IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 17, 2021, accepted August 13, 2021, date of publication September 3, 2021, date of current version September 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110143

Enhancements of Attention-Based Bidirectional LSTM for Hybrid Automatic Text Summarization

JIAWEN JIANG[1], HAIYANG ZHANG[2], CHENXU DAI[1], QINGJUAN ZHAO[3], HAO FENG[1], ZHANLIN JI[1,4], (Member, IEEE), AND IVAN GANCHEV[5,6,4], (Senior Member, IEEE)

[1]Department of Computer Science, College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063009, China
[2]Department of Computer Science, The University of Sheffield, Sheffield S10 2TN, U.K.
[3]Department of Computer Science and Engineering, Beihang University, Beijing 100191, China
[4]Telecommunications Research Centre (TRC), University of Limerick, Limerick, V94 T9PX Ireland
[5]Department of Computer Systems, Plovdiv University "Paisii Hilendarski", 4000 Plovdiv, Bulgaria
[6]Institute of Mathematics and Informatics (IMI), Bulgarian Academy of Sciences (BAS), 1040 Sofia, Bulgaria

Corresponding authors: Zhanlin Ji (zhanlin.ji@ncst.edu.cn) and Ivan Ganchev (ivan.ganchev@ul.ie)

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2017YFE0135700 and in part by Grant No. BG05M2OP001-1.001-0003 financed by the Science and Education for Smart Growth Operational Program of Bulgaria (2014–2020) and co-financed by the European Union through the European Structural and Investment Funds.

ABSTRACT The automatic generation of a text summary is a task of generating a short summary for a relatively long text document by capturing its key information. In the past, supervised statistical machine learning was widely used for the Automatic Text Summarization (ATS) task, but due to its high dependence on the quality of text features, the generated summaries lack accuracy and coherence, while the computational power involved, and performance achieved, could not easily meet the current needs. This paper proposes four novel ATS models with a Sequence-to-Sequence (Seq2Seq) structure, utilizing an attention-based bidirectional Long Short-Term Memory (LSTM), with added enhancements for increasing the correlation between the generated text summary and the source text, and solving the problem of out-of-vocabulary (OOV) words, suppressing the repeated words, and preventing the spread of cumulative errors in generated text summaries. Experiments conducted on two public datasets confirmed that the proposed ATS models achieve indeed better performance than the baselines and some of the state-of-the-art models considered.

INDEX TERMS Natural language processing (NLP), automatic text summarization (ATS), sequence-to-sequence (Seq2Seq) model, attention mechanism, bidirectional LSTM (Bi-LSTM), pointer network, coverage mechanism, mixed learning objective (MLO) function.

I. INTRODUCTION

With the advance of the mobile Internet in recent years, text information is showing a trend of explosive growth. With people continuously exposed to massive text information all time, it has become an imperatively important for them to be able to extract valuable content from text information quickly and accurately in order to get good reading experience and lower the time needed for digesting the obtained information. The Automatic Text Summarization (ATS) [1], aiming at automatically outputting concise and fluent text summaries by utilizing different algorithms, has provided an efficient and feasible solution. ATS aims to find the key points after understanding the text contents in order to generate smooth and flexible short texts by means of cutting and splicing, synonymous reporting, abbreviation, extension, etc.

ATS can be divided into three main types: (i) *extractive ATS* which can extract and combine important words from a source text in order to generate a text summary [2]; (ii) *abstractive ATS* which utilizes neural network algorithms as to generate a concise and aggregated text summary by means of synonymous substitution, person conversion, imitation, expansion, etc. [3]; and (iii) *hybrid ATS*, employing elements of both extractive and abstractive ATS. Compared with the extractive ATS, the abstractive ATS can produce text summaries closer to that produced by the people's thinking process. By utilizing the advantages of both extractive and abstractive ATS, however, the hybrid ATS has the potential to achieve even better results. Generally, it is a challenging task for a computer to understand the text contents and generate

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna D'Ulizia.

text summarization automatically [4]. It should not only have the ability to understand the content information of the whole text, but also to identify the importance of text words, make choices according to their importance, and copy, rearrange, and combine the choices as to meet the requirements of the automatic text generation.

