Postprint: Automatic News Summarization Generation for People' s Daily Corpus

Authors: Liang Yuan, Wang Dongbo, Huang Shuiqing

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] This study investigates the mainstream news media People' s Daily corpus, aiming to provide ideas and practical support for automatic text summarization research, with subsequent application to news and related text information processing, thereby contributing to research on knowledge aggregation services and information acquisition pathways. [Method/Process] Using the segmented People's Daily corpus from January 2015, June 2015, and January 2016 in the New Era People' s Daily corpus (NEPD) as experimental data, this research investigates extractive automatic summarization algorithms based on TF-IDF, TextRank, etc., as well as generative automatic summarization models based on pointer generator networks, and conducts analysis and evaluation of the summarization results. [Results/Conclusion] The study designs news extractive automatic summarization algorithms tailored for the People' s Daily corpus, constructs a pointer generator network model for news generative automatic summarization oriented toward the People' s Daily corpus, and evaluates the experimental results through ROUGE metrics (including three metrics: ROUGE-1, ROUGE-2, and ROUGE-L), thereby providing specific ideas for the application of the People' s Daily segmented corpus and offering corpus support and practical backing for research on news automatic summarization systems.

Full Text

Automatic Summary Generation of News for People's Daily Corpus

Liang Yuan[1],[2], Wang Dongbo[1],[2], Huang Shuiqing[1],[2]

[1]College of Information Management, Nanjing Agricultural University, Nanjing 210095
[2]Research Center for Humanities and Social Computing, Nanjing Agricultural University, Nanjing 210095

Abstract

[**Purpose/Significance**] This study focuses on the mainstream news media corpus of *People' s Daily*, aiming to provide ideas and practical support for automatic text summarization research that can be applied to news and related text information processing, thereby contributing to knowledge aggregation services and information access research. [**Method/Process**] The experimental corpus consists of segmented articles from *People' s Daily* in January 2015, June 2015, and January 2016, drawn from the New Era People's Daily (NEPD) corpus. Based on extractive summarization algorithms such as TF-IDF and TextRank, as well as a generative summarization model built on the pointer-generator network, we conducted experiments and evaluated the summarization results. [**Result/Conclusion**] We constructed an extractive automatic summarization algorithm for news and a pointer-generator network model for generative summarization, both tailored to the *People' s Daily* corpus. Evaluation using Rouge metrics (Rouge-1, Rouge-2, and Rouge-L) yielded fruitful experimental results, providing specific approaches for applying the segmented *People' s Daily* corpus and offering both data support and practical foundations for news automatic summarization system research.

Keywords: *People' s Daily*; Extractive Automatic Summarization; Generative Automatic Summarization; NEPD; Pointer-Generator Networks

1. Introduction

The explosive growth of online information has made information access more convenient while simultaneously reducing information utilization efficiency and increasing reading costs. Automatic summarization technology addresses this problem by compressing and refining information, providing an auxiliary means to improve knowledge acquisition efficiency [1]. Currently, automatic summarization primarily employs two approaches: extractive and generative. Extractive summarization, which emerged earlier, has matured through years of research by numerous scholars. With the introduction of machine learning into the field, generative summarization has once again gained development momentum.

News serves as a crucial channel for documenting social issues, disseminating information, and accessing current events. As the official newspaper of the Central Committee of the Communist Party of China, *People' s Daily* represents the primary medium for state-people communication and a bridge for cultural exchange both domestically and internationally, making research on its corpus particularly significant. Our experimental data originates from the New Era People' s Daily Segmented Corpus (NEPD) [2], which contains manually segmented and proofread *People' s Daily* articles, constituting high-quality refined data [3]. The NEPD enables rapid calculation of word frequencies and facilitates subsequent data preprocessing for various text processing tasks.

Responding to evolving news consumption trends and the need to distill large volumes of news text, this study focuses on the *People' s Daily* corpus to implement both extractive and generative news summarization algorithms, evaluating the results to improve information utilization efficiency, reduce reading costs, and provide methodological insights for automatic summarization technology and evaluation approaches.

