold approach.

In[5], using higher-order singular value decomposition (HOSVD) for extracting the concept of the words from a word-document and then select important sentences with more cosine similarity to this concept and Using a Word Net-based Semantic Similarity to calculate similarity between ranked sentences and redundancy elimination. In this paper, The experimental results show that their approach can out perform other state-of-the-art summarization approaches.

In[6],Author was Compared the manually-produced summaries that generated by experts and the automatically-produced summaries. Using Fuzzy method and Vector approach, author analyzes that automatically generated summaries can be produced much faster than human summaries and they are more economical, more appropriate and more efficient.

Kupiec, et al.[7] conducted an investigation on the use of Genetic Programming (GP) for solving the problem of Automatic Text Summarization. The paper also showed that GP can produce the summaries that have better quality than a number of statistical methods and also better than non-expert and less experienced human expert.

In[8], Author used Naïve-Bayes method for text summarization and evaluated method for number of applications. This method can be very efficient, so author used this method for many applications and recommended the features.

In[9], Ranking Algorithm uses of the neural network that ranks the sentences on the basis of weighting parameter. Author also used third party data sets to enhance sentence text features with statistical significance.

In[10], the author claims that existing approaches to summarization have always assumed feature independence. Also used log-linear models to remove assumptions and showed that the model produced better extracts than a naive-Bayes model.

Hidden Markov Model is used to extract the sentence from the document in [11]. In this paper, Author used extractive Summarization that extracts the sentences directly from the text document. This method is very easy to understand and efficient for any application.

In [12, 13], Author focused on the Sentence position that is very important of a single feature. In Text, "position method" that is just weighing a sentence by its position, arises from the idea that texts generally follow a predictable discourse structure.

III. Taxonomy of Automatic Text Summarization Techniques:

Automatic Text Summarization can be classified into single document text summarization and multi document text summarization (Fig 1)[3].

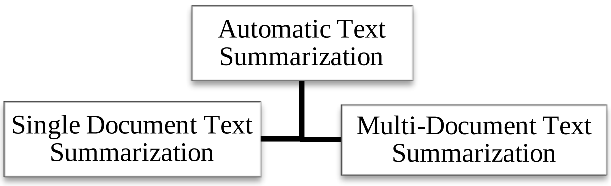

Automatic Text Summarization
Single Document Text Summarization
Multi-Document Text Summarization

Fig 1: Automatic Text Summarization models

Single-Document Summarization: Single-Document Summarization is extracting the information from a single text document. The biggest challenge in summarization is to identify or generalize the most important and informative sentences from a document because the information in the document is non-uniform usually [1].Multi-Document Summarization: Multi-Document Summarization is the extraction of information from multiple texts written about the same topic.

The flow of information in a given document is not uniform, which means that some parts are more important than others in Single-Document Summarization. The major challenge in Single-Document summarization lies in distinguishing the more informative parts of a document from the less ones. Though there have been instances of research describing the automatic creation of abstracts.

Automatic Text Summarization methods for single document (Fig 2) are listed below:

Naïve-Bayes Method [8]: Naïve-Bayes Method can be trained very efficiently in a supervised learning setting. In many practical applications, one can work with the naïve-Bayes model without accepting Bayesian probability or using any other Bayesian methods.

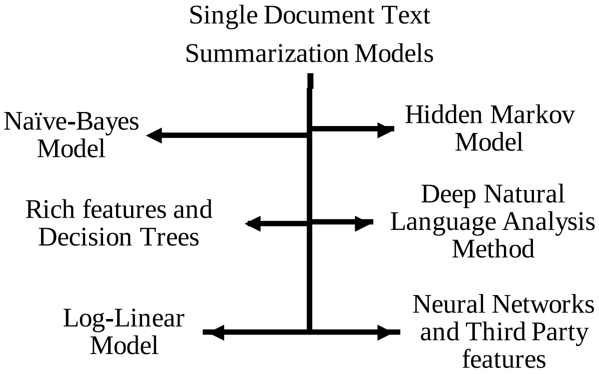

Single Document Text
Summarization Models
Naïve-Bayes Model
Hidden Markov Model
Rich features and Decision Trees
Deep Natural Language Analysis Method
Log-Linear Model
Neural Networks and Third Party features

Fig 2: Single Document Text Summarization methodologies

Rich Features and Decision Trees [12, 13]: The importance of a single feature, sentence position is must there. In Text, "position method" that is just weighing a sentence by its position, arises from the idea that texts generally follow a predictable discourse structure.

