Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 1267336, 12 pages
https://doi.org/10.1155/2023/1267336

WILEY | Hindawi

Research Article

SentMask: A Sentence-Aware Mask Attention-Guided Two-Stage Text Summarization Component

Rui Zhang,[1,2] **Nan Zhang**,[3] **and Jianjun Yu**[1]

[1]*Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China*
[2]*University of Chinese Academy of Sciences, Beijing 100083, China*
[3]*PetroChina Changqing Oilfield Company Technical Monitoring Center, Xi'an, Shanxi 710000, China*

Correspondence should be addressed to Jianjun Yu; yujj@cnic.ac.cn

Received 9 February 2023; Revised 22 July 2023; Accepted 31 July 2023; Published 22 August 2023

Academic Editor: B. B. Gupta

The text summarization task aims to generate succinct sentences that summarise what an article tries to express. Based on pretrained language models, combining extractive and abstractive summarization approaches has been widely adopted in text summarization tasks. It has been proven to be effective in many existing pieces of research using extract-then-abstract algorithms. However, this method suffers from semantic information loss throughout the extraction process, resulting in incomprehensive sentences being generated during the abstract phase. Besides, current research on text summarization emphasizes only word-level comprehension while paying little attention to understanding the level of the sentence. To tackle this problem, in this paper, we propose the SentMask component. Taking into account that the semantics of sentences that are filtered out during the extraction process is also worth considering, the paper designs a sentence-aware mask attention mechanism in the process of generating a text summary. By applying the extractive approach, the paper first selects the most essential sentences to construct the initial summary phrases. This information leads the model to modify the weights of the attention mechanism, which provides supervision for the generative model to ensure that it focuses on the sentences that convey important semantics while not ignoring others. The final summary is constructed based on the key information provided. The experimental results demonstrate that our model achieves higher ROUGE and BLEU scores compared to other baseline models on two benchmark datasets.

1. Introduction

With the rapid increase in the number of articles and papers, we have found ourselves drowned in the sea of documents. The time-consuming and energy-draining reading process can be avoided by creating a concise abstract of a text and transmitting the main concept to the reader. But summarizing articles automatically is a difficult process as it necessitates models to rewrite a long article into a concise and fluent version while preserving the essential information. In the area of automatic text summarization, extractive and abstractive methods are two primary paradigms. To produce a summary, the extractive[1] techniques select the salient phrases or sentences exactly from the original source, whereas the abstractive [2] techniques generate new phrases and sentences from scratch. However, because relevant information is spread throughout all sentences rather than contained in a few, extractive models suffer from a lack of semantics and cohesiveness in summary sentences, as well as redundancy in certain summary sentences. On the other hand, abstractive summarization models suffer from the slow encoding of long documents and the unreliability of the generated summaries.

Recently, some researchers have tried to combine these two methods in an extract-then-abstract way [3, 4]. The work [3] proposes a hybrid framework HYSUM for text summarization, which maintains salient content by switching rewriting sentences and copying sentences according to the degree of redundancy. The work [4] provides a hybrid abstractive-extractive method, which scans a document, produces prominent textual fragments that highlight its main ideas, and selects the important sentences

by calculating the BERTScore. These models design a two-stage pipeline to pick out salient sentences from a source document first and then rewrite the extracted sentences into a complete summary. However, most research using the extract-then-abstract framework generates summaries based solely on the extracted sentences, which loses robustness. In many cases, significant content might be filtered by the extraction model, causing severe information loss in the generation process.

Furthermore, it is difficult to comprehend and generalise articles due to their rigorous grammatical statements. To maintain the consistency of professional grammatical definitions and logic within original sentences, it is vital to preserve sentence-level information and semantics in summaries, which have also been ignored in previous works.

To overcome both of these issues while combining the benefits of both paradigms, in this paper, we propose SentMask, a novel sentence-aware mask attention-guided two-stage text summarization component, adaptively reducing the attention weight of filtered sentences by training neural networks. Taking Figure 1 for example, the existing methods generate the summary according to the selected sentences extracted by the extractors only. However, the sentences also contain some information, which should not be lost, such as "adverse events." Thus, the paper utilizes these sentences by reducing rather than deleting their attention weights.

An extractive summary is to extract important sentences to form a summary to achieve the function of summarizing the full text. During the extraction process, the model fully considers semantic information between sentences. The generative summary is to generate orderly words, form sentences, and then form a summary to highly summarise the entire article. During the generation process, the semantic information between words is fully considered by the model, but the emphasis on the semantic information between sentences is weakened. In order to make full use of the semantic information between each word and sentence, we employ an extractor to extract the initial summary and an abstractor to abstract the final summary. Therefore, our model takes into account both word-level and sentence-level information in the text generation process. Unlike selecting important words in other works that separate the semantics of the whole, the paper uses an extractor to select essential information at the sentence level, faithfully preserving the semantics of the whole sentence. In this way, with the above issues solved, our model can avoid syntactic errors and incoherent errors in summary sentences and ensure that the generated phrases are flexible and stable. To better leverage the results of the extractor algorithm and preserve the necessary global information, the paper proposes a sentence-aware mask attention mechanism in our model.

The paper evaluates the efficacy of our semisupervised and supervised SentMask models, respectively. The semi-supervised SentMask model consists of the TextRank algorithm [5] and sequence-to-sequence model (Seq2Seq) [6], while the supervised SentMask model consists of the MemSum algorithm [7] and BART [8] model. The paper leverages the extractor algorithm to extract important sentences for summarization. Based on its results, the paper then masks other sentences by reducing rather than deleting their attention weights. The noise reduction capability of our model is demonstrated by the weight reduction of the information in trivial sentences, which, to some extent, relatively increases the weight of important information.

The following are our primary contributions:

(1) The paper proposes a brand-new two-stage hybrid abstractive and extractive summary method. While acquiring the information of the salient sentences generated by the extractor, our abstractor also extracts knowledge in a specific way for the nonsalient sentences. Our method is implemented in semi-supervised and supervised versions, which include unsupervised and supervised extractors, respectively.

(2) The paper proposes a sentence mask module, a sentence-aware mask attention mechanism, and a mask-aware copy mechanism. The sentence mask module aims to transform a sample input into a mask matrix. The sentence-aware mask attention mechanism reduces the nonsalient sentences' attention weight rather than losing its information. The mask-aware copy mechanism copies only words from salient sentences since there could be noise throughout the article.

(3) The paper extensively evaluates SentMask on two benchmark datasets. The results of the experimental evaluation show that SentMask outperforms the current state-of-the-art in these evaluations.

2. Related Work

2.1. Traditional Summarization. Several traditional summarization approaches for automatic summary generation have been advanced over the years, incorporating a variety of statistical-based [9], topic-based [10], graph-based [5], and semantic-based [11] techniques. For instance, the work [9] brings improvements by involving sentence position, sentence length, and keyword sentence features. The work [10] proposes a term frequency-inverse document frequency algorithm, which measures the importance of keywords based on their frequency of occurrence and uses it to assess each sentence. The abstract is extracted from the highest-scoring sentences. Biased TextRank [5] is a method for capturing meaning closeness between graph nodes and a target text that depends on document representation models and similarity measurements. The latent semantic analysis [11] is an unsupervised technique that encodes text semantics based on the observed cooccurrence of words.