2. Literature Review

Early research by Mo Yan [4] and Wang Yongcheng [5] introduced concepts and algorithms for automatic literature abstracting and knowledge extraction. Subsequently, Wang Yongcheng and Xu Huimin [6] and Wang Zhijin [7] proposed and designed the OA Chinese literature automatic summarization system and a sentence-selection-based automatic text summarization system, respectively, while also reviewing the history, development, and significance of Chinese literature summarization. Shi Lei and Wang Yongcheng [8] studied English literature automatic summarization systems.

Building upon this foundation, automatic summarization research has advanced rapidly with continuous algorithmic innovation. Xiong Jiao et al. [9], Zhang Xiaodan and Hu Xuegang [10], Liu Xinghan and Huo Hua [11], Ji Wenqian et al. [12], Zeng Zhejun [13], and Liu Jing and Xiao Lu [14] employed graph models, vector space models, mutual information, continuous LexRank algorithms, and dependency parsing graph models for automatic summarization. Wang Shuai et al. [15] proposed a two-stage long text summarization method combining graph models and recurrent neural networks, testing it on large-scale financial texts. Wu Yun et al. [16] enhanced the frequency of feature words similar to titles to calculate word frequency matrices and sentence similarity, developing a word-sentence collaborative extraction algorithm. Chen Chen et al. [17] applied word-sentence collaborative ranking to propose a graph model-based automatic summarization algorithm. Ding Jianli et al. [18] employed multi-dimensional word embedding patterns with dual encoders incorporating dual-channel semantics for short text summarization. Feng Dujuan et al. [19] constructed a CGAtten-GRU model based on dual-encoder network architecture, achieving good results on large-scale Chinese short text summarization. Liao Tao et al. [20] proposed an event network representation for text events to perform automatic summarization. Xu Xintao et al. [21] improved the TextRank algorithm by integrating Doc2Vec and K-means to optimize thematic sentence extraction. Chen Haihua et al. [22] fused citation context features with support vector machine (SVM) models for academic text summarization. Huang Shuiqing et al. [23] designed an automatic text summarization system for computer science literature. Zhang Han and Zhao Yuhong [24] constructed a semantic graph-based model for medical multi-document summarization through normalized extraction of text and semantic relationships. Chen Zhimin et al. [25] and Li Fang and He Tingting [26] approached summarization from an information retrieval perspective, generating summaries based on user query expansion

and query-document collections.

Among these algorithms, research on topic segmentation and multi-feature fusion has been particularly prominent. Zhang Zheming et al. [27] proposed a high-quality long text summarization model combining topic awareness and communication agents. Chen Yanmin et al. [28] developed a topic-content fusion method producing coherent and fluent summaries through anaphora resolution. Luo Fang et al. [29] improved graph models by incorporating latent Dirichlet allocation (LDA) topic models to mine thematic semantic information, measuring and extracting text through multi-dimensional features including topic, statistical, and inter-sentence similarity features for deep thematic semantic mining. Du Xiuying [30] constructed a MapReduce automatic summarization architecture based on clustering and semantic similarity analysis for large-scale multi-document summarization, improving time performance, compression effectiveness, and summary quality. However, these methods primarily focus on extractive summarization, leaving considerable research space for generative approaches.

With the rapid development of big data and artificial intelligence, traditional automatic summarization is evolving from extractive to generative methods to produce higher-quality, more natural and fluent abstracts. Recent years have seen deep learning applied to generative summarization research. Wu Shixin et al. [31] built a generative model by introducing semantically aligned neural networks to the Sequence-to-Sequence model with attention, Pointer, and Coverage mechanisms. Fang Xu et al. [32] proposed a Chinese short text summarization algorithm combining core word correction with long short-term memory (LSTM) networks. Tang Xiaobo and Zhai Xiapu [33] improved the PageRank algorithm and employed a hybrid machine learning model incorporating sentence vectorization, classifier classification, sentence group division, and sentence reorganization for multi-document summarization. Tan Jinyuan et al. [34] and Zhang Kejun et al. [35] fused multiple deep learning models to propose the Bi-MulRnn+ and BERT-PGN generative summarization models, respectively, effectively improving accuracy and fluency. Li Weiyong et al. [36] and Xiao Yuanjun and Wu Guowen [37] also conducted research and implementation of Chinese generative summarization models based on deep learning.