With the development of deep learning, especially with the emergence of Recurrent Neural Networks (RNNs), the elaborated abstractive ATS models mainly rely on a large amount of data rather than using complex model structures to achieve better and rapid natural language processing (NLP) in multiple fields, such as machine translation, speech recognition, sequence generation, etc. However, due to the problem of "long-distance dependence" [5], a RNN model could lose some information after a certain number of steps, resulting in an inaccurate text summarization. By combining the attention mechanism with a RNN [6], more emphasis could be put on the important information, by selectively ignoring the other contents, so that the generated text summaries would be more targeted and accurate.

With a RNN, each word is entered into the model according to its order of appearance in the original text, which is good for recording the sequence information of the text but limits the speed of network training and text summary generation. If both the encoder and decoder adopt a Convolutional Neural Network (CNN) [7], [8], this will allow the model to use parallel computing in the phase of training, which will not only maintain accuracy but will also improve efficiency. At the same time, if multi-step attention is used [9], every layer of the semantic vector, generated in the decoder, would look forward, which could further improve accuracy.

Supervised statistical machine learning is widely used in the task of text summary generation, but due to its high dependence on the quality of text features, the generated summaries often lack accuracy and coherence. In addition, its computational cost and performance cannot easily meet the needs of massive texts in the 'big data' era. Therefore, by taking the attention-based bidirectional Long Short-Term Memory (LSTM) model as a foundation [10], [11], we aimed at improving it by employing an enhanced semantic network (ESN), a pointer network (PN), a coverage mechanism, and a mixed learning objective (MLO) function [12] as to improve the automatic generation of text summaries and reduce the word recurrence, e.g., generated by the traditional statistical methods.

In general, the models, proposed in this paper (except ESN), belong to the hybrid ATS, because in addition to using the core principles of the abstractive ATS, they also employ some elements of the extractive ATS, but only for the processing of words that are *unregistered* in the basic vocabulary. In each such case, a choice is made about copying the *unregistered* word from the source text or pointing to another similar word in the basic vocabulary, based on a probability produced by a *softmax* normalization. In other words, the proposed hybrid ATS models select the output terms by using a probability, based on both parts – the vocabulary and the source text.

The rest of the paper is organized as follows. The next section presents the related work done in this field. Section III describes the four novel ATS models proposed. Section IV presents the performance evaluation of the elaborated ATS models in comparison with the baselines and state-of-the-art models, and discusses the results obtained by the conducted experiments. Finally, Section V draws the conclusions and sets a future direction for research.

II. RELATED WORK

In 2014, the Sequence-to-Sequence (Seq2Seq) model, proposed by the Google's Brain team [13], officially started the research on end-to-end network in the NLP field. In this model, the encoder has a bidirectional LSTM structure (Bi-LSTM) [14], whereas the decoder has a unidirectional LSTM structure. However, the "long-distance dependence" problem resulted in insufficient accuracy of the generated text summarization.

In a paper published by Bahdanau *et al.* [15], an attention mechanism was applied to the NLP task for improving the accuracy of generated text summaries. With the attention mechanism, semantic coding S_i at time i is not a direct coding of the input sequence, but a weighted sum considering the importance of each element x_j in the sequence (of length T_x) as per the following formula:

$$S_i = \sum\nolimits_{j=1}^{T_x} b_{ij} f(x_j), \tag{1}$$

where $f(x_j)$ denotes the coding of element x_j, and b_{ij} is treated as a probability, reflecting the importance of element x_j to S_i, represented by the following *softmax* function:

$$b_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{jk})} \quad (\text{for } j = 1, \ldots, T_x), \tag{2}$$

where e_{jk} is the degree of matching between element x_j to be encoded and other elements x_k in the sequence. Higher the degree of matching, greater the influence of the element and greater the value of b_{ij}.