Traditional unsupervised text summarization models do not require any training data and generate the summary by accessing only the target documents. However, these traditional methodologies do the summarization task using manual design features, which shows poor generalization ability for new data.

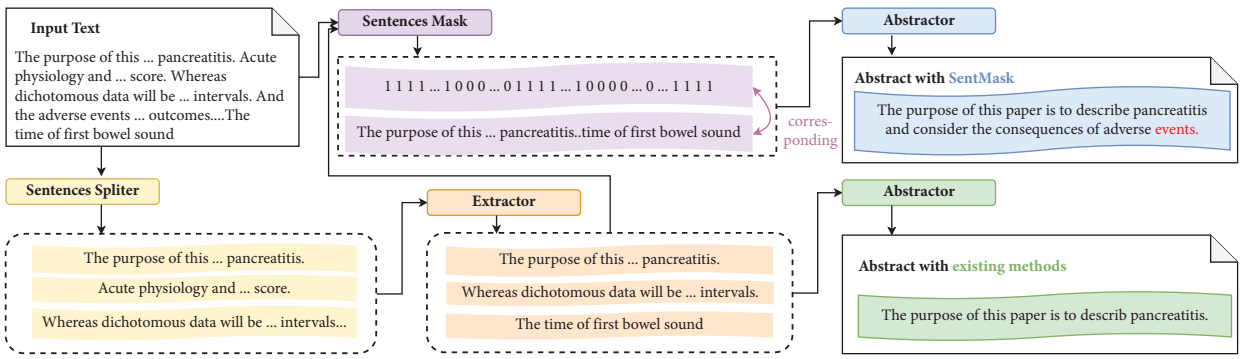

FIGURE 1: Sample summary of an article from the MS2 dataset corpus. Existing methods generate the summary based on the sentences selected by the extractor. While the paper reduces the attention weights of nonsalient sentences by using a mask attention matrix with a sentence-aware masked attention mechanism, the sentence mask module in our model is a transformation of a sample input into a mask matrix.

2.2. Neural Networks Summarization. The two most common types of study are extractive summarization and abstractive summarization. Extractive summarization methods commonly construct an encoder-decoder architecture, with the graph attention network [12] as an encoder and autoregressive [13] or nonautoregressive [14] decoders. The work [7] proposes a multistep extractive summariser based on reinforcement learning-based Markov decision processes, which considers information from the current extraction history.

In recent years, pretraining has been used in several varieties of transformer architecture in various ways, including encoder-only pretraining models like XLNet [15], decoder-only pretraining models like GPT [16], and encoder-decoder pretraining models like T5 [17] and BART [8]. For instance, the work [18] distills large pretrained sequence-to-sequence transformer models into smaller ones for faster inference and with the least amount of performance loss.

Two-stage document summarizing systems have been developed in recent studies. The first stage of this framework usually involves extracting some segments of the original text, and the second stage involves selecting or modifying these segments. There are various extract-then-abstract summarization methods such as extract-then-rewrite and extract-then-compress. In extract-then-rewrite models, the method [19] employs a coarse-to-fine approach inspired by humans, extracting all relevant sentences first and then decoding them simultaneously. The work [20] introduces a novel training signal that employs reinforcement learning to directly maximise summary-level ROUGE scores. In extract-then-compress models, the model [21] selects phrases from the document, identifies plausible compressions based on constituent parses, and rates those compressions using a neural network model to construct the final summary. The work [22] proposes a method for learning to select sentence singletons and pairs, which would subsequently be employed by an abstractive summariser to build a sentence-by-sentence summary, with singletons compressed and pairs fused.

Previous research using the extract-then-abstract framework generates summaries based solely on the extracted sentences, which loses semantic information in the filtered sentences, causing a severe information loss. To that end, the paper designs a sentence-aware mask attention-guided two-stage text summarization component, which captures the gist of the text.

3. Materials and Methods

In this section, the paper introduces our sentence-aware extract-then-abstract summarization framework in detail as illustrated in Figure 2. It consists of four components: (1) An extractor, an importance-aware content selection component that utilizes the TextRank or MemSum [7] algorithm to extract and organize salient sentences. (2) An abstractor, a Seq2Seq [6] or BART- [8] based abstract generation component with sentence-aware mask attention mechanism that compresses and rephrases both the extracted sentences and the original article to a succinct summary. (3) The sentence-aware mask attention mechanism, a modified version of the attention weight mechanism by masking the nonsalient sentences. (4) The mask-aware copy mechanism, a modified version of the copy mechanism by copying words from the salient sentences rather than the whole article. The paper describes these components in detail as follows.

3.1. Extractor. First, we split the article into sentences. Let x denote the original sentences of the article, which consists of a sequence of sentences $(x = u_1, u_2, \ldots, u_m)$. Each u_i consists of a sequence of words $u_i = (w_1^i, w_2^i, \ldots w_\alpha^i)$.

These sentences are constructed as a directed graph represented by a sentence similarity matrix with the TextRank algorithm or input to a multistep episodic Markov decision process with historical awareness using the MemSum algorithm. After the extractor algorithm, a score is calculated for each sentence, which represents the "importance" of the sentences. The sentences are sorted in reverse order of the score, and the first K sentences with the

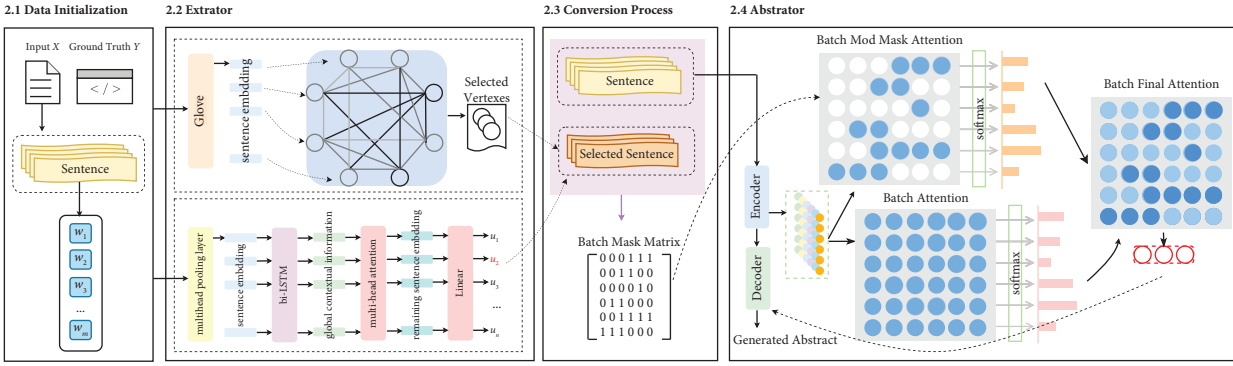

Figure 2: The architecture of the SentMask model.

highest scores are chosen to be the draft as the input of the abstractor to form the final summary.

x^E denotes the initial sentences extracted by the extractor algorithm, which belong to the sentences in x. $x^E = \text{extractor}(x)$, where extractor = {TextRank, Mem Sum}. The paper redescribed $x^E = \{r_1, r_2, \ldots, r_j\}$, where $r_j = u_q, q \in \{1, 2, \ldots, m\}$.

