

A Comparative Study on Various Approaches and Complexities of Text Summarization

Pramoda Devi B and Jagadish S Kallimani

Department of Computer Science and Engineering
M S Ramaiah Institute of Technology (MSRIT)
Bangalore, India

Article Info

Volume 83

Page Number: 10172 - 10183

Publication Issue:

May - June 2020

Abstract:

In today's digitalized world web is having at most information for users, additional to that newspaper, textbook and magazine are offline resource of information. Web consist of million and billions of resources in the form of documents in addition to that many documents in different domain are adding to it, thus data is growing exponentially. To understand and assimilate this data many applications of natural language processing came to existence namely, Text mining, Information Retrieval, Machine Translation, Question and answering, text summarization and many more. Research on text summarization started over seven decades and till now effective method or system is not available to generate summary as human. In this paper, surveys the recent literature on different automatic text summarization method and propose idea of abstractive text summarization form orphologically rich language, kannada which is still lacking in field of text summarization.

Keywords: Abstractive text summarization, extractive text summarization, semantic nature, Parts of Speech (POS) tagging, term frequency-inverse document frequency (tf-idf), Neural attention model, Naive-Bayes classification, ROust Clustering algorithm (ROCK), Recurrent Neural Network encoder-decoder and document classifier.

Article History

Article Received: 19 November 2019

Revised: 27 January 2020

Accepted: 24 February 2020

Publication: 18 May 2020

1. Introduction

Natural Language Processing(NLP) combines the efforts of artificial intelligence,computational linguistics and computer science to read and understand the natural language. There are wide range applications of NLP like Information Retrieval, Document Clustering, Word Sense Disambiguation, Short answer grading, text summarization. Real time examples of NLP are spam filter, autocomplete, spell check, voice text messaging, siri, Alexa and google assistant.As data is exploding in today's world it needs to process in some way to analyze it, NLP provide a beautiful way to do this by automatic text summarization. Summarization is the process of shortening or condensing the original text document which contains information of high-quality, such that it

conveys the meaning original document. Summary can be of different category, based on type of

generated summarythere are two types,Extractive Summarizationand

AbstractiveSummarization,Extractive text Summarization selects the important sentences, or it reuse the text in the original document to make summary. But this summary is un-cohesive, un-coherent.Abstractive text Summarizationrephrases the sentence from the original document. Since it considers semantic into consideration the problem of extractive summary is overcome i.e un-cohesive, un-coherency. The summary generated here is closer to human generated summary.Based on content of summary there are three types, Inductive summarization, evaluative summarization and Informative summarization.Indicative Summarization gives summary as important topic of

original text by reducing text by 90%. It helps the reader to decide to continue reading the document. In Evaluative Summarization, summary is considered as review or opinion of author on given topic or product and Informative Summarization gives summary which is longer than indicative summary and reduce the summary size to 70% than original document. Based on number of documents to summarize there is two types, Single-document and multi-document summarization. In Single-Document Summarization, system take only one document for summarization and in Multi-Document Summarization system take multiple document of same topic to generate a single summary. Based on target audience there are three types of summary, generic, query-focused and update summarization. In Generic summarization, summary is generated irrespective of domain. In Query-focused Summarization, summary is generated by picking information which is relevant to user query and in

Update Summarization, summary is generated by omitting the information that user already has and maintains novelty. Based on type of summarizer there are two types namely author and expert summarization. In Author Summarization, summary is written based on author or writer perspective and in Expert Summarization, summary is generated by domain expertise who has good knowledge and skills in that domain. Based on input and output language there are three types of summarization namely monolingual, multi-lingual and cross-lingual summarization. In Monolingual Summarization, system work on single language and generate summary in same language of input document, where as in Multilingual Summarization, system work on multiple language but generate summary in same language of input document and in Cross-lingual Summarization, system generate summary in different language of input document.