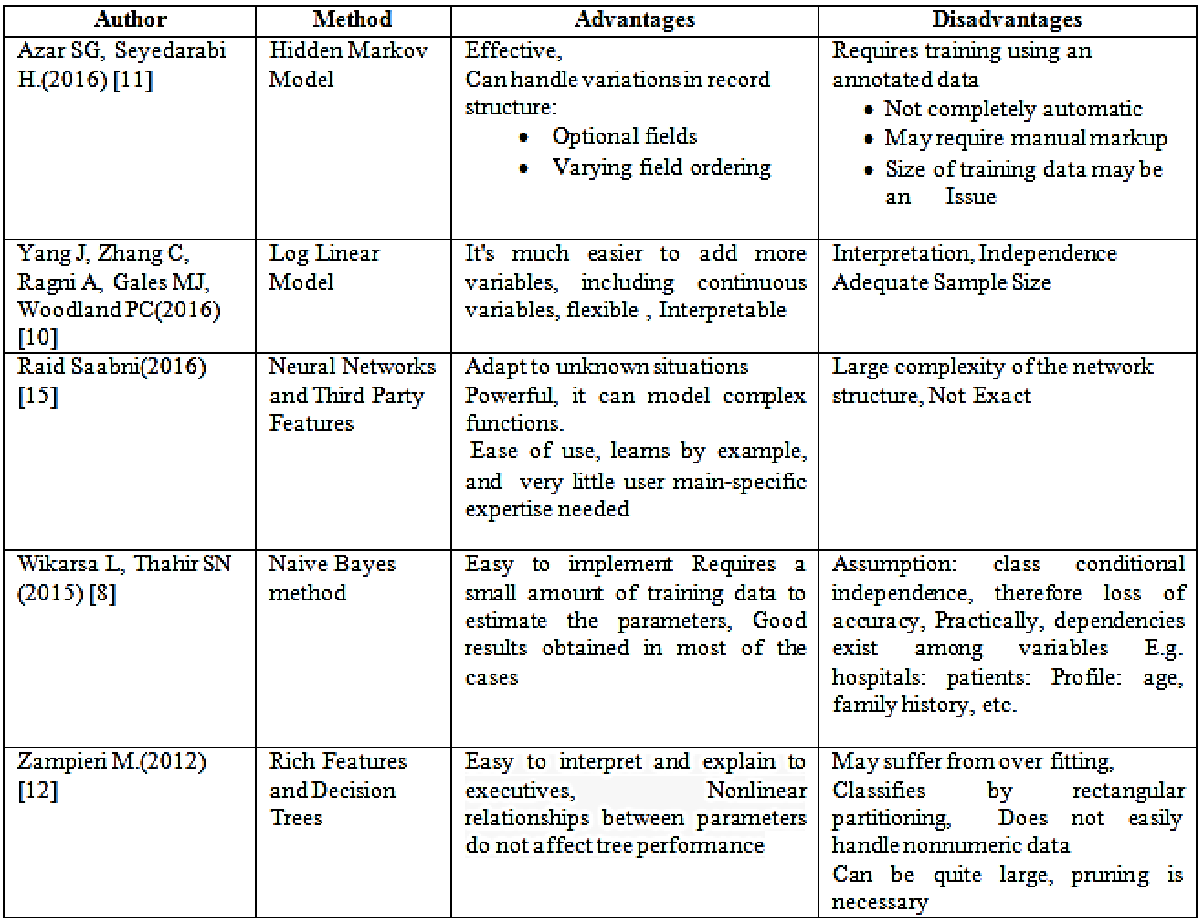
Deep Natural Language Analysis Method [15]: This method is used for to extract meaning from natural language text in machine readable form. Deep linguistic processing is also essential to the creation of natural language dialogue systems, which allow computers to understand and reply in natural language which is understandable by humans.

Neural Networks and Third Party Features [9]: Neural network ranking algorithm and third party datasets to enhance sentence features with statistical significance to tackle the problem of extractive summarization.

Log Linear Models [10]: The author claims that existing approaches to summarization have always assumed feature independence. The author used log-linear models to remove this assumption and showed that the system produced better extracts than a naive-Bayes model.

Hidden Markov Models [11]: Hidden Markov Model is used for to extract the sentence from the document. This method is useful for text summarization of single documents.