So far, the paper is discussing the sentence level. The extractor helps us to preserve the whole sentence semantics. The paper then converts this information to the word-level since the Seq2Seq and BART models would take the word-level information into account.

The paper utilizes a sentence mask module to transform a sample input into a mask matrix. The transformation of the input of the SentMask model is shown in Figure 3.

x^{mask} indicates whether the word is in the selected sentences. $x^{\text{mask}} = (u_1^{\text{mask}}, u_2^{\text{mask}}, \ldots, u_m^{\text{mask}})$, where $u_i^{\text{mask}} = (m_1^i, m_2^i, \ldots, m_\alpha^i), m_k^i$ is shown as follows:

$$m_k^i = \begin{cases} 1, i \in \{1, 2, \ldots, j\}, & k \in \{1, 2, \ldots, \alpha\}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where x^{mask} will be the essential component for us to perform a sentence-aware mask attention mechanism, as it conveys information about how important the word is. To make it clear, the paper reformulates $x = (x_1, x_2, \ldots, x_m)$ and $x^{\text{mask}} = (x_1^{\text{mask}}, x_2^{\text{mask}}, \ldots, x_m^{\text{mask}})$.

3.2. Abstractor. After obtaining the initial salient textual fragments representing the source article's key points by the extractor, the paper generated the summary with the assistance of these extracted sentences.

The paper uses a pretrained word representation to map each token to a vector. Then, the paper utilizes an abstractor to encode and decode the whole article, abstractor $\in$ {Seq2Seq, BART}. The decoder is initialized with the encoder's last hidden state. In Seq2Seq, our encoder and decoder are GRU-based. h_t is the encoder's hidden state and s_t is the decoder's hidden state at the time step t. The context vector is $c_t = \sum_i a_{t,i} h_i$.

$$\begin{aligned} h_t &= \text{GRU}(h_{t-1}, x_t), \\ s_t &= \text{GRU}(s_{t-1}, y_{t-1}, c_t). \end{aligned} \tag{2}$$

In the BART, our encoder and decoder are transformer architecture. h^E is the hidden state of the encoder, and h_t^D is the hidden state of the decoder at the time step t.

$$\begin{aligned} h^E &= \text{BART}^{\text{enc}}(x), \\ h_t^D &= \text{BART}^{\text{dec}}(y_{t-1}, h^E), \end{aligned} \tag{3}$$

where y_{t-1} is the word generated in the last step.

The paper uses a sentence-aware attention mechanism in both of our abstractors. In addition, the paper utilizes a mask-aware copy mechanism in the Seq2Seq.

3.3. Sentence-Aware Mask Attention Mechanism. Based on the attention mechanism, the paper proposes a sentence-aware attention mechanism in this paper, which is employed both in semisupervised and supervised modes. $a_{t,i}$ is the attention score obtained by our sentence-aware mask attention mechanism. It consists of two parts: standard word-level attention and sentence-aware masked attention on the sentence level. The word-level attention is calculated by the associated phrase attention. In the masked sentence attention, the paper forces the model to focus on the important sentences extracted by the extractor algorithm. By combining such attention scores together with a hyperparameter as the weight, the paper can not only emphasize information from important sentences but also not lose semantics in other sentences. The attention score calculation process in Seq2Seq is shown as follows:

$$a_{t,j}^{\zeta} = \text{softmax}\left(\mu_1^T \tanh\left(W_2 s_{t-1} + W_3 h_i\right) + \eta_i\right). \tag{4}$$

The attention score calculation process in the BART is shown as follows:

$$a_{t,j}^{\zeta} = \text{softmax}\left(Q \bullet K + \eta_i\right), \tag{5}$$

when $\zeta = \text{attn}$, η_i is the default attention mask. When $\zeta = \text{mask}$, η_i is shown as follows:

$$\begin{aligned} \eta_i &= \begin{cases} 0, & x_i^{\text{mask}} = 1, \\ \xi, & x_i^{\text{mask}} = 0, \end{cases} \\ a_{t,i} &= a_{t,j}^{\text{attn}} * \epsilon + a_{t,j}^{\text{mask}} * (1 - \epsilon), \end{aligned} \tag{6}$$

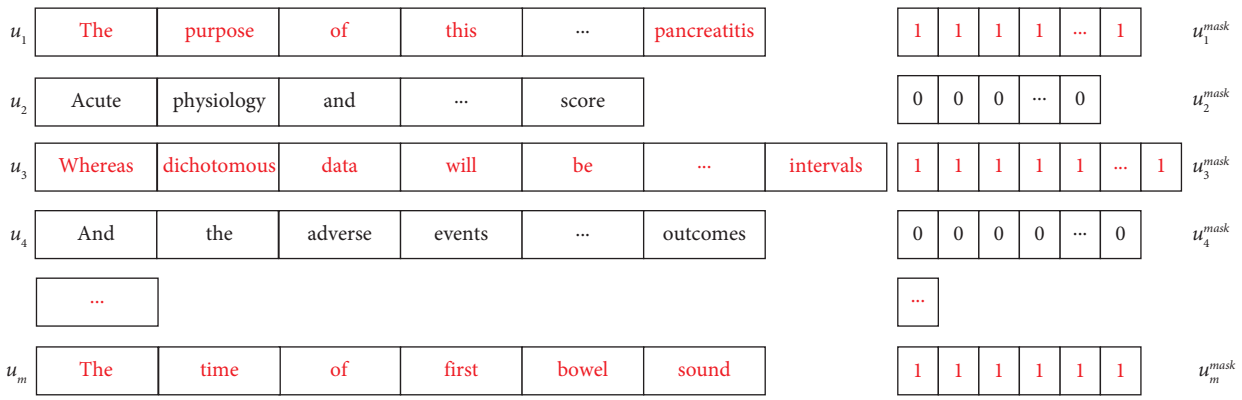

FIGURE 3: The transformation of a sample input, x, is on the left representing the original article and x^{mask} is on the right representing the corresponding sentence-aware mask matrix. The sentences marked in red (x^E) are the sentences extracted by the extractor.

where ξ and ϵ are the hyperparameters. The extension of the generation sources encourages the integrity of the sentence and increases the probability of correctness.

