International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Survey on Text Summarization Framework Using Machine Learning

Mrs. Kirti D Kulkarni, Prof. Jareena N. Shaikh
Department of Computer Engineering, Zeal College of Engineering and Research Pune, Maharashtra, India

Article Info
Volume 9, Issue 3

Page Number : 581-585

Publication Issue :
May-June-2022

Article History
Accepted : 01 May 2022
Published: 07 May 2022

ABSTRACT

An important natural language processing application, automatic text summarizing aims to condense a given textual content into a shorter model by using machine learning techniques. As media content transmission over the Internet continues to rise at an exponential rate, text summarization utilizing neural networks from asynchronous combinations of text is becoming increasingly necessary. Using the principles of natural language processing (NLP), this research proposes a framework for examining the intricate information included in multi-modal statistics and for improving the features of text summarization that are currently available. The underlying principle is to fill in the semantic gaps that exist between different types of content. In the following step, the summary for relevant information is generated using multi-modal topic modelling. Finally, all of the multi-modal aspects are taken into account in order to provide a textual summary that maximizes the relevance, non-redundancy, believability, and scope of the information by allocating an accumulation of submodular features.

Index Terms—word vectors, word analogies, fast text, Integer linear programming, text summarising, natural language processing.

I. INTRODUCTION

Now a days, there are large numbers of documents or information that is present related to any particular field. There are many sources out of which we can gather a lot of information that will be pertinent to our field of search. Much information is available at various sources like the internet. But, as we know that a huge amount of information cannot be always considered or taken into use. So, a precise amount of information is always considered and that information is drawn out from the original document that is huge in size. In other words, we can say that we pluck out the summary of the main document. A summary of any document is defined as a collection of essential data by collecting the brief statements accounting the main points of the original document. Therefore, Summarization of a text is a procedure of separating or getting the relevant data out of a very large document. It is the process of shortening the text document by using various technologies and methodologies to create a coherent summary including the major points of the

original document. There are various methods by which the summarization process can be carried out. While most summarization systems focus on only natural language processing (NLP), the opportunity to jointly optimize the quality of the summary with the aid of automatic speech recognition (ASR) and computer vision (CV) processing systems are widely ignored. On the other hand, given a news event (i.e., news topic), multimedia data are generally asynchronous in real life. Thus, Text summarization faces a major challenge in understanding the semantics of information. In this work, we present a system that can provide users with textual summaries to help to acquire the gist of asynchronous data in a short time without reading documents from beginning to end. The purpose of this work is to unite the NLP with neural network techniques to explore a new framework for mining the rich information contained in multimodal data to improve the quality of Text summarization.

Summarization is the process of compressing a long piece of material into a shorter version that retains the essential information. There are two types of summarization methods: extractive and abstractive. Extractive approaches create summaries solely from sections (typically full sentences) extracted straight from the source material, whereas abstractive methods may generate new words and phrases not found in the source text — as a human-written abstract normally does. Because copying huge portions of text from the original document provides baseline levels of accuracy, the extractive approach is easier.Text summarization has previously been split into two subtasks, namely sentence scoring and sentence selection, in prior publications that use extractive approaches. Sentence scoring is a technique for assigning an importance value to each sentence that has been extensively researched in the past. Extractive methods create summaries by reproducing parts of the source content (typically entire sentences), whereas abstractive methods may generate new words or phrases not found in the source document. Extractive summarization, which is commonly characterised as a sentence ranking or binary classification problem (i.e., sentences that are top ranked or predicted as True are selected as summaries), has received a lot of attention in the past.Content selection in summarization is normally performed by sentence (and, on rare occasions, phrase) extraction. Despite the fact that deep learning models are a significant component of both extractive and abstractive summarization systems, it is unclear how they accomplish content selection with only word and sentence embedding based features as input.One of the most difficult NLP tasks is summarization, which is defined as the process of generating a shorter version of a piece of text while keeping critical context information. The performance of sequence-to-sequence neural networks on summarization has lately improved significantly.

The availability of large-scale datasets, on the other hand, is critical to the effectiveness of these models. Furthermore, the length of the articles and the variety of styles might add to the complexity. Because news stories have their own distinct characteristics, systems trained solely on news may not be adequately generalised. Recent advances in machine learning have resulted in significant advancements in text summarization. Huge labelled summarization corpora, such as the CNN/Daily Mail dataset, have made it possible to train deep learning models with a large number of parameters. Recurrent neural network (RNN) and Arif Ur Rahman, the associate editor who coordinated the evaluation of this manuscript and approved it for publication, were among them. For text summarization, convolution neural networks (CNN) have been frequently employed. RNN is used in extractive approaches to evaluate sentence importance while simultaneously picking representative sentences.

A. Motivation

Text summarising is a technique for condensing information from a source text into a few representative sentences in order to construct a

coherent summary containing relevant information from source corpora. deep neural network-based summarization models has a number of serious flaws. To begin, a significant quantity of labelled training data is required. This is a common issue in low-resource languages where publicly available labelled data is lacking. So that we propose a model, Learning Free Integer Programming Summarizer (LFIP-SUM), which is an unsupervised extractive summarization model.

II. LITERATURE SURVEY

Abigail See et al: In this paper, authors demonstrated that a hybrid pointer generator design with coverage lowers inaccuracy and repetition. Authors tested our model on a fresh and difficult lengthy text dataset and found that it outperformed the abstract state-of-the-art result significantly. This model has a lot of abstract skills, but getting to higher degrees of abstraction is still a work in progress.

