Research on Automatic Text Summarization Algorithms Based on Deep Learning

Jiachen Sun[1,2,*]

[1]School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China.

[2]Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding 071003, China.

*Corresponding author Email: 1634046649@qq.com

Abstract

With the explosive growth of news text data in the Internet and social media, how to quickly get key information from massive text data has become a hot issue. Text summary technology can compress long text into concise, coherent and unappropriate short text, thus speeding up people's access to information. Helps people read news from the Internet and social media more efficiently. For this reason, this paper summarizes the development process, research status and the problems to be solved of text summary technology.

Keywords

Automatic Text, Algorithms, Deep Learning.

1. Rationale and Background

Rapid development of Internet with big data technology places us in an era of information explosion, while also causing problems of text information overload to become increasingly severe. Through the Internet we are able to quickly obtain massive information, but the web text usually contains a lot of redundant data. And, the title party problem is getting stronger since the Internet Meyer into business era, there are various reasons that users cannot directly discern the effective content of the text, increasing the difficulty of users to obtain the target information. Therefore, rapidly extracting key information from the massive text information has become the urgent need of people, and the research topic of text key information extraction has emerged.

AI technologies have entered the third wave of development in recent years, with many excellent technologies being proposed and the field at a rapid pace. Natural language processing technology, although the most difficult of the artificial intelligence techniques, has also yielded a lot of historic results. Especially with the promotion of researchers from top companies such as Google, Facebook and Stanford and colleges, natural language processing has opened a new starting point in the field of deep learning. Large scale pretrained models, such as Google released Bert [1], iPhone released Ernie [2], have opened the pretraining era in the natural language domain. Abstract the generation technology also achieved better results, as illustrated by news publics at hundreds of micromolar voice sound, results from Google search engines, generation of Aichi Video Tags, and so on, and these technical results not only liberated the enterprise employees' hands, but also greatly improved efficiency and quality.

So the methodological research on abstract generation technology is pivotal in the fields of public opinion, science and technology retrieval, and news. For accurate and high-quality extraction of text critical information, in order to alleviate the problem of information overload people face and quickly obtain information that is valuable to themselves from the Internet, this study will mainly focus on unstructured text documents for abstract generation algorithms.

2. A Dynamic Analysis of The Current Status and Aevelopment of Research at Home and Abroad

Text Abstract refers to the computer's analysis of the textual information from which to select sentences that reflect the contents of the text topic or that are formally and meaningfully coherent, nonredundant sentences are generated by the computer on the basis of its understanding of the textual information. Users can obtain critical information from the abstract without reading the entire text, thereby helping users improve reading efficiency.

Internationally, in 1952 IBM Corporation researcher H.P. Luhn [3] was the first to suggest that computer can be utilized for literature compression, in which it was proposed to rate the importance of that word throughout the text by calculating the word frequency in the text, and in which several sentence composition abstracts were picked out by taking the statistical result as a reference for measuring the importance of a sentence by counting the word frequency of the words contained in the sentence in the text. Rush et al and nallapati et al [4] were among the first scholars to apply neural network codec architectures to text summaries. In 2018, Paulus et al [5] proposed a depth boosting model (DRM) for generative textual abstracts that uses an intra attentional mechanism to deal with covering problems. In this mechanism, the decoder processes the words generated before. Narayan et al [6] proposed a generative summary model suitable for extreme generalizations based on convolutional neural networks, introducing thematic distribution conditions (tconvs2s). In May 2019, Jacob Devlin et al [7] proposed a new linguistic model: the Bert (bidirectionalencoder representations from transformants), which is represented using the transformer's two-way encoder. Berts are designed to pre train bidirectional representations of depth by jointly modulating context in all layers, through an additional output layer, making it possible to fine tune pre trained Bert models to create state-of-the-art models for various tasks such as question response and linguistic reasoning, without substantial modification of the task specific architecture. In the same year, Yang Liu et al [8] constructed a text Abstract pretrained encoder model by stacking several mutually exclusive transformer layers on the basis of the Bert encoder, which proposed a new fine-tuning method to optimize the encoder and decoder separately, and further improved the quality of the generated text summary. 2020, M Ramina et al. 9 passed keyword information in text format to an automated abstract model, which generated subject level abstracts using a Bert based bidirectional encoder. Ming Hsiang Su [10] et al used a text segmentation module to represent the input text in several parts using a Bert model and an LSTM pretrained two-way encoder. A summary model based on a Bert based summary model (bertsum) is then constructed, extracting the most important sentences from each segment. Yisong Chen et al [11] combined textrank with Bart models to increase the weight of key statements in the press, making the abstract more topical. With the development of graph neural network (GCN), Z. Liang et al [12] proposed a gate keeping graph neural attention network (ggnan) for abstract abstract. The proposed ggnan combines graph neural networks and the famous seq2seq to better encode complete graph structure information.

