Literature Review on Text Summarization

Parth Valani
Computer Science
Lakehead University 1116576
Thunder Bay, Canada
pvalani@lakeheadu.ca

Birva Patel
Computer Science
Lakehead University 1111092
Thunder Bay, Canada
bpatel24@lakeheadu.ca

Jaykumar Nariya
Computer Science
Lakehead University 1116571
Thunder Bay, Canada
jnariya@lakeheadu.ca

***Abstract*—In today's world, There is plethora of information or data available online and most of it is in text form. When We try to retrieve that information from the database, to read that much amount of data is very trivial task. Therefore, Automatic Text summarization came into demand. Manual creation of the summary is very complicated task. Text summarization means generate the useful information or gives the main idea of overall data as an output. This paper depicts survey of recent evaluations which took place in the field of text summarization.**

***Index Terms*—Text Summarization, Natural language processing, Abstrative summary, Extractive summary.**

I. Introduction

In the following review of literature confirms the researches in the field of text summarization over the years and techniques used to summarize the text or document. Research in the field of text summarization began in the 1950s and until now there is no system that can produce summaries such as professionals or humans. This paper aims to identify and analyze methods, datasets and trends in automatic text summarization research from 2015 to 2019.

There are approaches which are helpful to generate a summary – extraction and abstraction. Extraction is domain independent and picks up the important sentences and gives a summary while on the other hand, abstraction is domain dependent and takes the human knowledge by understanding the whole text and prepares a goal and produce a summary. Summarization is of two types.

1) Single document text summarization
2) Multi-document text summarization

The paper organization is as follows. Section 2 describes the related work done by the pioneers in summarization. Section 3 provides the classification among the methods used by single and multidocument summarization. Section 4 includes the challenges to summarize the text and Section 5 concludes this paper.

II. Early Work

The earliest work on the summarization of a single document concentrated on technical documents. Perhaps the most cited paper on description is the one from (Luhn, 1958), which outlines research done at IBM during the 1950s [1]. Throughout his thesis, Luhn proposed that the frequency of a given word in an article would provide a useful measure of its significance.There are several primary ideas put forward in this paper that have taken on the value of summarization in later work. Words were truncated into their root forms as a first step, and stop words were deleted. Luhn then compiled a list of words of content sorted by decreasing frequency, the index giving a meaning measure of the word.On a sentence level, a significance factor was derived that reflects the number of occurrences of significant words within a sentence, and the linear distance between them due to the intervention of non-significant words. All sentences are ranked in order of their significance factor, and the top ranking sentences are finally selected to form the auto-abstract.

Baxendale in 1958 focused on sentence position to find the salient features [2]. He took 200 paragraphs and examined that in 85% of paragraphs topic sentences are placed in the beginning while in rest 7% he found, it occurred in the last.

Edmundson(1969) describes a system which produces excerpts from documents. His primary contribution was to develop a typical structure for an experiment with extractive summarisation. [3]. At first, the author developed a protocol for creating manual extracts, that was applied in a set of 400 technical documents. From the previous two books, the two features of word frequency and positional value were integrated [3]. Two other features were used: the presence of cue words (presence of words like significant, or hardly), and the skeleton of the document (whether the sentence is a title or heading). Each of these features had weights added manually to score growing sentence. During evaluation, it was found that about 44% of the auto-extracts matched the manual extracts [4].

III. Techniques of Summarization

There are different techniques of text summarization. All the broadly classified techniques can be grouped into following two groups.

1) Extraction Based Summarization Techniques:
   Extractive summarization is extraction based summarization whose summary consists entirely of extracted content. Originally, the research focused on document management strategies with various approaches, such as sentence location or word frequency in text [2]. The experiment was then carried out using the Information Extraction (IE) Extraction technique for automatic summarizing on the grounds of increased accuracy and more specific results. [5]A system is built which takes

information extraction for automatic summation, called RIPTIDES, which works for news summary based on the user's chosen scenario template.