Qingyu Zhou et al: authors introduce unique neural network architecture for extractive document summarization in this paper, which addresses this problem by learning to score and pick sentences simultaneously. The most notable difference between our method and earlier methods is that it integrates sentence rating and selection into a single phase. It rates sentences based on the partial output summary and current extraction state each time it selects one. The suggested combination sentence scoring and selection strategy greatly outperforms the existing segregated method, according to ROUGE evaluation results.

Xingxing Zhang et al: In this paper, authors proposed a latent variable extractive summarization strategy that uses a sentence compression model to directly exploit human summaries. This approach outperforms a powerful extractive model in experiments, whereas applying the compression model on the output of our extractive system produces inferior results. Authors intend to investigate approaches to train compression models specifically for our summarising task in the future.

Chris Kedzie et al: In this study, An empirical research of deep learning-based content selection algorithms for summarization was given. Our findings imply that such models have significant limits in terms of their capacity to learn robust features for this task, and that more work on sentence representation for summarization is required.

Linqing Liu et al: Authors suggested an adversarial technique for abstractive text summarization in this study. Experiments revealed that this model was capable of producing more abstract, legible, and diversified summaries.

Jacob Devlin et al: In this paper, Recent empirical gains in language models owing to transfer learning have shown that rich, unsupervised pre-training is an important feature of many language comprehension systems. These findings, in particular, show that deep unidirectional topologies can benefit even low-resource jobs. Our key contribution is to extend these findings to deep bidirectional architectures, allowing a single pre-trained model to solve a wide range of NLP tasks.

T. Boongoen et al: This survey has presented classical and recently developed approaches to cluster ensemble.It begins with the formal terms used to define the problem. Following that, four main categories of consensus clustering approaches are explained in depth with examples. Following that, it describes extensions to three major components of a cluster ensemble framework: ensemble generation, representation and summarization, and consensus function. Many cluster ensemble approaches have been used for a wide range of applications and domain challenges due to their improved ability to deliver accurate data partitions.

Mahnaz Koupaee et al: In this paper, Authors present WikiHow,a new large-scale summary dataset made up of a variety of articles from the WikiHow knowledge base The aspects of WikiHow outlined in the research may provide additional challenges to summarization

systems. Authors expect that the new dataset will pique the interest of academics as a viable option for evaluating their systems.

Edouard Grave et al: In this work, Authors contribute word vectors trained on Wikipedia and the Common Crawl, as well as three new analogy datasets to evaluate these models, and a fast language identifier which can recognize 176 languages. Authors investigate the impact of several hyper parameters on the trained models' performance, demonstrating how to produce high-quality word vectors. Authors also show that, despite its noise, employing common crawl data can result in models with greater coverage and better models for languages with little Wikipedia. Finally, we find that the quality of the produced word vectors is substantially lower for low-resource languages, such as Hindi, than for other languages.

Zhilin Yang et al: In this study, XLNet is a generalised AR pre-training method that combines the benefits of AR and AE methods by using a permutation language modelling target. XLNet's neural architecture was designed to work in tandem with the AR goal, including the integration of Transformer-XL and the meticulous design of the two-stream attention mechanism. On a variety of tasks, XLNet delivers significant improvements above previous pre-training objectives.

III. CONCLUSION

In this survey, we studied various summarization techniques and algorithms.

IV. REFERENCES

[1]. MYEONGJUN JANG AND PILSUNG KANG, "Learning-Free Unsupervised Extractive Summeriza-tion Model ," 2021, arXiv:9321308 [Online].Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=9321308

[2]. A. See, P. J. Liu, and C. D. Manning, "Get to the point: Sum-marization with pointer-generator networks," 2017, arXiv:1704.04368. [Online].Available: http://arxiv.org/abs/1704.04368.

[3]. Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "Neural document summarization by jointly learning to score and select sen-tences," 2018, arXiv:1807.02305. [Online]. Available: http://arxiv.org/ abs/1807.02305.

[4]. X. Zhang, M. Lapata, F. Wei, and M. Zhou, "Neural latent extractive document summarization," 2018, arXiv:1808.07187. [Online]. Avail-able:http://arxiv.org/abs/1808.07187.

[5]. C. Kedzie, K. McKeown, and H. Daume, "Content selection in deep learning models of summarization," 2018, arXiv:1810.12343. [Online].

[6]. L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, "Generative adversarial network for abstractive text summarization," in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–2.

[7]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018,arXiv:1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[8]. T. Boongoen and N. Iam-On, "Cluster ensembles: A survey of approaches with recent extensions and applications," Comput. Sci. Rev., vol. 28, pp. 1–25, May 2018.

[9]. M. Koupaee and W. Y. Wang, "WikiHow: A large scale text summarization dataset," 2018, arXiv:1810.09305. [Online].

[10]. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in Proc. Int. Conf. Lang. Resour. Eval. (LREC), 2018, pp. 1–3.

[11]. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le,"XlNet: Generalized

autoregressive pretraining for language understand ing," in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 5753–5763.

[12]. C. Kedzie, K. McKeown, and H. Daume, "Content selection in deep learning models of summarization," 2018, arXiv:1810.12343. [Online]. Available: http://arxiv.org/abs/1810.12343

[13]. Z. Li, Z. Peng, S. Tang, C. Zhang, and H. Ma, "Text summarization method based on double attention pointer network," IEEE Access, vol. 8, pp. 11279–11288, 2020

[14]. J. Tan, X. Wan, and J. Xiao, "Abstractive document summarization with a graph-based attentional neural model," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, Jul. 2017, pp. 1171– 1181.

[15]. H. Kim and S. Lee, "A context based coverage model for abstractive document summarization," in Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC), Oct. 2019, pp. 1129–1132.

[16]. L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, "Generative adversarial network for abstractive text summarization," in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–2.