Multi-document
Query
Single document

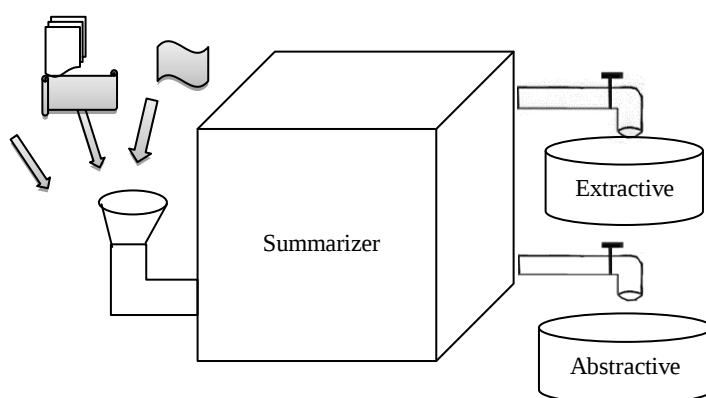


Figure 1: Overview of Summarization

The above figure 1 shows the overview of text summarization. Summarizer takes input as single document or multi-document or query. First job of summarizer is pre-processing using various approach like tokenization, stemming, lemmatization etc. After applying methodology like machine learning approach or term frequency-inverse document frequency (tf-idf), summarizer picks the sentence from the input document and forms extractive text

summarization, or it rephrases the sentences in the input document and form abstractive text

summarization in either of the summary type content should be less than half of the input document. The rest of the paper is structured as follows. Section two present current and future scope of this paper, section three gives objective of this paper, section four presents Related works, section five gives

benefits of summarization, conclusion and future work is covered in section six.

2. Current and Future Scope

Text Summarization is said to be the important research topic in field of data science, especially abstractive approach. The traditional approach to Abstractive text summarization using rule-based AI is been poorly used. But development in the field of Deep Learning with neural network has improved it. Jiang-Conrath semantic similarity matrix is an advanced approach which can be used in document clustering, summarization etc. which yields best results compared to other methods. Precision, recall, F-score etc. are the evaluation measures of text summarization, if these values are said to be high then those summarizations can be used in the fields of description answer evaluation system, short answer grading, question answering chat bots, automated content creations applications etc. Can use Reinforcement learning to improve the accuracy of text summarization. Apply text summarization to other Indian regional languages especially kannada language, which is rich in morphologically rich language.

3. Objectives of the Study

In this paper, detailed study and analysis on developed abstractive text summarization methods are attempted with following objectives.

- To understand different approach of text summarization
- To understand different methodology used for text summarization
- To know which domain and languages are considered for text summarization
- To analyze the accuracy of each method
- To know different type of text summarization
- To understand drawbacks and efficiency of each method

4. Related Work

This section shows the brief comparison of different text summarization approaches from a decade by

highlighting its aim, methodology used, conclusion, future work and remarks.

Table 1: Comparison Table of Summarization Approach

SL .no	Author	Publication name	Year of publicat ion	Title	Aim of project	Methodology used	Conclusion	Future work	Remark
1	Ahmet Aker, Trevor Cohn, Robert Gaizauskas	Association for Computational Linguistics	2010	Multi- document Summarizati on using A* Search and Discriminati ve Training	To find extractive text summariz ation up to given length.	A* search algorithm, discriminative training algorithm.	Find extractive text summarizati on of user specified length using A* algorithm and use discriminati ve training method to increase the quality of summary generated. This paper also addresses the search and training problem which is most	To conside r global feature and reduce the redund ancy in the generat ed summa ry	This paper uses machine learning approach and A* learning algorithm to find extractive text summarization for English language which lacks in semantic and generated summary has redundancy. This system generate summary which is optimal and efficient, and this model consider only local features.

							challenging part in extractive text summarization.		
2	Jayashree.R, SrikantaMurthy.K and Sunny.K	International Journal on Soft Computing(IJSC)	2011	Document Summarization in Kannada Using Keyword Extraction	To find Kannada text summarization by extracting keywords.	GSS coefficient, IDF, TF	Keyword is key factor to identify the document and helps in indexing. Extractive summary is generated by keyword extraction by considering pre-classified document.	To produce abstractive text summarization considering semantic nature and to classify the document and then go for summarization.	M number of sentences is included in summary, where m is user defined. Generated summary lacks in semantic.
	Alexander M.Rush, Sumit Chopra,	Association for Computational	2015	A Neural Attention	To find abstractive	Neural attention	Sentence-level	To produce	In this paper sentence level