With the Seq2Seq model, it is possible to generate unknown words and incomplete content. To solve these problems, Seq2Seq integrates keyword information. Through the use of an attention mechanism, the main information and keywords of the text source are considered at the same time to improve the amount of information by its generated summary [16]. Another way is to use the hidden state of the source and the participating summary vector to calculate the attention weight of the summary perception [17].

For discovering the relationship between the original sentences, Yang *et al.* proposed a hierarchical abstraction mechanism, whereby a self-attention is used first to discover the relationship between sentences, followed by a replication mechanism to solve the out of vocabulary (OOV) problem [18].

In NLP, learning representation is a pioneering research, especially when applied to Seq2Seq, in which the generated output completely depends on learning to represent the source sequence. In order to get rich semantic representation, the encoder is equipped with an asynchronous bidirectional parallel structure. The decoder is different from the classical attention-based works, and uses self-awareness context selection mechanism as to generate text summaries in a more efficient way [19].

In 2015, Hu *et al.* applied the RNN and RNN context model on 2.4 million short-text data contained in the LCSTS corpus [20], which has provided the basis for generating Chinese text summarization. For dealing with noise and unimportant information, Ma *et al.* proposed the SRB model [21], whereby the source text is represented by a gated attention encoder, while the digest representation is generated by the decoder. Moreover, the similarity between representations is maximized during the model training process. For the problem of unregistered words, Gu *et al.* proposed CopyNet [22], a model of Seq2Seq type, which skilfully aided the decoder with a copy mechanism. In CopyNet, the encoder is a bidirectional RNN [23] and decoder is divided into two parts using: (i) a generate mode, determining the output based on semantics; and (ii) a copy mode, determining the copy according to the position of the input text. During the decoding process, when the model updates the state at any particular moment, it not only uses the prediction information at the previous moment, but also uses the coding information of the prediction output at the encoding stage.

In [24], a CNN was used in the encoder to realize word serialization, which has not only retained the accuracy, but also sped up the computation. In this model, the encoder employs CNN and Bi-LSTM, and the decoder has a Bi-LSTM structure with a *softmax* function. The resultant 'CNN-Bi-LSTM' model is depicted in Figure 1.

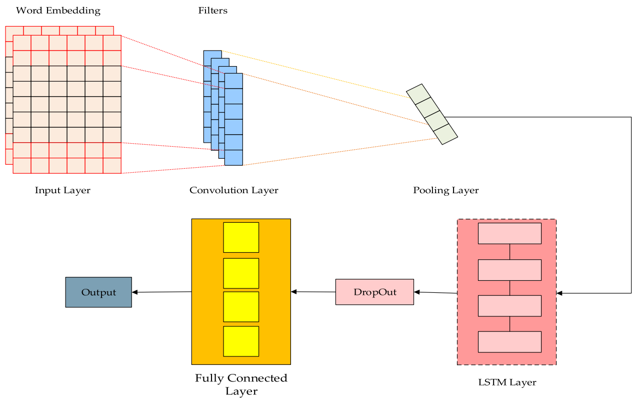


FIGURE 1. The 'CNN-Bi-LSTM' model.

Due to different lengths of sentences in NLP, the size of the input matrix of the CNN is uncertain, as it depends on the number of characters at the input. The convolution layer is essentially a feature extraction layer. A hyperparameter F is set to specify how many feature extractors (filters) are used. For a filter, one can imagine a $k * d$ moving window from the input matrix, where k is the window size specified by the filter and d is the length of the embedding word. At a certain moment, through a nonlinear transformation of the neural network, the input value in this window is converted into a certain characteristic value. As the window continues to move backward, the characteristic value corresponding to this filter is generated continuously to form the filter feature vector. This is, in short, the process of extracting features from the convolution layer. Each filter operates this way to form a different feature extractor. The pooling layer performs dimensionality reduction operations on the features of the filter to form the final features. Generally, a fully connected layer is used after the pooling layer to complete the classification process.