In the news domain, Guan Lihe [38] analyzed the 思路 and 流程 of Chinese online news automatic summarization with experimental verification. Han Yongfeng et al. [39] addressed information redundancy in summarization, proposing an improved multi-document news summarization method based on event extraction. Shen Zhou et al. [40] established a news literature topic extraction rule base and built a rule-based automatic summarization system for news. Li Mengshuang et al. [41] proposed a summarization algorithm based on mutual information calculation of word-sentence semantic features for topic segmentation and key sentence extraction. Wang Kaixiang and Ren Ming [42] designed a query-based news summarization algorithm and compared it with six methods including TF-IDF, TextRank, and LDA. Huang Xiaojiang et al. [43] automati-

cally generated comparative summaries for news topics based on collaborative graph ranking. Ke Xiu and Wang Huilin [44] fused multiple algorithms including anaphora resolution, text external features, and graph ranking to implement multi-document news summarization for Chinese, English, and Bengali. Ye Lei et al. [45] similarly employed graph ranking, proposing a multi-feature fusion method for Chinese-Vietnamese bilingual news summarization.

Beyond news, user-generated content on platforms like Weibo and forums also holds significant research value, though automatic summarization faces challenges from high redundancy and noise characteristics [46], which scholars continue to address [47-50].

This literature review reveals that automatic summarization technology has continuously evolved with technological advances and user needs, progressing from rule-based and statistical methods to deep learning, and from plain text to dynamic video. News automatic summarization remains critically important for meeting people' s news consumption needs in fast-paced modern life. However, applied research has concentrated primarily on extractive methods, with domain-specific, highly accurate generative models and systems still underdeveloped. Therefore, this study investigates automatic summarization for the *People*' *s Daily* corpus, completing summarization tasks through both traditional and deep learning algorithms to address the time-consuming nature of reading long news texts and low information utilization efficiency, while also providing support for knowledge aggregation services in news media and offering new perspectives for news dissemination and cultural inheritance.

3. Methodology

Natural language processing (NLP) remains a traditionally active research field, sustained not only by new technologies but also by its reputation as one of the "most difficult AI subfields." The automatic summarization task represents a major challenge that researchers continuously strive to overcome, particularly as speed-reading becomes an essential reading mode. Current automatic summarization methods are primarily categorized as extractive or generative. Extractive methods mainly apply keyword-sentence ranking, while generative methods rely more on deep learning models. In our experiments, extractive summarization employs traditional algorithms based on keyword frequency for sentence weighting and TextRank, while generative summarization references the Chinese Text-Summarizer-Pytorch-Chinese model built on pointer-generator networks [50].

3.1 Extractive Automatic Summarization Our extractive approach primarily uses word frequency and clustering to identify keywords, then scores sentences based on these keywords, ranking them to determine summary sentences. This method originates from H.P. Luhn' s IBM paper *The Automatic Creation of Literature Abstracts* [51], which proposed using clusters to represent

keyword clustering results—where a cluster is a sentence fragment containing multiple keywords, as illustrated in Figure 1 [Figure 1: see original paper].

The cluster weight calculation formula [52] is:

$$\text{Cluster Weight} = \frac{\text{Number of Keywords in Cluster}}{\text{Cluster Length}} \quad (1)$$

Here, cluster length refers to the number of words in the sentence fragment. For example, from our *People' s Daily* corpus:

Segmented text: "经过/全国/各族/人民/共同/努力/，/ '/十二五/' /规划/圆满/收官/，/广大/人/民/群众/有/了/更/多/获得感"

If we consider "' 十二五' 规划圆满收官" as one cluster with length 6, where "十二五", "规划", and "收官" are keywords, and "广大人民群众有了更多获得感" as another cluster with length 8, where "人民", "群众", and "获得感" are keywords, the cluster weights would be $3/6 = 0.5$ and $3/8 = 0.375$, respectively. Sentences are then ranked by weight, and a threshold is applied (set to 10 in this study, extracting the top 10 most important sentences) to generate the final summary. Similar to TextRank, this algorithm derives from PageRank, treating sentences as web pages and using sentence similarity matrices with a set threshold to obtain high-scoring sentences as the summary—an unsupervised extractive method.