In extraction based summarization the sentences and keywords are selected that can be included in the summary. The sentences and words are selected from the main text. In this, the main text is divided into the sentences or words and select and reject them according to significance score. The selected sentences and words are included in the final summary.The extract field is more well-researched, in contrast to abstracts which have more challenging problems and require extensive natural language processing

2) Abstraction Based Summarization Techniques:

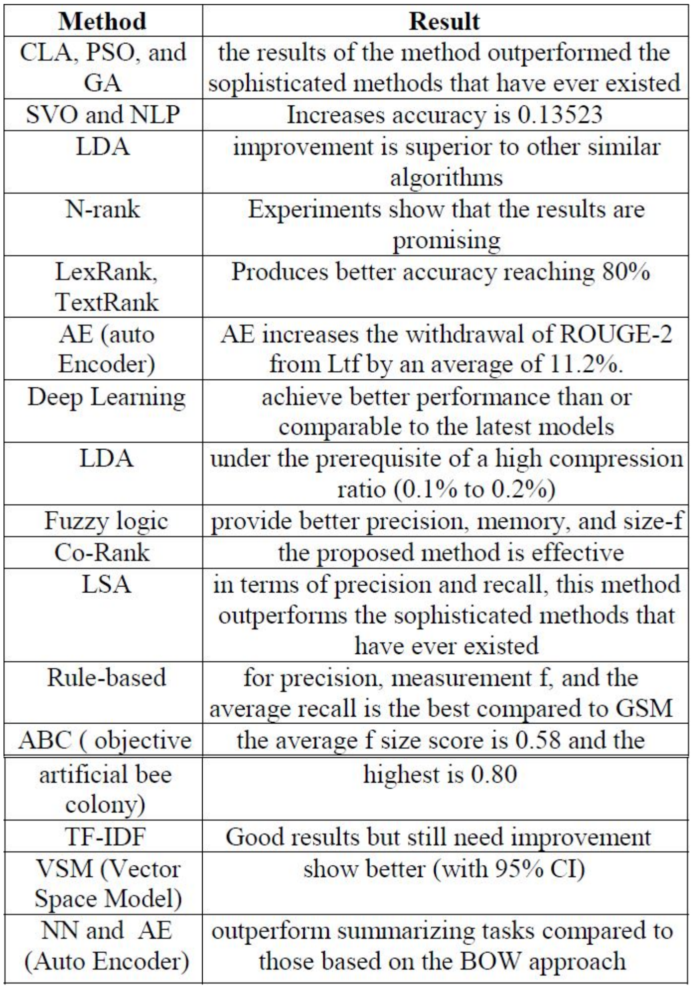
Abstractive summarization is a summary system that produces new phrases or uses words that are not in the original text. For perfect abstractive summaries, the model must really understand the document and then try to express that understanding briefly using new words and phrases [5] or arrange them in different forms. In comparison to abstracts which have more challenging problems and require extensive natural language processing, the extract area is more well researched. In general, abstractive summarization methods are grouped into two categories: Linguistics (syntax) and semantic approaches. Summary with syntax method includes lead and body methods, tree-based methods and information item-based methods . While abstractive summarization with semantic methods used on ontology-based methods and template-based methods.

The new sentences that reflect some text information from the main text are created in abstraction based summarization. That we achieve through the development of natural language. This generates more human like information in which sentences represents the main idea of the text in summary rather the including exact sentences or words from the main text in the final summary like Extraction based summarization techniques.

Also it can be classified Based on Number of Documents:Single Document and Multi Document.

- Single Document: A single document summary system will produce a summary based on one document source. Multiple subpapers with multiple paragraphs may form a single text. The material listed in each sub-document underlines all the different aspects around the same subject. [5] Kamal made automatic text summarization using Key Concepts in single documents and Hans et al. Automatic text summary built using TF-IDF for text summary use in single documents.
- Multi Document: Text summarizing of multi-documents is a method that involves a large amount of information in different sources of documents related to having only important material or key ideas in the text [5]. Multi document summarization can also be interpreted as a summary of documents covering the same topic from a document or information taken from several sources. A major problem for researchers is the management of related features from one text to a multidocument. John et al. A multi-document text summarization experiment was conducted, and the results showed that the current state beat in terms of recall and accuracy.

IV. Classification Methods



Method	Result
CLA, PSO, and GA	the results of the method outperformed the sophisticated methods that have ever existed
SVO and NLP	Increases accuracy is 0.13523
LDA	improvement is superior to other similar algorithms
N-rank	Experiments show that the results are promising
LexRank, TextRank	Produces better accuracy reaching 80%
AE (auto Encoder)	AE increases the withdrawal of ROUGE-2 from Ltf by an average of 11.2%.
Deep Learning	achieve better performance than or comparable to the latest models
LDA	under the prerequisite of a high compression ratio (0.1% to 0.2%)
Fuzzy logic	provide better precision, memory, and size-f
Co-Rank	the proposed method is effective
LSA	in terms of precision and recall, this method outperforms the sophisticated methods that have ever existed
Rule-based	for precision, measurement f, and the average recall is the best compared to GSM
ABC (objective artificial bee colony)	the average f size score is 0.58 and the highest is 0.80
TF-IDF	Good results but still need improvement
VSM (Vector Space Model)	show better (with 95% CI)
NN and AE (Auto Encoder)	outperform summarizing tasks compared to those based on the BOW approach