The research on text automatic abstract technology in China began in the 1980s, and multiple shallow features of text from research and development on subject groups led by Professor Yongcheng Wang [13] in 1997 scored sentences in the text and developed "" Chinese automatic abstract experiment system "" (sjtucaa), followed by the development of "" Chinese Literature Automatic Abstract System CAE "" and "" OA Chinese Literature Automatic Abstract System "". The system integrates indicating phrase, position, title, etc. In recent years, many scholars have fully considered the influence of semantic factors on the text Abstract in order to improve the issues such as information redundancy and diversity of the areas to which the information belongs. Du Xiuying [14] and others constructed multi text Abstract Based on extracting subject sentences using cloud computing platform, and at the same time proposed a multi text automatic abstract method using clustering theory and semantic similarity analysis as technology. Shen Hua [15] et al incorporated information such as sentence length and similarity of sentences in the text into the eigensemantic vector calculation of sentences. Hou rive [16] and others proposed to integrate thematic key information into the original encoder decoder structure using a multi attention mechanism to reinforce the model's understanding of the original theme. Lee Daozhou [17] et al proposed a generative text summary algorithm that inputs raw text to the encoder at the encoder end and generates a fixed length semantic vector in combination with a bidirectional gating cycle unit, assigning weights to each input word using an attentional mechanism to reduce the loss of details of input sequence information. Xiaomalay Jiang et al [18] mined the thematic information of original articles by drawing keywords and integrated them explicitly into the attention mechanism, so that the model, in the presence of global thematic information guidance, generates theme oriented abstracts in a context aware manner. Zhongxiang Cai et al.19 proposed an automatic text summary model, tri PCN, by using the transformer model to extract multi-level global text features in the decoding stage and fusing a pointer network. A copying mechanism for introducing pointer generation network models keyword information can be copied directly from the news text as the news title is generated. Yang Tao [20] and others have introduced motif models into the long text abstract task using a combination of abstraction and generative style. Overall, compared with the statistical based methods, the text abstract methods based on deep learning algorithms are more complex [21], and the resulting text Abstract is more comprehensive, objective, comprehensible and readable, which will be the trend of future research.

3. Key Tssues to be Solved

(1) Previous automatic summarization techniques tend to ignore important information and produce a lot of redundant information. Sequence-to-sequence models do not adequately model important information in sentence summary tasks. This study will better remove redundant information and reduce the omission of important information by incorporating text features, generator pointers, and hierarchical attention into the previous algorithm architecture. Therefore, this study will further explore the sequence-to-sequence model design for the characteristics of summary tasks.

(2) Although the above methods improve the quality of the generated summary, there is still room for improvement. Text themes are composed of multiple sub-themes, and each sub-theme has a different location and probability, so they need to be distinguished to produce a high-quality summary. This study will continue to investigate how keywords can be introduced into automatic summarization algorithms to improve the quality of summary generation by incorporating topic information.

(3) In addition to extracting key information to remove redundancy, the fact accuracy, language structure, lexicon, grammar, coherence and other aspects of sentences are also very important in text generation. In the current automatic summarization algorithm for Chinese text, the output sentences are not satisfactory. Sentences are often inconsistent and inconsistent. This study will strive to improve the language quality of the output summary.

References

[1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. ar Xiv preprint ar Xiv:1810.04805, 2018.

[2] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. ar Xiv preprint ar Xiv:1904.09223, 2019.

[3] Luhn H P. The automatic creation of literature abstracts [J]. IBM Journal of research
and development. 1958, 2 (2): 159–165.

[4] Nallapati R, Zhou B, Gulcehre C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond [J]. ar Xiv preprint ar Xiv:1602.06023. 2016.
networks [J]. ar Xiv preprint ar Xiv:1704.04368. 2017.

[5] Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization[J]. ar Xiv preprint ar Xiv:1705.04304. 2017.

[6] Narayan S, Cohen S B, Lapata M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization [J]. ar Xiv preprintar Xiv:1808.08745. 2018.

[7] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. ar Xiv preprint ar Xiv:1810.04805. 2018.

[8] Liu Y, Lapata M. Text summarization with pretrained encoders [J]. ar Xiv preprintar Xiv:1908.08345. 2019.

[9]M. Ramina, N. Darnay, C. Ludbe and A. Dhruv, "Topic level summary generation using BERT induced Abstractive Summarization Model," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 747-752, doi: 10.1109/ICICCS48265.2020.9120997.

[10]M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi: 10.1109/TASLP.2020.3006731.

[11]Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 2005-2010, doi: 10.1109/IAEAC50856.2021.9390683.

[12]Z.Liang, J.Du, Y.Shao, H.Ji,Gated Graph Neural Attention Networks for abstractive summarization,Neurocomputing,Volume 431,2021,Pages 128-136,ISSN 0925-2312,

[13] Su Haijv, Wang Yongcheng. Automatic compilation of abstracts of Chinese scientific and technological documents [J]. Journal of the China Society for Scientific and Technical Information. 1989, 8 (6):433–439.

[14] Du Xiuying. Multi — document Automatic Summarization Based on Clustering and Semantic Similarity Analysis on Cloud Computing Platform [J]. JOUＲNAL OF INTELLIGENCE:167–172.

[15] Shen Huadong, Peng Dunlu. AM-BRNN: Automatic Text Summarization Extraction Model Based on Deep Learning [J].Journal of Chinese Computer Systems. 2018, 39 (06): 1184–1189.

[16] Hou Liwe,Hu Po,Cao Wenlin. Automatic Chinese Abstractive Summarization WithTopical Keywords Fusion [J]. Acta Automatica Sinica. 2019, 045 (003): 530–539.

[17] Li Dazhou, Yu Pei, Gao Wei, Ma Hui. Abstractive Chinese text summarization based on encoder-decoder model[J].Computer Engineering and Design, 2021, 42(03):696-702.

[18] Jiang Xiaoping. Research on chinese Abstractive Summarization via Fusing Topic Information[D].Cental China Normal University,2020.

[19] Cai Zhongxiang, Sun Jianwei. News Text Summarization Model Integrating Pointer Network.[J]. Journal of Chinese Computer Systems,2021,42(03):462-466.

[20] Yang Tao, Xie Qing, Liu Yongjian, Liu Pingfeng.Research on Topic-Aware Long Text Summarization Algorithm.[J].Computer Engineering and Applications,2022,58(20):165-173.

[21]J. Wang and Z. Yan, "A Hybrid Attention Based Autoencoder for Chinese Text Abstractive Summarization," 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2020, pp. 2141-2145, doi: 10.1109/ITAIC49862.2020.9339084.