3.2 Generative Automatic Summarization The pointer-generator network for automatic summarization is illustrated in Figure 2 [Figure 2: see original paper]. This model concentrates on important vocabulary in the text through self-attention mechanisms to generate new words. Rather than copying original words, it balances vocabulary probabilities, word distributions, and attention distributions to determine candidate word weights and obtain final distributions.

Currently, few Chinese-oriented pointer-generator network models exist. Therefore, we adapted the Text-Summarizer-Pytorch-Chinese construction approach by adjusting the pre-training corpus to NEPD and updating the vocabulary accordingly for NEPD-specific pre-training and model building.

4. Experiments

4.1 Experimental Corpus As described by its editors, "*People' s Daily* is an authoritative and serious comprehensive daily newspaper that responds to news events with its editorial strength, reporting major domestic and international events" [54]. As a mainstream media outlet serving as the "eyes, ears, and voice" and a "bridge and bond" for the Party and people, its textual information value is self-evident. The *People' s Daily* corpus has long been an important data source for researchers. The Institute of Computational Linguistics at Peking University built the first large-scale modern Chinese annotated corpus [55], and in 2019, Nanjing Agricultural University' s Research Center for Humanities and

Social Computing processed articles from 2015-2018 to construct the New Era People' s Daily Corpus (NEPD) [56].

This study uses the January 2015, June 2015, and January 2016 segmented news corpora from NEPD as experimental data. The original corpus is shown in Figure 3 [Figure 3: see original paper].

4.2 Data Preprocessing According to our research needs, we segmented each news article from the source corpus. The processed texts, shown in Figure 4 [Figure 4: see original paper], were prepared for subsequent summarization extraction and generation. After data cleaning (including removal of data without standard summaries), we obtained 2,748 news articles from January 2015, 2,796 from June 2015, and 2,748 from January 2016, totaling 6,292 data entries used as our research subjects.

4.3 Experimental Environment and Parameter Settings For generative model training and testing, we used Ubuntu 16.04 with 16GB DDR4 RAM, 4GB GDDR5 VRAM, an Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz, and an NVIDIA Quadro K1200 GPU. Model parameters are listed in Table 1 .

Table 1: Generative Automatic Summarization Model Parameters

Parameter	Value
hidden_{dim}	512
emb_{dim}	256
batch_{size}	200
max_{{{enc}}_{{steps}}}	100
max_{{{dec}}_{{steps}}}	20
beam_{size}	4
min_{{{dec}}_{{steps}}}	3
vocab_{size}	40000
rand_{{{unif}}*{{init}}*}{mag}	0.02
trunc_{{{norm}}*{{init}}*}{std}	1e-4
eps	1e-12
max_{iterations}	5000000

4.4 Experimental Design Our research comprises two parts: extractive automatic summarization algorithm and generative automatic summarization model, both tailored for the *People' s Daily* corpus.

The extractive algorithm involves eight steps: (1) obtain segmented *People' s Daily* corpus; (2) preprocess texts (remove special characters, spaces, blank lines); (3) remove stopwords and perform word frequency statistics (no segmentation needed as NEPD is pre-segmented); (4) calculate sentence weights using features including title keywords and sentence length; (5) rank sentences by

weight; (6) select appropriate threshold to extract summary sentences; (7) generate summary; (8) evaluate against standard summaries using Rouge-1, Rouge-2, and Rouge-L.

