Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

Comprehensive Survey on Abstractive Text Summarization

Kirti Wanjale[1], Paritosh Marathe[2], Vedant Patil[3], Sandesh Lokhande[4], Hrishikesh Bhamare[5],
[2,3,4,5]Students, BE (Computer Engineering),
[1] Associate Professor, Department of Computer Engineering,
Vishwakarma Institute of Information Technology, Pune, India

***Abstract*— Over the past few years, we have seen the rise of Automation for the purpose of human convenience. Using ML learning approach, we inch ever closer towards achieving a general purpose AI. The field of Artificial Intelligence (AI) can roughly be divided into three parts namely Machine Learning (ML), Computer Vision and Natural Language Processing (NLP). NLP involves the understanding and handling of human language of which Automatic Text Summarization is an important part. Text summarization is the process of shortening a lengthy document into a short summary. It creates fluent and coherent information while maintaining the context (meaning) of the information. It is a difficult task for human beings to generate a manual summary since it requires a rigorous analysis of the entire document. In order to reduce human efforts and time, automatic summarization techniques prove to be helpful. Text summarization has broadly two techniques, namely Extractive text summarization and abstractive text summarization. Extractive technique relies on extraction of key words, whereas in abstractive text summarization technique utilizes the principles of deep learning to generate the required summary.**

Keywords*—- *Text Summarization, Extractive Text Summarization, Abstractive Text Summarization.

INTRODUCTION

Text summarization is the process of reducing the size of the original document while preserving its original information and the summary generated is less than half of the main text of the original document. Summarization can be seen as a two-step process. The first step is the extraction of vital concepts or sequence formation from the source text by building associate intermediate vector or file of sorts. This step may also include any pre-processing required to be done on the text including tokenization, tagging or otherwise. The second step uses this intermediate file to generate a summary. News blaster can be considered as an example of a text summarizer that helps users to find the news that's of most interest to them. The system has the ability to collect, cluster, categorize and summarize news from multiple sites on the web on a daily basis. A summarization machine can be viewed as a system that accepts either a one document or multiple documents or a query as an input and can extract or abstract a summary based on that input.

I. LITERATURE SURVEY

Recent advancements in technology have enabled us to use ML models and perform exceptionally well in tasks unusually performed by humans. Higher computational power has empowered the use of complex sometimes non-linear models to be tasked on tasks such as text summarization, where they yield exceptionally high accuracies sometimes even better than humans.

Previous researches have introduced various summarization techniques which are cited in this section. Most of the researches concentrate on sentence extraction rather than generation for text summarization. Automatic text summarization as aptly mentioned in Mehdi Allahyari et al. [4] is the task of automatically producing a concise and fluent summary without human intervention. This is achieved by a machine using various techniques some of which incorporate NLP and Deep learning (neural networks) as the base.

There are two base divisions, Extractive summarization and Abstractive summarization.

Qaiser, Shahzad & Ali, et al. [3], the use of Term Frequency-Inverse Document Frequency (TF-IDF) is discussed in examining the relevance of key-words to documents in corpus. Extractive text summarization uses a statistical based approach to select important sentences or words from the document. Statistical approach can summarize the document using features like term frequency, location, and title, assigning weights to the keywords and then calculating the score of the sentence and selecting the highest scored sentence into the summary. The study is focused on how the algorithm can be applied to a number of documents. First, the working principle and flow which should be followed for implementation of TF-IDF is illustrated. Secondly, to verify the findings from executing the algorithm, results are presented, and then strengths and weaknesses of TF-IDF algorithm are evaluated. This paper also tackles such weaknesses.

John X. Qiu, Hong-Jun Yoon, et al. [5], we see an implementation of Extractive summarization using convolutional Neural Network (CNN) for extracting ICDO-3 topographic codes from a corpus of breast and lung cancer pathology reports. The purpose was the extraction of information from the export and therefore helps in faster understanding of a report. However, the basic task was to extract words from the document and place them and not to consider the grammatical nuance, flow of words (sequences) and carry over the context. Alexander M. Rush, Sumit

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

Chopra, et al. [11], we see the use of Long Short Term Memory (LSTM) [11] as a base unit in an encoder decoder model for summarization of news articles. Ramesh Nallapati, Bowen Zhou, et al. [7] introduces a model called Sequence to Sequence (Seq2Seq) proposed in [10] which makes use of an encoder decoder arrangement with an attention mechanism for summary generation which again makes use of LSTM [11] as its basis. Abigail See, Peter J. Liu, et al [6], we see an implementation of a pointer generator network which utilizes a pointer network [9] with our traditional approach in [6] for a decoder. Alexander M. Rush, Sumit Chopra, et al. [11] proposes a more data driven approach as compared with the traditional generative approach. Using the model described in [7] as a base this paper draws a parallel between a Convolutional encoder a neural network language model (NNLM) encoder (as compared to a traditional LSTM [11]) and an attention based encoder.

II. ABSTRACTIVE TEXT SUMMARIZATION

Abstractive summarization is a more efficient and accurate in comparison to extractive summarization. It can retrieve information from multiple documents and create an accurate summarization of them. Its popularity lies in its ability of developing new sentences to tell the important information from the source text documents. An abstractive summarizer displays the summarized information in a coherent form that is easily readable and grammatically correct. Readability or linguistic quality is an important catalyst for improving the quality of a summary. Abstractive text summarization method generates a sentence from a semantic representation and then uses natural language generation techniques to create a summary that is closer to what a human might generate. Here we are concentrating on the generative approach for abstractive text summarization.