3	Jason Weston	Linguistics		Model for Sentence Summarization	Sentence level summarization with data-driven approach.	model, Extractive tuning.	abstractive text summarization based on recent development in Neural machine translation. This system uses attention-based model to generate summary. To translate the rare words in abstractive model is a difficult task, to resolve this issue they came up with tuning a small set of additional features.	Paragraph-level summary.	summarization is beautifully explained, using Neural attention-based mechanism. This model is compared with various other summarization model and this model does the best.
	Yogesh kumar Meena,	International Conference on	2015	Domain Independent	Aims at producing	Machine learning	Concluded that no	Can extend	Some of the methods works

4	Dinesh Gopalani	Intelligent Computing, Communication and Convergence(I CCC)		Framework for Automatic Text Summarization	platform independent for generating summarization for all domains and even for abstractive and extractive text summarization	algorithm (No specific algorithm mentioned), TF, IDF	single method can be applicable for all domain thus proposes framework classifies the input document and then select summarization method. Once input data categorized properly, we can apply either abstractive or extractive summarization technique.	to sub-category of document and can apply on standard datasets	efficiently in particular domain and may not give accurate result in another domain. The proposed method is not depended on any domain thus it should have knowledge of corpus to select the summarization.
	Jagadish S. Kallimani, K.G Srinivasa and B. Eswara Reddy	Association for Computational Linguistics	2016	Statistical and analytical study of guided	To develop abstractive text summariz	TF/IDF, IE rules	Template based summary is generated by classifying	Speech output summary for the	An attempt to do template based abstractive text summarization

5				abstractive text summarization	ation based on template.		the input document using TF and extracting the important information using IE rules. Comparison of abstractive and extractivemethod summarizes and conclude that abstractive type summary outperforms.	current work	especially for Indian regional language. Since template is independent of each domain and topic, for each topic we need to generate independent template, which is tedious process.
	Sumit Chopra, Michael Auli, Alexander M.Rush	Association for Computational Linguistics	2016	Abstractive Sentence Summarization with Attentive Recurrent Neural	To produce abstractive text summarization of input	Recurrent Neural Networks	This model is trained with Gigaword corpus to produce headlines	Not specified	This system is extension of abstractive text summarization of [3]. It outperforms for DUC-2005

6				Networks	sentences using conditional Recurrent Neural Network.		based on first line of news article. This model can be used to train on large number of datasets.		datasets.
7	Jianpeng Cheng, Mirella Lapata	Association for Computational Linguistics	2016	Neural Summarization by Extracting Sentences and Words	Aims at producing extractive text summarization by extracting word or sentence.	Neural network-based reader or encoder and attention-based content extractor.	This system uses data-driven approach to produce extractive text summarization using neural network and continuous sentence feature for single document.	To make use of neural networks, to consider structural information and to adopt information theoretical approach using unsupervised	By understating the meaning by encoder or reader and extracting the important sentence by attention-based content extractor, this model produces the extractive summary for single document.

								learnin g to generat e summa ry	
8	Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo and Ichiro Sakata	Association for Computational Linguistics.	2017	Extractive Summarizati on Using Multi-Task learning with Document Classificatio n	Single document summariz ation with small amount of reference summaries	Recurrent Neural Network encoder- decoder and document classifier	This system uses multi- tasking by doing both sentence extraction for summarizati on and document classificatio n based on the certain domain	Not specifi ed	This framework summarizes single document with relatively small reference for training.
9	Anusha B S, Harshitha P, Divya Ramesh, Uma D, Lalithnarayan C	International Journal of Computer Science Application	2019	Multi- Classificatio n and Automatic Text Summarizati on of Kannada News Articles	To produce automatic summary of news article from several source	Naive-Bayes classification, ROCK clustering algorithm	Summary is generated for news article for three category i.e General news, sports and politics	To produc e abstrac tive text summa rization by conside ring	Produce summarization but did not considered the semantic of sentence thus yield summary with less accuracy.