In June 2017, the Google team proposed that a complete end-to-end processing could be realized without using a CNN or a RNN, but only using the encoder/decoder attention in a self-attention manner [25]. In this model, both the encoder and decoder employ a multi-headed self-attention mechanism. The resultant model, called here 'Seq2Seq + Attention', is shown on Figure 2. The encoder/decoder attention could be used to establish the corresponding relationship between the original contents and the target phrases and sentences, whereas self-attention pays more attention to the structure of the word pairs in a sentence. The encoder maps the original text vector sequence $x = [x_1, \cdots, x_{T_x}]$ to a context vector. The decoder uses the hidden state h_i as the input of the query at each time step of decoding, and queries the hidden state of the encoder [17]. Each time step calculates the weight of the corresponding query and performs a weighted average to obtain the context vector u, which is used to represent the original text information most relevant to the output of the current time step. When entering the next time step, the corresponding decoded words are inputted into the cyclic neural network of the decoder together with the context vector, so that the current most concerned information can be extracted, instead of relying entirely on the hidden state of the previous time step.

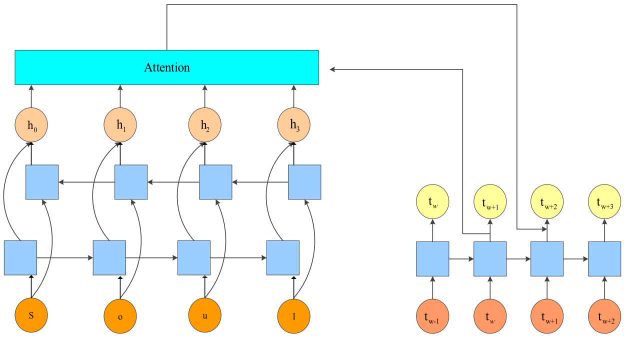


FIGURE 2. The 'Seq2Seq + Attention' model.

In theory, the 'Seq2Seq + Attention' model can be generalized for processing longer sentences, but currently it is still limited in the field of translation.

A pointer network is a variant of the Seq2Seq model. The network does not perform sequence conversion but generates a series of pointers to the elements of the input sequence. In ATS, pointer networks are used mainly to solve the OOV problem and the word sparseness issue [26].

Another successfully used model is the attention-based bidirectional LSTM model, presented in [27], called here 'Bi-LSTM + Attention'. In this model, the encoder employs an attention mechanism with Bi-LSTM, and the decoder has a Bi-LSTM structure. The model consists of five layers (Figure 3): an input layer used to input data into the model, an embedding layer used to map each word in a low-latitude space, a Bi-LSTM layer that uses a bidirectional LSTM for advanced feature extraction and weight vector generation, an attention layer that is able to generate sentence-level features by combining a weight vector with lexical features, and an output layer.

Basically, the 'Bi-LSTM + Attention' model adds an attention mechanism on top of the Bi-LSTM layer, for the purposes of applying an attention weighting. The added attention mechanism can differentiate the weight of each word and thus enable the whole sequence to get the key information more easily.

The 'Bi-LSTM + Attention' and 'Seq2Seq + Attention' models are used as a basis in this paper for elaborating four new ATS models, described in the next section, utilizing an enhanced semantic network (ESN), a pointer network (PN), a coverage mechanism, and a mixed learning objective (MLO) function [12], respectively, as to improve the automatic generation of text summaries.

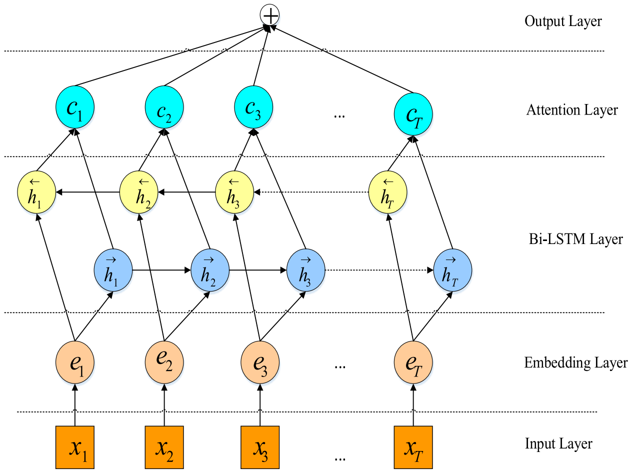


FIGURE 3. **The 'Bi-LSTM + Attention' model.**

The main contributions of the paper could be summarized as follows:

1) A novel ESN model is proposed, using a hidden-layer state at both the encoder and decoder to work out the semantic similarity and a loss function for paying more attention to the source text during the model training, for improving the correlation between the generated text summary and the source text.