The generative model construction involves seven steps: (1) obtain segmented corpus; (2) preprocess texts and adjust training format; (3) build pre-trained model using *People*' *s Daily* corpus; (4) incorporate features: introduce custom vocabulary from NEPD as a user dictionary and add title features; (5) train generative model with parameter tuning and iterative training; (6) generate summaries using final model; (7) evaluate with Rouge metrics.

5. Results and Analysis

5.1 Evaluation Metrics As no standard summary corpus exists for *People*' *s Daily*, we used keyword frequency-based extractive results and Baidu AI Cloud' s news summary API outputs as reference standards. Baidu' s service automatically extracts text based on deep semantic analysis models to generate news summaries of specified lengths [57].

Rouge (Recall-Oriented Understudy for Gisting Evaluation) is a common metric for evaluating summarization and machine translation [58]. It calculates similarity between standard and automatic summaries:

$$\text{Rouge-N} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}n \in S} \text{Count}(\text{gram}_n)} \quad (2)$$

The denominator counts n-grams in reference summaries, while the numerator counts overlapping n-grams. We used Rouge-1, Rouge-2, and Rouge-L, where Rouge-L employs longest common subsequence (LCS):

$$\text{LCS}(X, Y) = \text{length of longest common subsequence} \quad (3)$$

$$R_{\text{lcs}} = \frac{\text{LCS}(X, Y)}{m} \quad (4)$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X, Y)}{n} \quad (5)$$

where m and n are lengths of reference and automatic summaries (typically word counts), and R_{lcs} and P_{lcs} are recall and precision. The β value is typically large, making Rouge-L primarily consider recall.

The internal P, R, F metrics are calculated as:

$$\text{Precision} = \frac{\text{Correctly identified pairs}}{\text{Correctly identified pairs} + \text{Incorrectly identified pairs}} \quad (6)$$

$$\text{Recall} = \frac{\text{Correctly identified pairs}}{\text{Correctly identified pairs} + \text{Unidentified pairs}} \quad (7)$$

$$F\text{-Measure} = \text{Harmonic mean of precision and recall} \quad (8)$$

5.2 Extractive Summarization Results In extractive experiments, we scored sentences using keyword frequency and cluster-based keyword extraction, ranking them to extract summaries. Using frequency-based extractive results as reference standards and cluster-based results as automatic summaries, we performed Rouge evaluation. Partial results are shown in Figure 5 [Figure 5: see original paper], with comprehensive results in Table 2 .

Table 2: Extractive Automatic Summarization Evaluation Results

Metric	Rouge-1	Rouge-2	Rouge-L
Mean	0.8447	0.8257	0.8446

Overall, extractive summarization performed well, providing adequate summarization of original texts. However, since Rouge calculates similarity between complete sentences, mismatched extractions yield extremely low similarity scores, potentially causing artificially low Rouge metrics. Future work will adjust reference summary accuracy and improve evaluation methods.

5.3 Generative Summarization Results For generative experiments, we preprocessed all news corpora and accessed Baidu AI Cloud' s news summary API. Due to platform text length limitations, we filtered 7,967 compliant news texts. We combined corpora from January 2015, June 2015, and January 2016 for training. Custom vocabulary from NEPD was introduced to improve training effectiveness and summary fluency. Evaluation results are shown in Table 3 .

Table 3: Generative Automatic Summarization Evaluation Results

Metric	Rouge-1 (%)	Rouge-2 (%)	Rouge-L (%)
Mean	0.8447	0.8257	0.8446

Table 4 presents sample outputs. Different algorithms produce varying summary content, but overall fluency remains acceptable. Extractive summaries,

drawn directly from original sentences, exhibit higher intra-sentence readability but lower inter-sentence coherence. They contain richer content but poorer generalization, with redundancy and lower sentence association. Generative summaries demonstrate semantic understanding, producing more concise content better aligned with news summary characteristics and more flexible original text summarization, though occasionally suffering from word repetition or incomplete coverage.

Rouge metrics, while intuitive and 简洁, have limitations. They reward original expressions with higher scores [53], typically yielding higher scores for extractive than generative summaries. This is particularly problematic for generative evaluation. Future research will explore multiple evaluation approaches, including human-generated reference summaries and manual scoring, for more accurate assessment.