								more than three categories	
10	Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat et al	Association for Computational Linguistics.	2019	A Summarization System for Scientific Documents.	Aims to produce summary for Computer science publication.	Science-Parser, Elastic-search	Generates summary for each sub-categories of paper and integrate this to get summary of whole paper.	In current work they considered only three entity and future work to add on the entities and increase the corpus of paper	It is first system to generate summary for research paper. So, researcher can surf paper easily, thus reduced the researcher burden. In this paper they took input form researchers and scholar that how they browse and go through the paper.

5. Benefits of Summarization

As the data is exploding in internet today, it is difficult to analyze and digest the large amount of textual data quickly. User must surf the whole document to get the desired information that he/she required. It is difficult for the user to extract the important information manually from the large document. Automatic summarization come into existence in order to resolve the above issue. Summarization is the process of compressing the original textsuch that it preserves the most importantdata and contains size less than original document. It reduces the time of user and get glimpse of the data quickly. In researcher point of view, it will help a lot while selecting the research paper to read further.

6. Conclusion and Future Work

Text summarization helps the user to find most important information by condensing the input document, thus saving the precious time of user. Lot of work done on the text summarization from past 1950's and there is no particular system which generate the accurate summarythus research is going on. Comparison of different approach of text summarization in shown in table 1. By inferring the above comparison table,text summarization is done on English language and less work undergone in text summarization for regional language especially kannada. The above papers commonly use extractive approach of text summarization, which lacks in semantics.As a future work one can concentrate on abstractive summarization of kannada by considering its semantic into consideration.

References

1. Ahmet Aker, Trevor Cohn and Robert Gaizauskas, "Multi-Document Summarization using A* search and Discriminative Training", Proceeding to Conference on Empirical Method in Natural Language Processing, ACL pages 482-491, MIT, Massachusetts, USA, 9-11 October 2010
2. Jayashree R, Murthy, Srikantha K and Sunny. K "Document Summarization In Kannada Using Keyword Extraction", International Journal on Soft Computing, Vol 2, Number 4, Pg- 81, Year, 2011
3. Alexander M. Rush, Sumit Chopra and Jason Weston "A Neural Attention Model for Sentence Summarization", Proceeding to Conference on Empirical Method in Natural Language Processing, ACL pages 379-389, Lisbon Portugal 17-21, September 2015.
4. Yogesh Kumar Meena and Dinesh Gopalani "Domain independent framework for automatic summarization", Procedia Computer Science,2015.
5. Jagadish S. Kallimani, K G Srinivasa and B Eswara Reddy, "Statistical and analytical study of guided abstractive text summarization", Current Science, Volume 110, No. 1, 10 January 2016.
6. Sumit Chopra, Michael Auli and Alexander M. Rush, "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks", Proceeding of NACCL-HLT, ACL pages 93-98, San Diego, California, June 12-17, 2016.
7. Jianpeng Cheng and Mirella Lapata, "Neural Summarization by Extracting Sentences and Words", Proceeding of the 54th Annual Meeting of the Association for Computational Linguistics, pages 484-494, Berlin, Germany, August 7-12, 2016.
8. Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo and Ichiro Sakata, "Extractive Summarization Using Multi-Task Learning with Document Classification", Proceeding to Conference on Empirical Method in Natural Language Processing, ACL pages 2101-2110 Copenhagen, Denmark, September 7-11, 2017
9. Divya Ramesh, Anusha B S and Harshitha P, "Multi-Classification and Automatic Text Summarization of Kannada News Articles", International Journal of Computer Application Volume 181- No.38, January 2019.
10. Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora PeledNakash, OdelliaBoni,HaggaiRoitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin et al, "A Summarization System for Scientific Documents", Proceeding to the EMNLP and the 9th IJCNLP, ACL pages 211-216 Hong Kong, China November 3 -7 2019.