For summary output, the final vocabulary distribution in BART at time step t is $P = \text{Dense}(h_t^D)$, where Dense is a dense layer, while the preliminary vocabulary distribution in Seq2Seq at time step t is defined as follows:

$$P_{\text{vocab}} = \text{softmax}\left(\text{FeedForward} \bullet \left(\text{concat}\left(s_t, c_t, y_{t-1}\right)\right)\right). \tag{7}$$

3.4. Mask-Aware Copy Mechanism. The copy mechanism in the Seq2Seq, according to [23], uses the encoder's representation of words to select a word in the inputs instead of choosing from the whole vocabulary. When dealing with important words, this technique may be more reliable than generating from all vocabulary. Due to the hidden state of a word being governed by its full context and lexical auxiliary feature collectively, the model can consistently produce great terms in the target vocabulary. The paper makes a modification to the original copy mechanism. The paper only copies words from important sentences since there could be noise throughout the article. By limiting the scope, the model can easily find the most possible word to generate. P_{copy} is calculated as follows:

$$\begin{aligned} a_{t,j}^{\text{copy}} &= \text{softmax}\left(\mu_2^T \tanh\left(W_4 s_{t-1} + W_5 h_i\right) + \eta_i\right), \\ c_t^{\text{copy}} &= \sum_i a_{t,i}^{\text{copy}} h_i, \\ P_{\text{copy}} &= \sigma\left(W_6 \bullet \left(\text{concat}\left(c_t^{\text{copy}}, y_{t-1}\right)\right)\right), \end{aligned} \tag{8}$$

where μ_2^T, W_4, W_5, and W_6 are trainable parameters. And σ means the sigmoid function.

The final prediction is obtained by merging the copy probability and the output of the decoder.

$$\begin{aligned} P &= \left(1 - P_{\text{copy}}\right) P_{\text{vocab}} + P_{\text{copy}} \sum_i a_{t,i} \delta\left(y_t | x_i\right), \\ \delta\left(y_t | x_i\right) &= \begin{cases} 1, & \text{if } y_t == x_i, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{9}$$

In conclusion, our SentMask model extends the Seq2Seq and BART models, respectively, with an important sentence-guided masked attention strategy that enables the model to leverage both word-level information and sentence-level information for final sequence generation. Taking advantage of containing more condensed semantics at the word-level and keeping the original sentence grammar at the sentence level, our SentMask model promotes the capacity of capturing the gist of the input text, either semisupervised or supervised.

4. Results and Discussion

4.1. Dataset. To comprehensively investigate our proposed model, we employ two benchmark datasets for evaluation, which are common options in previous research, including the Multi-Document Summarization of Medical Studies benchmark dataset (MS2) and the AESLC dataset. The paper declares both of them are open access, where the MS2 dataset can be downloaded at https://paperswithcode.com/dataset/ms-2 and the AESLC dataset can be downloaded at https://github.com/ryanzhumich/AESLC. The statistical details of the two datasets are shown in Table 1. The following are brief summaries of these benchmark datasets.

4.1.1. MS2 [24]. MS2 dataset is a scientific literature dataset with about 470k pages and 20k summaries. The paper removes the contents that are excessively long or too short, and 20,434 papers are ultimately acquired as our corpus, with 16,112 documents for training, 2,277 for validation, and 2,045 for testing.

4.1.2. AESLC [25]. The AESLC dataset is obtained from the Enron dataset, including many emails from staffers in the Enron Corporation, which are composed of 517,401 e-mail messages from 150 user mailboxes. After filtering and deduplicating, the paper obtains the final AESLC dataset.

4.2. Implementation and Evaluation Details. This method is suitable for any encoder-decoder model based on a neural network, including pretrained language models. In this

Table 1: Details of the statistics of datasets.

Model	Train	Validation	Test	All
MS2	16,112	2,277	2,045	20,434
AESLC	14,436	1,960	1,906	18,302

paper, we implement our SentMask based on Seq2Seq and BART, respectively, which is sufficient to demonstrate the effectiveness of the method. The paper sets $\xi = -1e6$. The paper uses Pytorch to implement our model.

To demonstrate the performance of the proposed SentMask model, the paper compares the SentMask model to many baselines with the same model size for a fair comparison, including the Lead3 algorithm, TextRank algorithm, GenCompareSum model [4], Seq2Seq model, Presumm model [26], Global Encoding model [27], Pointer-Generator model [23], Transformer [28], AESLC baseline [25], and BART [8].

There are some descriptions of the baselines as follows.

4.2.1. Lead3 Algorithm. Lead3 algorithm takes the top K sentences.

4.2.2. TextRank Algorithm. The TextRank algorithm determines each sentence's score based on how similar the sentences are to one another and then selects the top K scoring sentences.

4.2.3. GenCompareSum Model [4]. GenCompareSum model is a hybrid extraction method, which generates salient text fragments representing their main points and selects the most important sentences in the document by calculating using BERTScore.

4.2.4. Seq2Seq Model. Seq2Seq is an encoder-decoder architecture, which consists of LSTM or GRU.

4.2.5. Presumm Model [26]. The Presumm model is based on the BERT model, which can express the semantics of the document and obtain the representation of the sentence and improve the quality of the summary through the fine-tuning method.

4.2.6. Global Encoding Model [27]. The Global Encoding model is a Seq2Seq model, which employs a gated convolutional unit in the encoder for global encoding.

4.2.7. Pointer-Generator Model [23]. Pointer-generator is an encoder-decoder model solving the OOV problem by controlling the pointer to make the model copy the token from the original context.

4.2.8. Transformer [28]. It is a brand-new, uncomplicated network architecture, which consists of attention mechanism techniques.

4.2.9. AESLC Baseline [25]. AESLC baseline is a multisentence extractor and a multisentence abstractor.

4.2.10. BART [8]. BART is a transformer-based model, which employs a bidirectional encoder with a number of denoising pretraining objectives.

For the evaluation of the quality of the experiment, the paper comprehensively evaluates the quality of the summary generated by these baseline models from both intelligent and human evaluation perspectives. Automated overview evaluation metrics, including ROUGE [29] and BLEU [30], are used to evaluate the quality of text summarization. In particular, the BLEU evaluation metric is an enhanced N-gram assessment metric, and its N-gram weights can be defined here to conveniently fit the models for different purposes and more accurately determine the consistency of the model.

4.3. Automated Evaluation. The experimental results on the MS2 and AESLC datasets are shown in Tables 2 and 3, respectively. The results show that the proposed SentMask model performs remarkably well in two text summarization datasets, demonstrating the effectiveness of our masked sentences attention mechanism.

Meanwhile, the improvements confirm that not only further refining information from the original text can be captured by the structure of a multilayer neural network but also the expression capacity that enables the model to generate summaries with few grammatical errors is improved by adding updated encoding information.

4.4. Human Evaluation. To further assess the quality of the summaries produced by the SentMask model, the paper conducted a human evaluation using three typical indicators, informativeness, fluency, and faithfulness. The following are brief summaries of these human evaluation metrics.

4.4.1. Informativeness. The informativeness of the summary is determined by how accurately it summarises the material in the original article.

4.4.2. Faithfulness. Faithfulness evaluates how well the facts in the summary match those of the original article.

4.4.3. Fluency. The summary's fluency is determined by how few serious grammatical faults it contains.

The paper hires five native English speakers and randomly chooses 300 news stories from the MS2 and AESLC datasets to evaluate the summaries of these baseline models and the SentMask model on three different aspects. The score ranges from 1 (poor) to 5 (outstanding).