2) A novel 'DA-PN' model is proposed, based on a joint application of the attention distribution at both ends (encoder and decoder), along with integrating the attention distribution in the current time step and introducing a pointer network at the decoder, for solving the problem of unregistered words.

3) A novel 'DA-PN + Cover' model is proposed, based on a coverage mechanism, integrating a coverage vector, calculated from the attention distribution at both the encoder and decoder in the previous time step, along with adding a loss function, for avoiding the repetition of words in the generated text summaries.

4) A novel 'DA-PN + Cover + MLO' model is proposed, by setting mixed learning objectives, along with introducing a self-critical gradient algorithm, setting a global reward, and taking the evaluation index as part of the model iteration, for preventing the spread of cumulative errors in the generated text summaries.

III. PROPOSED MODELS

The automatic text summarization aims at generating a concise and clear text summary automatically after the original (source) text content is inputted into the utilized algorithm [26]. The input text is a sequence containing $\mathcal{M}$ words and $x = [x_1, x_2, \cdots, x_i, \cdots, x_{\mathcal{M}}]$ columns, while the output summarization is a sequence containing $\mathcal{N}$ words and $y = [y_1, y_2, \cdots, y_j, \cdots, y_{\mathcal{N}}]$ columns. The length $L_{\mathcal{N}}$ of the text summary is shorter than the length $L_{\mathcal{M}}$ of the source text, i.e., $L_{\mathcal{N}} < L_{\mathcal{M}}$. The words x_i and y_j all come from a fixed vocabulary V of size $|V|$. Each word of the source text could be described as a one-hot vector $x_i \in \{0, 1\}^V$ for $i \in \{1, \cdots, M\}$, that is, for the i^{th} word, a vector of length V is used to represent it, whereby the i^{th} element is 1 and all other elements are 0.

In order to strengthen the relationship between the source text and the generated text summary, in this section we propose first to use an enhanced semantic network (ESN). Secondly, to handle out-of-vocabulary (OOV) words, we propose to use a pointer network (PN) integrating a decoder attention, which can connect the context vector of the decoder with the context vector generated from the source text, for jointly regulating the selection of the source text and the additional lexicon. Thirdly, for the repeated insertions, we propose to add a multi-attention coverage mechanism with simultaneously using the coverage vectors of both the encoder and decoder to affect the attention degree. Moreover, the attention information of the decoder can be inputted to the mapping layer of the input, so that the model could pay attention to the past information and thus reduce the recurrence phenomenon. Finally, aiming at the readability of generated text summaries, we propose to utilize a mixed learning objective (MLO) function [12] for setting a global reward, so that the specific discrete gradient could be maximized during scoring, thus enhancing the readability of the generated text summaries.

A. ESN MODEL

Based on the hypothesis of the existence of strong correlation between the generated text summary and the source text, a novel ESN model is proposed here, which works out the semantic similarity between the encoder and decoder, and maximizes the semantic relevance during the model training for performance enhancement (Figure 4).

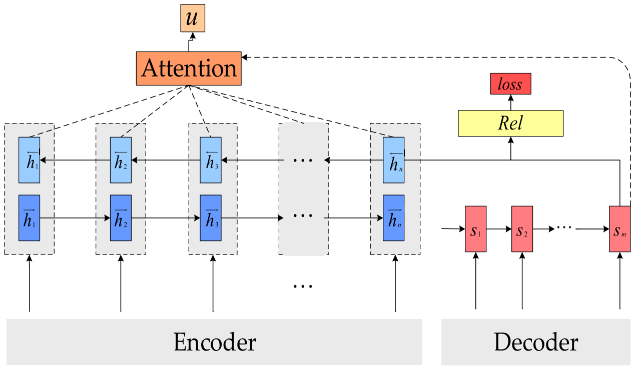


FIGURE 4. The proposed ESN model.