6. Conclusion

Automatic summarization distills long texts into concise versions, enabling rapid browsing and comprehension while saving reading costs and improving knowledge utilization efficiency—particularly valuable given today' s massive information resources. Using NEPD corpora from January 2015, June 2015, and January 2016, we designed keyword frequency-based and cluster-based extractive algorithms and constructed a pointer-generator network model for generative summarization. Both achieved good Rouge scores and produced coherent summaries.

Future work will refine algorithms, improve models, enhance reusability, and develop better evaluation methods incorporating multiple text features, human-generated reference datasets, and manual scoring to improve summary fluency and readability.

References

[1] Wang Shuai, Zhao Xiang, Li Bo, et al. TP-AS: A two-phase automatic summarization method for long texts[J]. Journal of Chinese Information Processing, 2018, 32(6): 71-79.

[2] Huang Shuiqing, Wang Dongbo. Construction, performance, and application of the New Era People' s Daily segmented corpus (Part 1)—Corpus construction and evaluation[J]. Library and Information Service, 2019, 63(22): 5-12.

[3] Huang Shuiqing, Wang Dongbo. Diachronic analysis of word frequency in Central No. 1 Document based on People' s Daily corpus[J]. Journal of Library and Information Science in Agriculture, 2020, 32(3): 4-9.

[4] Mo Yan, Wang Yongcheng. Automatic compilation of Chinese literature abstracts[J]. New Technology of Library and Information Service, 1993(3): 10-12.

[5] Wang Yongcheng. Automatic compilation of literature abstracts and automatic extraction of knowledge[J]. New Technology of Library and Information Service, 1993(3): 13-28.

[6] Wang Yongcheng, Xu Huimin. OA Chinese literature automatic summarization system[J]. Journal of the China Society for Scientific and Technical Information, 1997(2): 49-53.

[7] Wang Zhijin. Sentence-selection-based automatic text summarization method and its evaluation[J]. New Technology of Library and Information Service, 1998(1): 46-51, 58.

[8] Shi Lei, Wang Yongcheng. Research on English literature automatic summarization system[J]. Journal of the China Society for Scientific and Technical Information, 1999(6): 504-508.

[9] Xiong Jiao, Wang Mingwen, Li Maoxi, et al. Multi-document automatic summarization based on term-sentence-document three-layer graph model[J]. Journal of Chinese Information Processing, 2014, 28(6): 201-207.

[10] Zhang Xiaodan, Hu Xuegang. Research on redundancy processing in automatic summarization based on vector space model[J]. Journal of Hefei University of Technology (Natural Science), 2010, 33(9): 1355-1358.

[11] Liu Xinghan, Huo Hua. Text automatic summarization based on mutual information[J]. Journal of Hefei University of Technology (Natural Science), 2014, 37(10): 1198-1203.

[12] Ji Wenqian, Li Zhoujun, Chao Wenhan, et al. An improved automatic abstracting system based on LexRank algorithm[J]. Computer Science, 2010, 37(5): 151-154, 218.

[13] Zeng Zhejun. Research on multi-document automatic summarization optimization algorithm based on continuous LexRank[J]. Computer Applications and Software, 2013, 30(10): 209-212, 217.

[14] Liu Jing, Xiao Lu. Research on multi-topic text summarization based on dependency parsing[J]. Journal of Intelligence, 2014, 33(6): 167-171.

[15] Wang Shuai, Zhao Xiang, Li Bo, et al. TP-AS: A two-phase automatic summarization method for long texts[J]. Journal of Chinese Information Processing, 2018, 32(6): 71-79.

[16] Wu Yun, Yang Changchun, Mei Jiajun, et al. Word-sentence collaborative automatic summarization extraction method[J]. Computer Engineering and Design, 2018, 39(9): 2776-2779, 2838.