Table 4 findings demonstrate that, in terms of informativeness, fluency, and faithfulness, our SentMask model outperforms other baseline models, which illustrates the value of the sentence-aware mask attention mechanism.

Table 2: Details of the ROUGE and BLEU evaluation values in the baseline models on the MS2 dataset.

Models	RG1	RG2	RGL	BE	BE1	BE2	BE3	BE4
Semisupervised								
Lead3	25	10.5	21.78	4.77	12.12	5.76	3.47	2.14
TextRank	18.54	5.58	16.03	2.88	10.29	3.38	1.83	1.07
GenCompareSum	29.83	14.22	24.96	7.71	14.15	7.22	2.99	2.99
TextRank + Seq2seq	24.58	10.78	20.15	6.22	12.33	9.28	3.29	2.09
SentMask	**46.36**	**24.56**	**41.81**	**16.35**	**39.47**	**22**	**13.19**	**8.04**
Supervised								
Seq2Seq	34.45	19.25	31.64	8.27	15.67	9.54	6.02	4.41
Pointer-gen	35.34	16.28	31.43	7.63	18.31	9.08	5.64	3.62
Global encoding	29.67	14.53	24.34	12.54	30.1	15.01	10.32	6.17
Presumm	35.99	16.88	30.72	13.69	32.4	16.32	10.35	6.41
BART	52.97	33.41	49.15	28.85	54.22	34.83	25.64	19.41
SentMask	**55.38**	**35.97**	**51.8**	**30.94**	**55.51**	**36.75**	**27.69**	**21.48**

The best values in the metric are in bold. RG1: ROUGE-1; RG2: ROUGE-2; RGL: ROUGE-L; BE: BLEU; BE1: BLEU1; BE2: BLEU2; BE3: BLEU3; BE4: BLEU4.

Table 3: The ROUGE results on the AESLC dataset.

Model category	Models	ROUGE-1	ROUGE-2	ROUGE-L
Semisupervised	TextRank	11.32	3.88	10.14
	GenCompareSum	10.14	3.85	9.53
	TextRank + Seq2Seq	10.09	3.71	9.45
	SentMask	**22.3**	**10.78**	**22.11**
Supervised	Transformer	15.04	7.39	14.93
	Pointer-gen	17.02	5.45	15.78
	AESLC	23.67	10.29	23.44
	BART	27.24	14.04	26.79
	SentMask	**27.58**	**14.15**	**27.06**

Table 4: The human evaluation results. The score is calculated on an average of the scores for 300 news articles from the MS2 and AESLC datasets that were supplied by 5 volunteers. The score of each volunteer, which goes from 1 to 5, is the assessment of every news article.

Models	MS2 dataset			AESLC dataset		
	INFOR	FAITH	FLU	INFOR	FAITH	FLU
Semisupervised						
TextRank	3.584	3.5666	3.5726	2.5893	2.572	2.582
GenCompareSum	3.6293	3.6306	3.2687	2.6493	2.6506	2.646
TextRank + Seq2Seq	3.6266	3.6393	3.636	2.6406	2.644	2.6526
SentMask	**3.8326**	**3.8393**	**3.8433**	**2.8106**	**2.8233**	**2.8273**
Supervised						
AESLC	3.682	3.6853	3.6833	2.7033	2.7066	2.7046
Transformer	3.816	3.862	3.8366	2.8533	2.8166	2.8206
Pointer-gen	3.8146	3.83	3.822	2.8066	2.862	2.8586
BART	3.8693	3.8833	3.8533	2.88	2.8873	2.8746
SentMask	**3.9333**	**3.9373**	**3.9406**	**2.9466**	**2.954**	**2.9493**

The best values in the metric are in bold. INFOR: informativeness; FAITH: faithfulness; FLU: fluency.

4.5. Ablation Study. To obtain a more scientifically accurate explanation, an ablation study is conducted by removing some components of our model to verify their contribution. The paper conducts the ablation study with the semisupervised model and a supervised model, respectively, on the MS2 dataset. The paper conducts several experiments and ablation tests as follows.

4.5.1. SentMask-T. It is our proposed semisupervised model. The sentences are first generated by the TextRank algorithm and then passed through the proposed SentMask neural network.

4.5.2. TextRank. TextRank is a graph-based ranking model for natural language processing, which finds the most relevant sentences in an article.

4.5.3. SentMask-C. It is our proposed supervised model. The MemSum algorithm generates the initial selected sentences and passes through the proposed SentMask neural network.

4.5.4. MemSum. MemSum is a historical-aware multistep episodic Markov decision process algorithm.

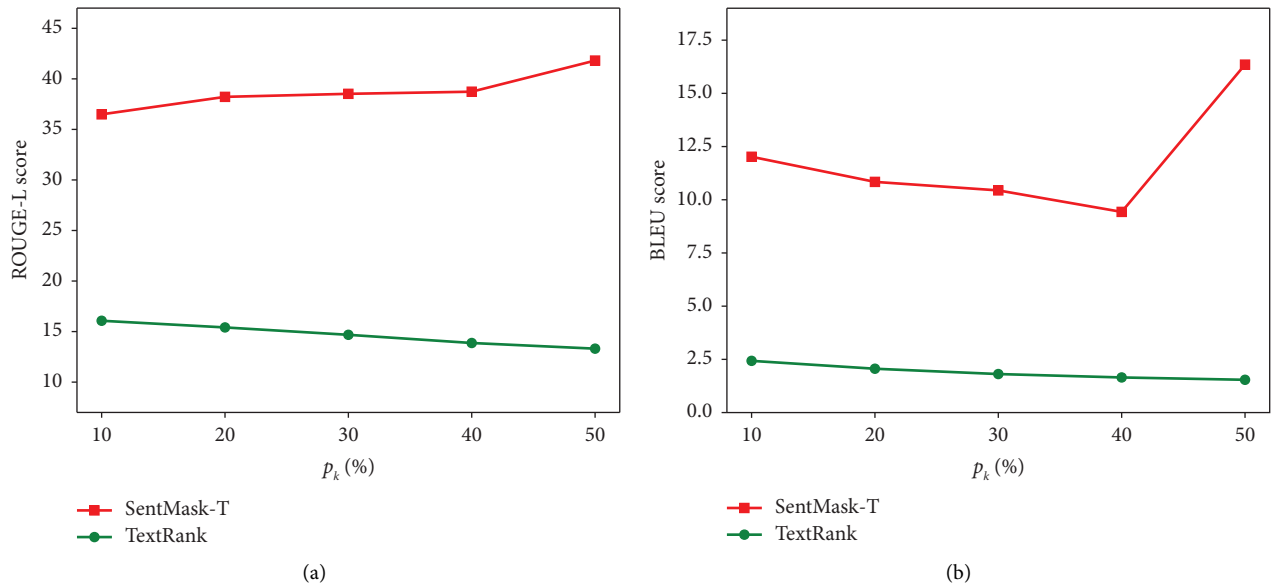

FIGURE 4: The ROUGE-L score and BLEU score of ablation models of semisupervised model with different P_K, where the x-axis represents the different P_K, and the y-axis represents the value of scores. (a) ROUGE score. (b) BLEU score.