Based on the comparison table, table 2, the best method now is VSM with 95% accuracy. Along with other approaches, the current approach can be modified or hybridised to generate methods of greater precision. Or you can use a method approach from other fields besides summarization to be used in summarization with results that are not inferior to existing methods like as Kaichun Yao et al. Using a neuro-computing approach to generate new DQN methods(Deep Q-Network). The DQN model uses two different architectures, namely the CNN-RNN hierarchy network and the RNN-RNN hierarchy.That's because not only does the DQN model generate knowledge features that reflect the DQN status, it also lists potential actions from the sentence in the DQN text.

A. Naive-Bayes Methods:

Given a training set of documents with hand-selected document extracts, develop a classification function that estimates the probability a given sentence is included in an extract. New extracts can then be generated by rating the sentences according to this likelihood and selecting a number of the top scoring ones listed by the user [6].

The features were compliant to (Edmundson, 1969), but additionally included the sentence length and the presence of uppercase words. Each sentence was given a score according to , and only the n top sentences were extracted. To evaluate the system, a corpus of technical documents with manual abstracts was used in the following way: for each sentence in the manual abstract, the authors manually analyzed its match with the actual document sentences and created a mapping (e.g. exact match with a sentence, matching a join of two sentences, not matchable, etc.). They then tested the auto-extracts against this mapping. Review of features showed that the best performed was a framework using only the location and the cue features, along with the sentence length feature.

B. Hidden Markov Model:

Hidden Markov Model(HMM) is an another method to extract a sentence from a document. The Hidden Markov model has fewer assumptions of freedom, as compared to a naive Bayesian approach. The HMM in particular does not conclude that the likelihood of sentence I being in the summary is independent of whether sentence i-1 is in the summary.

Conroy and O'leary (2001) modeled the problem of extracting a sentence from a document using a hidden Markov model (HMM). The basic reason for using a sequential model is to take the local dependencies between sentences into account. Only three features were used: the sentence location in the document, the number of terms(built into the state structure of the HMM) in the sentence and the likeliness of the terms given the terms in the document.

C. Log-linear model approach:

Making use of Log-linear models could experientially show that the summarization system produced better summaries than the Naïve-Bayes model.The summaries generated from such a model, using the standard F-score, were evaluated. The features included word pairs, length of sentence, location of sentence, and discourse features such as the introduction inside or the conclusion inside.

Osborne claims that existing approaches to summarization have always assumed feature independence. The author used log-linear models to obviate this assumption and showed empirically that the system produced better extracts than a naive-Bayes model, with a prior appended to both models. Let c be a label, s the item we are interested in labeling, fi the i-th feature, and _i the corresponding feature weight.

Just two valid labels exist in this domain: either the sentence is to be removed, or it is not. The weights were conditioned by descending of the conjugate gradient.

D. Neural Networks:

In 2001-02, DUC issued a task of creating a 100-word summary of a single news article. However, the best performing systems in the evaluations could not outperform the baseline with statistical significance. This extremely strong baseline has been analyzed by Nenkova and corresponds to the selection of the _ rst n sentences of a newswire article. This surprising result has been attributed to the journalistic convention of putting the most important part of an article in the initial paragraphs. After 2002, the task of single-document summarization for newswire was dropped from DUC [7].

Svore et al. propose an algorithm based on neural nets and the use of third party datasets to tackle the problem of extractive summarization, outperforming the baseline with statistical significance. He trained a model from the labels and the features for each sentence of an article, that could infer the proper ranking of sentences in a test document. The ranking was accomplished using RankNet , a pair-based neural network algorithm designed to rank a set of inputs that uses the gradient descent method for training.

E. Fuzzy Logic Based Text Summarization:

This method considers each features of a text such as sentence length, similarity to key word and others as the input of fuzzy system. Then, it develops and enters all the rules required for summarization, in the knowledge base of system. Instead, for each sentence in the output a value from zero to one is obtained based on the characteristics of the sentence and the rules available in the knowledge base. The value obtained as an output defines the measure of the final summary meaning of the sentence. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using IF-THEN rules according to the feature criteria. Text summarization based on design of the Fuzzy logic system usually includes selecting Fuzzy rules and membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system. The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base.Crisp inputs are converted into linguistic values in the fuzzifier using a membership function that is used for the linguistic input variables [8]. After fuzzification, the inference engine look into to the rule base containing fuzzy IFTHEN rules to obtain the linguistic values. In the final step the defuzzifier converts the output linguistic variables from the inference to the final crisp values using the membership function to represent the final sentence score.