In the ESN model, the last hidden-layer state at the encoder is considered as a semantic vector G_e of the source text. Moreover, if the encoder adopts the bidirectional LSTM, one can get $G_e = \left(\vec{h}_N, \overleftarrow{h}_N\right)$, namely, the features of the source sequence can be obtained in both the forward and backward direction. At the same time, the final hidden-layer state, obtained from the unidirectional LSTM at the decoder, is treated as a semantic vector G_d of the generated text summary. The semantic similarity Rel_{ESN} can be expressed by calculating the cosine similarity [28] of G_e and G_d. As the source text and generated text summary are located in the same semantic space, the cosine similarity can effectively measure the distance between them. Then, the semantic similarity could be calculated by taking into account the feature vectors of the decoder and encoder in both directions along with the calculated cosine similarity, and averaging the value to prevent excessive noise:

$$Rel_{ESN} = \frac{1}{2}\left(\frac{\vec{h}_N * G_d}{\left\|\vec{h}_N\right\| \|G_d\|} + \frac{\overleftarrow{h}_N * G_d}{\left\|\overleftarrow{h}_N\right\| \|G_d\|}\right). \qquad (3)$$

Then, a new loss function can be formed from the original loss function $loss_t = -\log P(w_t^*)$ and the calculated semantic similarity, as per [29], as:

$$loss = \frac{1}{N}\sum_{t=1}^{N} loss_t - \lambda Rel_{ESN}, \qquad (4)$$

where $P(w_t^*)$ denotes the probability distribution of the final prediction vocabulary, λ denotes the regulating factor which can be valued according to the experience, because the encoder carries information of both the forward and backward direction and is able to represent the feature of the source text more comprehensively. This way, by introducing semantic similarity into the loss function, the probability of generating more accurate text summary could be increased during the model training, thus improving the correlation between the generated text summary and the source text.

B. 'DA-PN' MODEL

Aiming at solving the problem of unregistered words, a novel 'DA-PN' model (Figure 5) is proposed in this subsection, utilizing decoder attention (DA) based on a pointer network (PN) [29]. At each time step, this model decides whether to copy a (unregistered) word from the source text or point to another similar word existing in the basic vocabulary. The former action is controlled by a probability normalized with *softmax*, whereas the latter relates to predicting the words from the input data by the attention distribution of the decoder. If both are combined, unregistered words can also be predicted if they appear in the source text but are not included in the basic vocabulary.

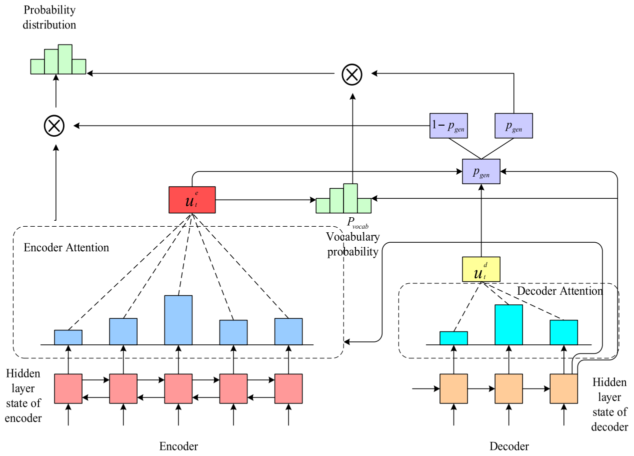


FIGURE 5. The proposed 'DA-PN' model.