To investigate how the hyperparameters affect the model's performance, the paper tries different hyperparameter settings in our ablation study. An essential hyperparameter is the number of sentences with the highest scores extracted by the extractor algorithm, K.

The paper performs a set of experiments with a different selection of K to uncover its influence on the quality of the generating sentence. There are two ways to control K in the extractor algorithm, one is to control the percentage of selected sentences and the other is to set K itself. The settings of the two ways are described as follows.

percent = p_k in the extractor algorithm; the first p_k of sentences is selected as subsequent input sentences and as nonmasked sentences. In our experiments, the paper tries different $p_k \in \{50\%, 40\%, 30\%, 20\%, 10\%\}$.

top = K in the extractor algorithm; the first K sentences are selected as the subsequent input sentences and as the nonmasked sentences. The paper tries different $K \in \{5, 4, 3, 2, 1\}$.

For the semisupervised model, the ROUGE-L score and the BLEU score of the ablation models with different p_k are shown in Figure 4. The ROUGE-L score and BLEU score of ablation models with different top $-K$ are illustrated in Figure 5. For the supervised model, the ROUGE-L score and BLEU score of ablation models with different p_k are shown in Figure 6. The ROUGE-L score and BLEU score of ablation models with different top $-K$ are illustrated in Figure 7.

Overall, the ablation models, either the semisupervised model or the supervised model, perform poorly in terms of the ROUGE-L score and BLEU score, demonstrating the effectiveness of the sentence-aware masked attention mechanism in our SentMask model. From the eight figures, with different K, the line trend of the results of the semisupervised SentMask model is more turbulent, while that in the supervised SentMask model is relatively stable. Thus, the performance of the semisupervised SentMask model is influenced by the parameter K significantly, while the supervised model is slightly influenced. In addition, selecting the proper number of sentences is a crucial decision for our model. Comparatively speaking, it can be observed that the best setting of hyper-parameter K is to select the first 50% of sentences of source articles, either the semisupervised model or the supervised model.

4.6. Effect of the Hyper-Parameter. To demonstrate our model robustness with different parameters, the paper tries different ϵ from 0.6 to 0.95 for the semisupervised model and the supervised model on the MS2 dataset. According to the results in Figure 8, the proposed SentMask performs well regardless of the value of ϵ. SentMask-T performs best when $\epsilon = 0.9$ in the MS2 dataset and SentMask-C performs best when $\epsilon = 0.95$ in the MS2 dataset. Note that the model mainly carries out the task of generating text abstracts, so the proportion of information from the attention mechanism represented by the masked sentences strategy should be less than that of the original attention mechanism.

4.7. Case Study. Table 5 shows an example of summaries generated by different models.

In this example, the original article provides verification of acupuncture's efficacy and safety in relieving abdominal pain and distension associated with acute pancreatitis. The primary idea of this paper is definitely about acupuncture's high efficacy and safety, and the research object is abdominal pain and distension for acute pancreatitis.

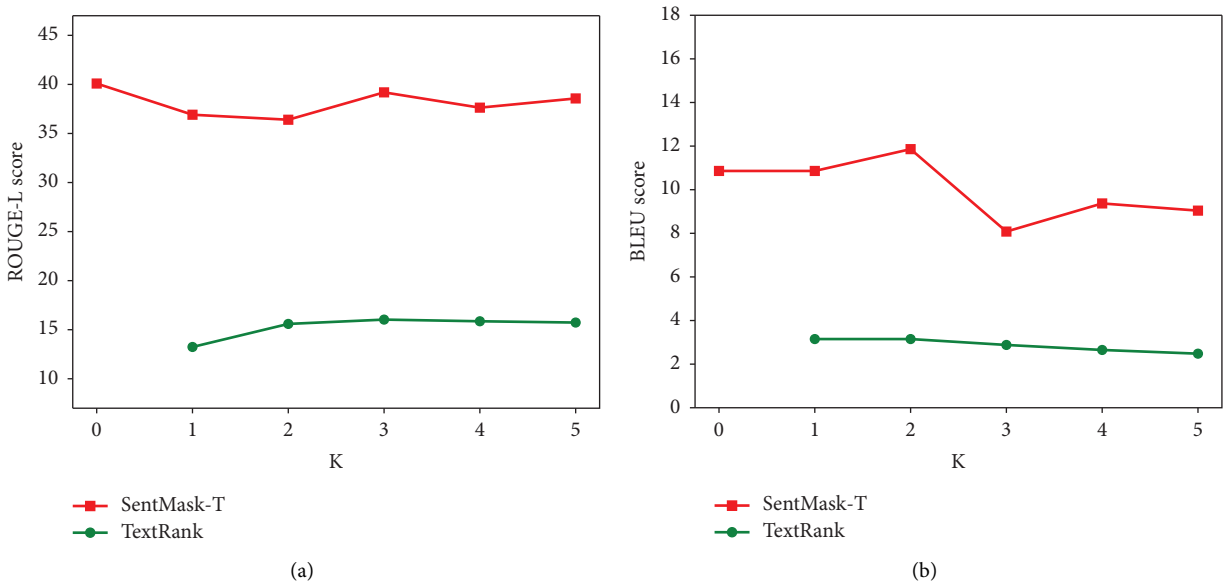

Figure 5: The ROUGE-L score and BLEU score of ablation models of semisupervised model with different top – K, where the x-axis represents the different top – K, and the y-axis represents the value of scores. (a) ROUGE score. (b) BLEU score.