Table 1 List of Methods of Text summarization of Single Document



Author	Method	Advantages	Disadvantages
Azar SG, Seyedarabi H.(2016) [11]	Hidden Markov Model	Effective, Can handle variations in record structure: • Optional fields • Varying field ordering	Requires training using an annotated data • Not completely automatic • May require manual markup • Size of training data may be an Issue
Yang J, Zhang C, Ragni A, Gales MJ, Woodland PC(2016) [10]	Log Linear Model	It's much easier to add more variables, including continuous variables, flexible , Interpretable	Interpretation, Independence Adequate Sample Size
Raid Saabni(2016) [15]	Neural Networks and Third Party Features	Adapt to unknown situations Powerful, it can model complex functions. Ease of use, learns by example, and very little user main-specific expertise needed	Large complexity of the network structure, Not Exact
Wikarsa L, Thahir SN (2015) [8]	Naive Bayes method	Easy to implement Requires a small amount of training data to estimate the parameters, Good results obtained in most of the cases	Assumption: class conditional independence, therefore loss of accuracy, Practically, dependencies exist among variables E.g. hospitals: patients: Profile: age, family history, etc.
Zampieri M.(2012) [12]	Rich Features and Decision Trees	Easy to interpret and explain to executives, Nonlinear relationships between parameters do not affect tree performance	May suffer from over fitting, Classifies by rectangular partitioning, Does not easily handle nonnumeric data Can be quite large, pruning is necessary

IV. CONCLUSION

In this paper, a brief introduction of various methods of Automatic Text Summarization has been described. All methods that are introduced of text summarization are in Single Document that only applies to a single document. This paper described the advantages and disadvantages of the methods.

REFERENCES

[1] S.A.Babara and PallaviD.Patil, Improving Performance of Text Summarization, International Conference on Information and Communication Technologies (ICICT 2014),2015 Elsevier

[2] Hingu D, Shah D, Udmale SS. Automatic text summarization of Wikipedia articles. In Communication, Information & Computing Technology (ICCICT), 2015 International Conference on 2015 Jan 15 (pp. 1-4). IEEE.

[3] Suneetha S. Automatic text summarization: The current state of the art. International Journal of Science and Advanced Technology. 2011 Nov;1(9):283-93.

[4] PadmaLahari E, Kumar DS, Prasad S. Automatic text summarization with statistical and linguistic features using successive thresholds. In Advanced Communication, Control and Computing Technologies (ICACCCT), 2014 International Conference on 2014 May 8 (pp. 1519-1524) IEEE.

[5] Biyabangard A, Abadeh MS. Word concept extraction using HOSVD for automatic text summarization. InAI& Robotics (IRANOPEN), 2015 Apr 12 (pp. 1-6) IEEE.

[6] Uy NQ, Anh PT, Doan TC, Hoai NX. A study on the use of genetic programming for automatic text summarization. In Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on 2012 Aug 17 (pp. 93-98). IEEE.

[7] Kupiec J, Pedersen J, Chen F. A trainable document Summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval 1995 Jul 1 (pp. 68-73). ACM.1995

[8] Wikarsa L, Thahir SN. A text mining application of emotion classifications of Twitter's users using the Naive Bayes method. In2015 1st International Conference on Wireless and Telematics (ICWT) 2015 Nov 17 (pp. 1-6). IEEE.

[9] Raid Saabni, Recognizing handwritten Single Digits and Digit Strings Using Deep Architecture of Neural Networks. In Artificial Intelligence and Pattern Recognition (AIPR), International Conference on 2016 Sep (pp. 1-6). IEEE.

[10] Yang J, Zhang C, Ragni A, Gales MJ, Woodland PC. System combination with log-linear models. In2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 Mar 20 (pp. 5675-5679) IEEE.

[11] Azar SG, Seyedarabi H. Continuous Hidden Markov Model based dynamic Persian Sign Language recognition. InElectrical Engineering (ICEE), 2016 24th Iranian Conference on 2016 May (pp. 1107-1112) IEEE.

[12] Zampieri M. Evaluating Knowledge-poor and Knowledge-rich features in automatic classification: A case study in WSD. In Computational Intelligence and Informatics (CINTI), 2012 IEEE 13th International Symposium on 2012 Nov 20 (pp. 359-363). IEEE.

[13] Ochotorena CN, Yap CA, Dadios E, Sybingco E. Robust stock trading using fuzzy decision trees. In2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) 2012 Mar 29 (pp. 1-8). IEEE.

[14] www.pce.uw.edu/courses/deep-processing-techniques-for-natural-language-processing