Since the human language is sequential in nature, it is found that RNN's and models based on them have a higher performance as compared to other approaches. Sequence to Sequence (Seq2Seq) (Ilya Sutskever et. al. 2014) is one of the base models in abstractive text summarization. It turns one sequence to another (e.g. machine translation). It achieves this by the use of DNN's. This approach specifically uses LSTM to avoid the problem of back propagation and for long carry of sequences. The output of one cell is the input of another and so on. This allows it to learn the sequence thorough a sentence. It normally employs encoder-decoder architecture. The encoder turns an input into its corresponding hidden vector containing the item and its value. This hidden vector is passed on to the decoder which reverses the process, i.e. turning the hidden vector into a corresponding item output utilizing the previous output as input.

Seq2seq models (see Fig. 1) [10] have been successfully applied to a variety of NLP tasks, such as machine translation, headline generation, text summarization and speech recognition. Inspired by the success of neural machine translation (NMT), (Bahdanau et al. 2014) introduced the concept of a "attention" model, which introduced a conditional probability at the decoder end effectively allowing the decoder to decide which part of the source input to pay attention to. This should a significant performance boost in comparison to the base enc-dec model. The dataset used was WMT. (Rush et al.) [11] further proposed a neural attention seq2seq model with a Neural Network Language Model(NNLM) attention-based encoder and a attention based beam decoder to the abstractive sentence summarization task, which has achieving a performance improvement over previous methods. The model was trained on DUC-2003, DUC-2004 and Gigaword. S Chopra et al. [9] further extended this model by replacing the feed-forward NNLM with a recurrent neural network (RNN). The model is also equipped with a convolutional attention-based encoder and a RNN (Elman or LSTM) decoder, and outperforms other state-of-the art models on a commonly used benchmark dataset, i.e., the Gigaword corpus. Ramesh Nallapati et al. [7] introduced several novel elements to the RNN encoder-decoder architecture to address critical problems in the abstractive text summarization. They also established benchmarks for these models on a CNN/Daily Mail dataset [16], which consists of pairs of news articles and multi-sentence highlights (summaries). These applied optimizations have led to further increase in performance.

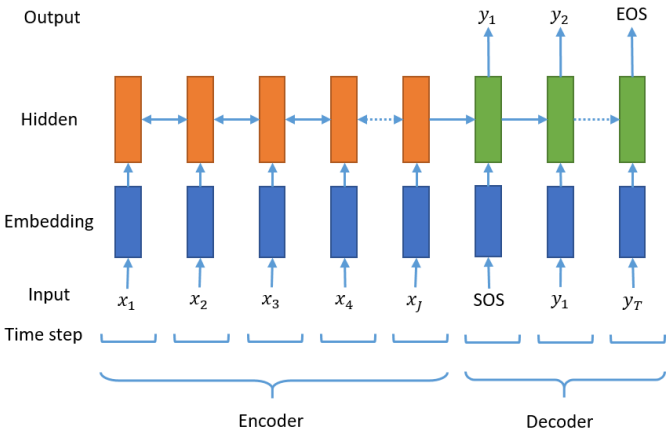

Figure 1: Basic seq2seq model

III CONCLUSION

The problem of automatic text summarization is an old problem with attempts and explanation going back a decade or two. Due to the introduction of powerful hardware and the resulting increase in computational power we can implement the more computationally hard algorithms to achieve the result. Hence there is a gradual incline towards abstractive summarization techniques utilizing non-linear models (deep learning) rather than the statistical (but computationally less expensive) extractive summarization techniques. While generating an abstract using abstractive summarization method still remains a difficult task. Abstractive summarization methods produce more coherent, less redundant and information rich summary.

Due to the above reasons the study of abstractive summarization techniques proves to be more useful.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

ACKNOWLEDGMENT

The author acknowledges the research done by the researchers referred in the paper for their useful research and in sharing their methodology. The author also acknowledges the efforts provided by Ass. Prof. Mrs. Kirti Wanjale towards guidelines related review work and proofreading of the paper.

REFERENCES

[1] Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, Hongyan Li, Generative Adversarial Network for Abstractive Text Summarization,2018.

[2] Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria5," Recent Trends in Deep Learning Based Natural Language Processing", November 2018.

[3] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, Text Summarization Techniques: A Brief Survey,2017, Cornell.

[5] John X. Qiu, Hong-Jun Yoon, Paul A. Fearn, and Georgia D. Tourassi, Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports,2017, IEEE Journal of Biomedical and Health Informatics.

[6] Abigail See, Peter J. Liu, Christopher D. Manning, Get to the Point: Summarization with Pointer-Generator Networks, Stanford University, 25 Apr 2017.

[7] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça˘glar G¨ulçehre, Bing Xiang, Beyond Abstractive Text Summarization using Sequence-to-sequence RNNs and, Aug 2016, Cornell.

[8] Jianpeng Cheng, Mirella Lapata, Neural Summarization by Extracting Sentences and Words, 1 July 2016.

[9] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2016, pp. 93–98.

[10] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the CNN/Daily Mail reading comprehension task," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers), vol. 1, 2016, pp. 2358–2367.

[11] Alexander M. Rush, Sumit Chopra, Jason Weston, A Neural Attention Model for Abstractive Sentence Summarization, 2015 Cornell.

[12] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp.379–389.

[13] Tsung-Hsien en, Milica Ga si c, Nikola Mrk si c, Pei-Hao Su, David Vandyke, Steve Young, "Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems", August 2015.

[14] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to Sequence Learning with Neural Networks, 2014.

[15] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.

[16] Sepp Hochreiter, Jurgen Schmidhuber, "LONG SHORT-TERM MEMORY", 1997.