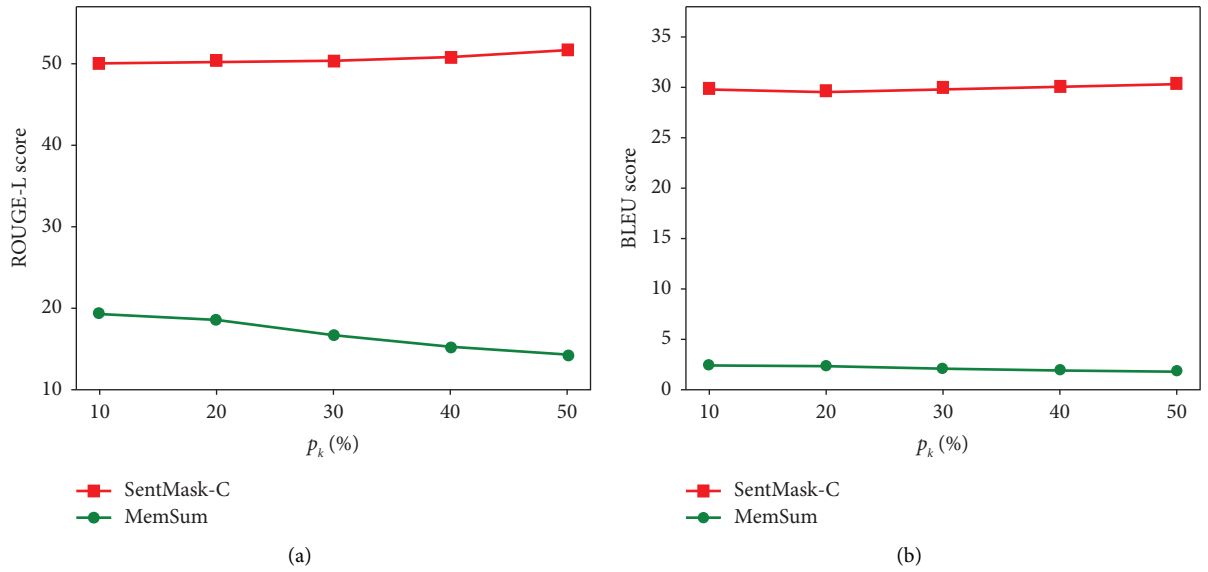

Figure 6: The ROUGE-L score and BLEU score of ablation models of supervised model with different P_K, where the x-axis represents the different P_K, and the y-axis represents the value of scores. (a) ROUGE score. (b) BLEU score.

However, the baseline models generate an inappropriate summary to varying degrees. In detail, the summary of the Lead3 algorithm contains duplicate information that does not represent the true abstract of this article, such as "Methods and Analysis"

The TextRank algorithm has a risk of ranking redundant sentences high and generates condensed sentences that are semantically similar sentences, such as "safety of acupuncture" which appears twice in the summary text.

The Seq2Seq model creates a summary that solely comprises information related to acupuncture, not the efficacy or safety of acupuncture. Furthermore, it made the mistake of redundantly repeating the word "acupuncture." The pointer network model generates an excessive number of words, emphasizing "acupuncture's effect" rather than "its efficacy and safety." Meanwhile, the trial method does not need to be included in the abstract of the paper. According to the summary of the Global Encoding model, "orthostatic hypotension and cardiovascular" is a component of the entire text, but not the main information. The main objective of the summary given by the Presumm model is "home-based ventilation in intensive care," which is inconsistent. The summary generated by the BART model focuses on "pancreatitis" rather than "efficacy," which is inappropriate.

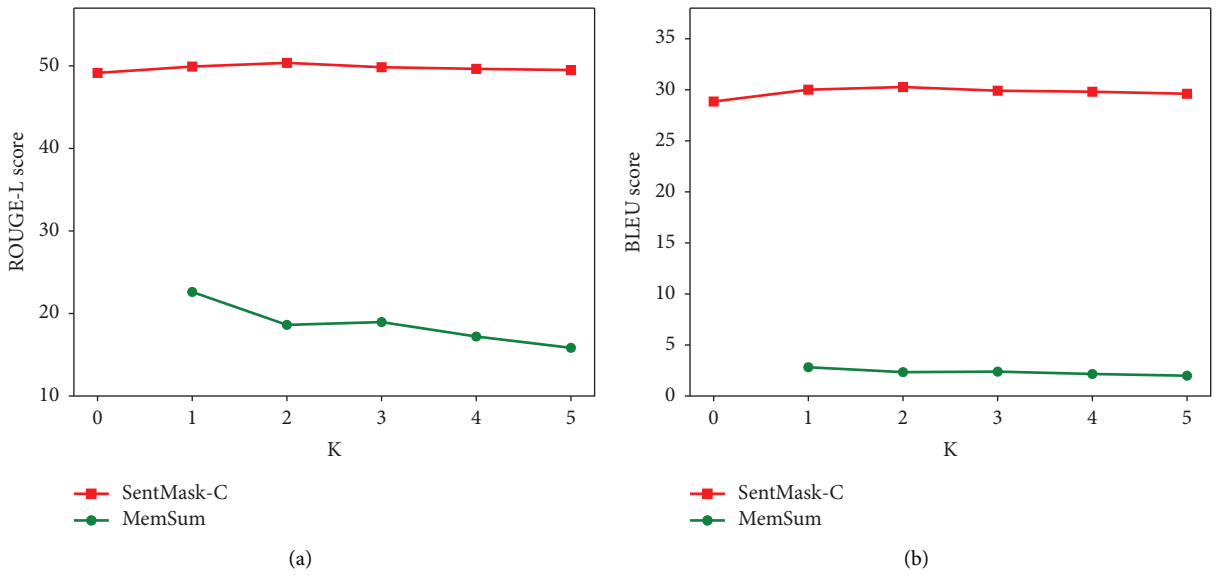

Figure 7: The ROUGE-L score and BLEU score of ablation models of supervised model with different top − K, where the x-axis represents the different top − K, and the y-axis represents the value of scores. (a) ROUGE score. (b) BLEU score.

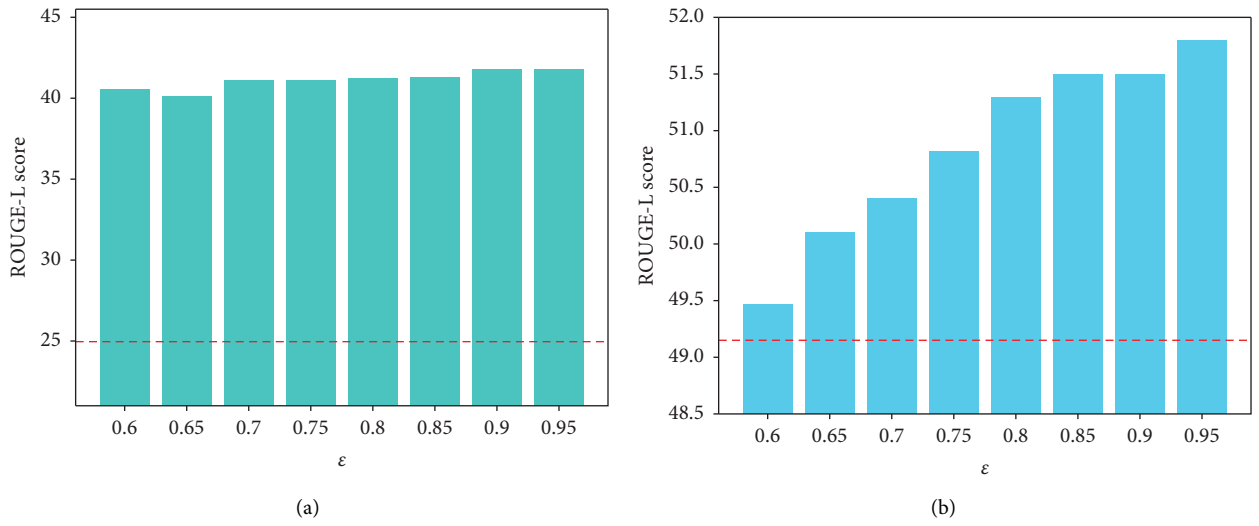

Figure 8: Results of SentMask model with different parameter ϵ on the MS2 dataset, where the x-axis represents different parameters ϵ, and the y-axis represents ROUGE-L scores. The red horizontal line on the figure is the line of the baseline result on different model categories. (a) SentMask-T model. (b) SentMask-C model.