As shown in Figure 5, the 'DA-PN' model has two input layers, respectively for copying a word from the basic vocabulary or from the source text. Each dimension represents one word. A switching mechanism *p_gen* [30] is used for generating a probability for controlling the input source. The decoder attention can compensate the information weakening caused by long sequences, thus enabling the model to locate the key information more accurately. The attention distribution, corresponding to the probability distribution of each word in the source text and thus revealing which words are more important in the prediction process, could be calculated as in [29]:

$$\alpha^t = softmax\left(V_\alpha tan\, h\left(W_\alpha s_t^T\right)\right), \qquad (5)$$

where W_α and V_α denote the weight parameters, and s_t denotes the hidden-layer state of the decoder at time step t. The weighted sum of the attention distributions obtained from (5) and the decoder state before the current time step constitutes the final decoder context vector, as per [31]:

$$u_t^d = \sum\nolimits_{j=1}^{t-1} \alpha_j^t s_j. \tag{6}$$

The probability of the switch p_{gen} can be worked out based on the encoder/decoder context vector and the current hidden-layer state, as in [29]:

$$p_{gen} = \sigma(W_y y_{j-1} + W_e u_t^e + W_d u_t^d + bt_{ptr}), \tag{7}$$

where W_y denotes the weight matrix of the previous word y_{j-1}, W_e denotes the weight matrix of the encoder hidden state at the current time step, u_t^e denotes the final encoder context vector, u_t^d denotes the final decoder context vector, W_d denotes the weight matrix of the decoder hidden state at the current time step, and bt_{ptr} denotes the bias term.

The weight matrix and bias term in (7) are parameters that can be updated during the model training iteration. The linear weighted sum of these parameters is nonlinearly activated by the *sigmoid* function and mapped between 0 and 1 as a soft switch to control the source of the input layer, based on the information of the two parts – the source text and the vocabulary. The vocabulary distribution of the current time step can be expressed, as per [29], as follows:

$$P_{vocab} = softmax(V' (V [s_t, w] + bt) + bt'), \tag{8}$$

where s_t is the current state sequence, V is the weight matrix, bt is the bias term for the training iteration, and w is the predicted word.

Then the probability of the final prediction for word w is:

$$P_{final}(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum\nolimits_{i:w_i=w} a_i, \tag{9}$$

where p_{gen} is the probability for controlling the input source, $P_{vocab}(w)$ is the distribution of words in the final output vocabulary, and $\sum_{i:w_i=w} a_i$ represents the encoder context vector. Basically, (9) refers to the sum of two products: (i) the product of the probability of generating a new word and the distribution of words in the vocabulary; and (ii) the product of the probability of the occurrence of a word in the source text and the probability of copying it from there. This way, the vocabulary can be expanded to form a union of all the words in the original corpus and in the initial vocabulary. Then (9) can be used to express the probability distribution of the expanded vocabulary. This way the limitation of using a preset basic vocabulary, as in the Seq2Seq model, can be easily overcome.

With the proposed 'DA-PN' model, it is also much easier to find suitable words for copying from the source text. For this, a relevant weight just needs to be assigned to each word. Moreover, words can be selected from the expanded vocabulary, which is created on demand by copying words, unregistered in the basic vocabulary, from the source text. For instance, as low-frequency words, names of people or places may not be present in the basic vocabulary but could be extracted from the source text for inclusion in the text summary.

In addition, with 'DA-PN': (i) a smaller vocabulary could be used, which results in saving a computation power and storage; (ii) the model training is faster; and (iii) fewer epochs[1] are needed to achieve identical performance to that of 'Seq2Seq + Attention'. Consequently, simultaneous attention can be paid to both the encoder and decoder, and the generated text summary more accurately matches the corresponding source text.

C. 'DA-PN + COVER' MODEL

This subsection proposes to combine the 'DA-PN' model, presented in the previous subsection, with a coverage mechanism integrating multi-attention. The resultant model, named 'DA-PN + Cover', is depicted on Figure 6. By using the attention distribution of the encoder and decoder in all the previous time steps, two coverage vectors are worked out: (i) c_t^d, representing the attention to the target sequence, i.e., the attention to the words generated in the process of decoding; and (ii) c_t^e, representing the attention to the words in the source text. The degree of coverage for a certain word is the sum of all attentions it has obtained at a particular moment. At the initial stage, both coverage vectors are set to 0, because there is no coverage occurring at that time.