[17] Chen Chen, Zhang Lu, Wu Zhi' ang. Automatic summarization algorithm based on word-sentence collaborative ranking[J]. Journal of Jiangsu University (Natural Science Edition), 2016, 37(04): 443-449.

[18] Ding Jianli, Li Yang, Wang Jialiang. Short text automatic summarization method based on dual encoders[J]. Computer Applications, 2019, 39(12): 3476-3481.

[19] Feng Dujuan, Yang Lu, Yan Jianfeng. Research on text automatic summarization based on dual-encoder structure[J]. Computer Engineering, 2020, 46(6): 60-64.

[20] Liao Tao, Liu Zongtian, Wang Xianchuan. Research on event-based text representation method[J]. Computer Science, 2012, 39(12): 188-191.

[21] Xu Xintao, Chai Xiaoli, Xie Bin, et al. Chinese text summarization extraction based on improved TextRank algorithm[J]. Computer Engineering, 2019, 45(3): 273-278.

[22] Chen Haihua, Huang Yong, Zhang Jiong, et al. Research on academic text automatic summarization technology based on citation context[J]. Digital Library Forum, 2016(8): 43-48.

[23] Huang Shuiqing, Li Zhiyan, Liang Gang. Research and implementation of automatic summarization system for computer science literature[J]. Library and Information, 2006(3): 93-97.

[24] Zhang Han, Zhao Yuhong. Construction of medical multi-document summarization extraction model based on semantic graph[J]. Library and Information Service, 2017, 61(8): 112-119.

[25] Chen Zhimin, Jiang Yi, Zhao Yao. Automatic summarization technology based on user query expansion[J]. Computer Application Research, 2011, 28(6): 2188-2190.

[26] Li Fang, He Tingting. Research on query-oriented multi-mode automatic summarization[J]. Journal of Chinese Information Processing, 2011, 25(2): 9-14.

[27] Zhang Zheming, Ren Shuxia, Guo Kaijie. Text summarization model combining topic awareness and communication agents[J]. Journal of Xidian University, 2020, 47(3): 77-83.

[28] Chen Yanmin, Wang Xiaolong, Liu Yuanchao, et al. An automatic summarization method based on article topic and content[J]. Computer Engineering and Applications, 2004(33): 194-196.

[29] Luo Fang, Wang Jinghang, He Daosen, et al. Research on text automatic summarization method fusing topic features[J]. Computer Application Research, 2021, 38(1): 129-133.

[30] Du Xiuying. Multi-text automatic summarization method based on clustering and semantic similarity analysis[J]. Journal of Intelligence, 2017, 36(6): 167-172.

[31] Wu Shixin, Huang Degen, Li Jiuyi. Research on generative automatic summarization based on semantic alignment[J]. Acta Scientiarum Naturalium

Universitatis Pekinensis, 2021, 57(1): 6.

[32] Fang Xu, Guo Yi, Wang Qi, et al. Core word corrected Seq2Seq short text summarization[J]. Computer Engineering and Design, 2018, 39(12): 3610-3615.

[33] Tang Xiaobo, Zhai Xiapu. Multi-document automatic summarization based on hybrid machine learning model[J]. Information Studies: Theory & Application, 2019, 42(2): 145-150.

[34] Tan Jinyuan, Diao Yufeng, Qi Ruihua, et al. Chinese news text automatic summarization generation based on BERT-PGN model[J]. Computer Applications, 2021, 41(1): 127-132.

[35] Zhang Kejun, Li Weinan, Qian Rong, et al. Text automatic summarization scheme based on deep learning[J]. Computer Applications, 2019, 39(2): 311-315.

[36] Li Weiyong, Liu Bin, Zhang Wei, et al. A Chinese generative automatic summarization method based on deep learning[J]. Journal of Guangxi Normal University (Natural Science Edition), 2020, 38(2): 51-63.

[37] Xiao Yuanjun, Wu Guowen. Research and implementation of automatic summarization generation algorithm based on Gensim[J]. Computer Applications and Software, 2019, 36(12): 131-135.