Compared with these baseline models, the summary of our model is more coherent and semantically relevant to the source text. Our model focuses on information on the efficacy and safety of acupuncture rather than itself and points out that this is a systematic review and meta-analysis in its generated summary. Meanwhile, all the words generated from our model are the target words of the standard dataset, maintaining a high degree of conciseness.

Therefore, our model can better consider the grammatical word-level and sentence-level appearances simultaneously by masking the sentences to advise the generator. This indicates that the masked sentence attention in our

Table 5: Comparison of the output of 8 summarization models on MS2 dataset.

Reference (truncated): efficacy and safety of acupuncture on relieving abdominal pain and distension for acute pancreatitis.
Lead3: the purpose of this study is to evaluate the efficacy and safety of acupuncture on relieving abdominal pain and distension in acute pancreatitis. *Methods and Analysis*. we will electronically search PubMed, MEDLINE, Embase, and Web of Science.
TextRank: efficacy and safety of acupuncture and safety of acupuncture abdominal pain in acute pancreatitis pain: a systematic review and meta-analysis.
Seq2Seq: acupuncture on relieving abdominal pain: a systematic review.
Pointer-gen: the effect of acupuncture on abdominal pain and distension in acute pancreatitis: systematic review and meta-analysis of randomized controlled trials and trial sequential.
Global encoding: association between orthostatic hypotension and cardiovascular risk of cardiovascular disease: a systematic review and meta-analysis.
Presumm: effect of home-based ventilation in intensive care: a systematic review of randomized controlled trials.
BART: acupuncture for acute pancreatitis: protocol of a systematic review and meta-analysis.
SentMask: efficacy of acupuncture in treating pain in acute pancreatitis: a systematic review and meta-analysis.

model is able to capture substantial semantics and minimize noise information from the source article by inserting an original sentence pointer.

5. Conclusions

In this paper, we propose SentMask, a novel extract-then-abstract method for text summarization. By utilizing the sentence-aware mask attention mechanism, our method avoids information loss caused by the extraction model. Besides, the paper utilizes a sentence-level extractor, which can preserve sentence-level semantics during generation. Experimental results, the semisupervised model and the supervised model, both demonstrate our model can generate comprehensive summaries without suffering information loss.

In terms of our future work, the paper attempt to extend our solution in various directions. One possible direction is to take into account the varied connections among the words and sentences in articles. The paper will explore using the similarity of phrases, especially critical phrases, to further explore semantic relationships.

Data Availability

The data used to support the findings of this study are included in the article [24, 25]. The MS2 and AESLC datasets can be derived from the websites https://paperswithcode.com/dataset/ms-2 and https://github.com/ryanzhumich/AESLC.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this article.

References

[1] Y. Gupta, P. S. Ammanamanchi, S. Bordia et al., "The effect of pretraining on extractive summarization for scientific documents," in *Proceedings of the Second Workshop on Scholarly Document Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2021.

[2] W. Li, X. Xiao, J. Liu, H. Wu, H. Wang, and J. Du, "Leveraging graph to improve abstractive multi-document summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020.

[3] L. Xiao, L. Wang, H. He, and Y. Jin, "Copy or rewrite: hybrid summarization with hierarchical reinforcement learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 9306–9313, 2020.

[4] J. A. Bishop, Q. Xie, and S. Ananiadou, "Gencomparesum: a hybrid unsupervised summarization method using salience," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, Dublin, Ireland, 2022.

[5] A. Kazemi, V. Perez-Rosas, and R. Mihalcea, *Biased Textrank: Unsupervised Graph-Based Content Extraction*, COLING, Gyeongju, Korea, 2020.

[6] R. Nallapati, B. Zhou, C. D. Santos, C. aglar Gulc ehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Berlin, Germany, 2016.

[7] N. Gu, E. Ash, and R. H. Memsum, "Extractive summarization of long documents using multi-step episodic Markov decision processes," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 6507–6522, Association for Computational Linguistics, Dublin, Ireland, 2022.

[8] M. Lewis, Y. Liu, N. Goyal et al., "Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020.

[9] M. A. Ali, A. Al-Dahoud, and B. Hawashin, "Enhanced feature-based automatic text summarization system-usingsupervised technique," *International Journal of Computers and Technology*, vol. 15, no. 5, pp. 6757–6767, 2016.

[10] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159–165, 1958.

[11] J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I.-H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing and Management*, vol. 41, no. 1, pp. 75–95, 2005.

[12] R. Jia, Y. Cao, H. Tang, F. Fang, C. Cao, and S. Wang, "Neural extractive summarization with hierarchical attentive heterogeneous graph network," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, 2020.

[13] A. Jadhav and V. Rajan, "Extractive summarization with swap-net: sentences and words from alternating pointer networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Melbourne, Australia, 2018.

[14] K. Arumae and F. Liu, "Reinforced extractive summarization with question-focused rewards," in *Proceedings of the ACL 2018, Student Research Workshop*, Association for Computational Linguistics, Melbourne, Australia, 2018.

[15] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized Autoregressive Pretraining for Language Understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Association for Computational Linguistics, Vancouver, Canada, 2019.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 9, no. 8, 2019, https://gwern.net/doc/ai/nn/transformer/gpt/2019-radford.pdf.

[17] C. Raffel, N. Shazeer, A. Roberts et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[18] S. Zhang, X. Zhang, H. Bao, and F. Wei, "Attention temperature matters in abstractive summarization distillation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 127–141, Dublin, Ireland, May, 2022.

[19] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, July, 2018.

[20] S. Bae, T. Kim, J. Kim, and S.-G. Lee, "Summary level training of sentence rewriting for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 10–20, Hong Kong, China, November, 2019.

[21] J. Xu and G. Durrett, "Neural extractive text summarization with syntactic compression," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3292–3303, Hong Kong, China, November, 2019.

[22] L. Lebanoff, K. Song, F. Dernoncourt et al., "Scoring sentence singletons and pairs for abstractive summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Italy, January, 2019.

[23] A. See, P. J. Liu, and C. D. Manning, "Get to the point: summarization with pointer-generator networks," 2017, https://arxiv.org/abs/1704.04368.

[24] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "Ms2: multi-document summarization of medical studies," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7494–7513, Punta Cana, Dominican Republic, November, 2021.

[25] R. Zhang and J. Tetreault, "This email could save your life: introducing the task of email subject line generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 446–456, Florence, Italy, July, 2019.

[26] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, November, 2019.

[27] J. Lin, X. Sun, S. Ma, and Q. Su, "Global encoding for abstractive summarization," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, no. 2, Melbourne, Australia, July, 2018.

[28] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] C.-Y. L. Rouge, "A package for automatic evaluation of summaries," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July, 2004.

[30] K. Papineni, S. Roukos, T. Ward, and Z. Wei-Jing, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association forComputational Linguistics (ACL)*, Philadelphia, July, 2002.