Witte and Bergle present a fuzzy-theory based approach to co-reference resolution and its application to text summarization.Automatic co-reference determination between noun phrases is filled with confusion. We demonstrate how fuzzy sets can be used to create a new co-reference algorithm that specifically captures this ambiguity and allows us to identify varying degrees of co-reference. Patil and Kulkarni also use Fuzzy Logic to score sentences after feauture selection and pre-processing step. They use eight features for text summarization: title word, sentence length, sentence position, numerical data, thematic words, sentence to sentence similarity, term weight and Proper Nouns.

F. Hybrid approaches:

- Fuzzy logic and LSA based approach:
  This hybrid approach uses fuzzy logic as a summarization sub-task improved the quality of summary by a great amount. It enhances the summary output by integrating the latent semantic analysis into the fuzzy logic framework derived from the sentence function to extract semant relationships between concepts in the original text.
- Graph and neural network based approach
  This method works on the sentence extraction-based text summarization task use the graph based algorithm to calculate importance of each sentence in document and most important sentences are extracted to generate document summary.Using a recurrent neural network, it uses Disambiguation Part of Language [9]. System recognises the most important phrases using various shallow linguistic features; it considers the degree of communication between the text units to reduce the resulting summary's weak linking sentences. The process can be described in three parts:
  i. preprocessing: Preprocessing Parse the document and generate sentences.
  ii. Graph Building: This represents a sentence as a node with all its properties and methods to handle with its behavior.
  iii. Sentence Ranking Algorithm: The basic approach of Sentence Rank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally.

V. Problems with the text summarization

The Problems with extractive text summarization are:
1. Usually longer sentences chosen for summary, so redundant portions of the sentences for summary are also included and space is used.
2.If the summary length is not long enough, the relevant information scattered in different statements can not be captured using extractive description.
3. Clashing details may not be portrayed correctly.
4. Sentences often have pronouns in it. If used out of context they lose their referents. If unrelated sentences are clubbed together, confusing anaphors understanding can result in misrepresentation of original information.
5. The same issue is with summarization of multi-documents, since text extraction is achieved on different sources. Post-processing may be used to address these problems, for example, replacing pronouns with their background, replacing relative temporal expression with actual dates etc.
Problems with the abstract text description are: Problem representation is the problem for abstract summary. Unit functionality depends on how carefully problem is described.The system can't summarise things which aren't properly represented in issue [6].

VI. Conclusion

Owing to fast growth of technology and use of Internet, there is information overload. This problem can be solved if strong text summarizers are available which produce a document summary to support user.This area for study is inspiring and creative, and has a variety of applications.A review of the literature focused mainly on the pioneering work of great personalities who contributed to this area and the methods used over the years that can be useful in future.. Also, a brief classification is explained by various methods. Like Bayesian Classifier, Hidden Markov Model, Neural Networks and Fuzzy Logic.

References

[1] H. P. Luhn, “The automatic creation of literature abstracts,” *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[2] P. B. Baxendale, “Machine-made index for technical literature—an experiment,” *IBM Journal of research and development*, vol. 2, no. 4, pp. 354–361, 1958.

[3] H. P. Edmundson, “New methods in automatic extracting,” *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[4] O. Tas and F. Kiyani, “A survey automatic text summarization,” *PressAcademia Procedia*, vol. 5, no. 1, pp. 205–213, 2007.

[5] A. P. Widyassari, A. Affandy, E. Noersasongko, A. Z. Fanani, A. Syukur, and R. S. Basuki, “Literature review of automatic text summarization: Research trend, dataset and method,” in *2019 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2019, pp. 491–496.

[6] S. Saziyabegum and P. S. Sajja, “Literature review on extractive text summarization approaches,” *International Journal of Computer Applications*, vol. 156, no. 12, 2016.

[7] N. Bhatia and A. Jaiswal, “Literature review on automatic text summarization: single and multiple summarizations,” *International Journal of Computer Applications*, vol. 117, no. 6, 2015.

[8] R. Rani and S. Tandon, “Chat summarization and sentiment analysis techniques in data mining,” in *2018 4th International Conference on Computing Sciences (ICCS)*. IEEE, 2018, pp. 102–106.

[9] E. Hovy, C.-Y. Lin *et al.*, “Automated text summarization in summarist,” *Advances in automatic text summarization*, vol. 14, 1999.