[38] Guan Lihe. Research on Internet news text automatic summarization[J]. Computer Engineering and Design, 2007(14): 3518-3520, 3545.

[39] Han Yongfeng, Xu Xuyang, Li Bicheng, et al. Multi-document automatic summarization for online news based on event extraction[J]. Journal of Chinese Information Processing, 2012, 26(1): 58-66.

[40] Shen Zhou, Wang Yongcheng, Xu Yizhen, et al. Research and practice of an automatic summarization system for news literature[J]. Computer Engineering, 2000(9): 70-73.

[41] Li Mengshuang, Zan Hongying, Jia Huizhen. Research on Weibo news automatic summarization based on multi-features and Ranking SVM[J]. Journal of Zhengzhou University (Science Edition), 2017, 49(2): 44-48.

[42] Wang Kaixiang, Ren Ming. Research on query-based news multi-document automatic summarization technology[J]. Journal of Chinese Information Processing, 2019, 33(4): 93-100.

[43] Huang Xiaojiang, Wan Xiaojun, Xiao Jianguo. Comparative news automatic summarization based on collaborative graph ranking[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 31-38.

[44] Ke Xiu, Wang Huilin. Construction and implementation of multi-language multi-document automatic summarization system based on hybrid method[J]. Research on Library Science, 2013(2): 66-72.

[45] Ye Lei, Yu Zhengtao, Gao Shengxiang, et al. Multi-feature fusion method for Chinese-Vietnamese bilingual news summarization[J]. Journal of Chinese

Information Processing, 2018, 32(12): 84-91.

[46] Gao Yongbing, Wang Yu, Ma Zhanfei. Research on personal event automatic summarization based on CR-PageRank algorithm[J]. Computer Engineering, 2016, 42(11): 64-68.

[47] Chen Zhuoqun, Wang Ping. Research on extractive summarization method for Chinese Weibo[J]. Information Science, 2015, 33(3): 130-134.

[48] Gao Yongbing, Zhong Zhenhua, Wang Yu, et al. Research on Chinese Weibo automatic summarization technology based on hybrid method[J]. Computer Engineering & Science, 2016, 38(6): 1199-1204.

[49] Jia Xiaoting, Wang Mingyang, Cao Yu. Research on Weibo text summarization generation based on weighted topic distribution representation[J]. Journal of Northeast Normal University (Natural Science Edition), 2020, 52(1): 69-74.

[50] Text-Summarizer-Pytorch-Chinese[EB/OL]. [2021-07-07]. https://github.com/LowinLi/Text-Summarizer-Pytorch-Chinese.

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

[52] Peng Min, Gao Binlong, Huang Jimin, et al. Weibo automatic summarization based on high-quality information extraction[J]. Computer Engineering, 2015, 41(7): 36-42.

[53] Ruan Yifeng. Application of TF-IDF and cosine similarity (Part 3): Automatic summarization[EB/OL]. [2021-07-07]. http://www.ruanyifeng.com/blog/2013/03/automatic_{summariz

[54] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[J]. arXiv preprint arXiv:1704.04368, 2017.

[55] Cheng Shuang. Three changes in *People' s Daily* expansion[EB/OL]. [2021-07-07]. https://baike.baidu.com/redirect/7e44WWpuHPxVjlVjuIAMGFXvpzQ0nX6dtcm9N58nsqPgZqu9Xe5

[56] Yu Shiwen, Zhu Xuefeng, Duan Huiming. Processing specifications for large-scale modern Chinese annotated corpus[J]. Journal of Chinese Information Processing, 2000(6): 58-64.

[57] Huang Shuiqing, Wang Dongbo. Review of domestic corpus research[J]. Journal of Information Resources Management, 2021, 11(3): 4-17, 87.

[58] Baidu AI Cloud. News summary[EB/OL]. [2021-07-07]. https://cloud.baidu.com/product/nlp_{apply}/ne

[59] Lin C Y. ROUGE: a package for automatic evaluation of summaries[C]//Text summarization branches out, 2004: 74-81.

Note: Figure translations are in progress. See original paper for figures.