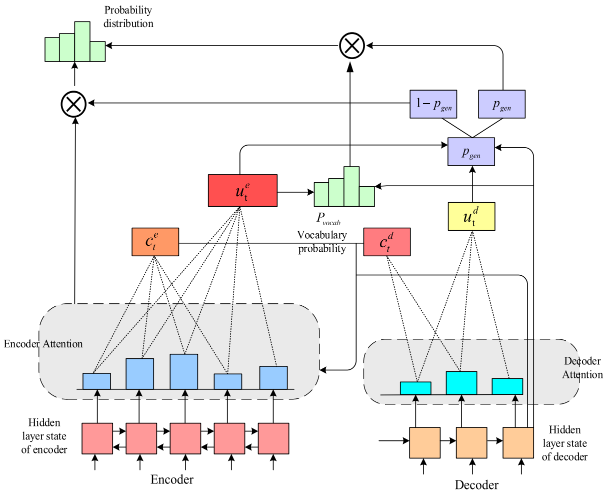


FIGURE 6. The proposed 'DA-PN + Cover' model.

Adding an attention mechanism to the decoder can effectively lead to focusing on the features of the generated text summary. However, for the problem of word recurrence, one should pay attention to the whole sequence, so it should be controlled with a global vector. In addition, the hidden layer of the decoder uses the context vector of the past time for calculation of the current input, so that the information obtained in the past can be used during the calculation, thus strengthening the past and current time relationship.

In the current time step, in order to use the source text information in the past and the target text summary's words

[1] *Epoch* - a hyperparameter, defining the number of times a model has worked through the entire training set.

generated, the two vectors calculated in the previous step should be integrated in the attention mechanism as per the following formula:

$$e_j^t = v^T \tanh\left(W_h h_j + W_s s_t + W_e c_t^e + W_d c_t^d\right) + bt_{atten}, \tag{10}$$

where v^T is the weight matrix of the iterative nonlinear activation function used in training, W_h is the weight matrix of the iterative hidden state in training, h_j is the hidden state, W_s is the weight matrix of the iterative current state sequence in training, s_t is the current state sequence, W_e represents the weight matrix of the encoder hidden state at the current time step, W_d represents the weight matrix of the decoder hidden state at the current time step, c_t^e is the coverage vector of the encoder, c_t^d is the coverage vector of the decoder, and bt_{atten} is the bias term of the training iteration.

This way, the output of the current time step is affected by the previous source text and the generated text summary's words, so as to avoid paying attention to the same information and thus avoiding the recurrence. Meanwhile, the context vectors of the two kinds of attention can be integrated into the probability distribution of the vocabulary, which makes it easier for the model to locate the important words when calculating the probability distribution of words. Finally, one needs to introduce an additional loss term to punish the coverage vector c_{ij} and the new attention allocation a_{ij}, and then get the coverage loss function formula. In the task of text summary generation, one does not require all original texts to be covered, but only needs to punish the recurrence between the attention distribution and coverage distribution (i.e., select the smaller value of the two vectors), so $cover\_\_loss_t$ is bounded. The final loss function consists of the original loss and coverage loss as follows:

$$loss_t = -\log\left(p\left(w_t^*\right)\right) + \beta cover\_\_loss_t, \tag{11}$$

where $p(w_t^*)$ denotes the probability distribution of the final predicted vocabulary, $cover\_\_loss_t$ denotes the coverage loss function, and β denotes the weight of the coverage mechanism in the overall loss.

In (11), the weight of the coverage mechanism in the overall loss can be regulated by controlling the value of β. Moreover, by introducing the distribution of the attention at both ends (encoder and decoder), as well as using additional loss items, the model can effectively suppress the possible repeated fragments and improve the automatic generation of text summaries.

D. 'DA-PN + COVER + MLO' MODEL

In order to prevent the spread of cumulative errors in generated text summaries, this subsection proposes to add a mixed learning objective (MLO) [12] to the 'DA-PN + Cover' model. The resultant model is named 'DA-PN + Cover + MLO'. The evaluation indicator ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [32] is used as a part of the model iteration, along with a global reward. However, as ROUGE itself is not differentiable, it cannot be used directly for gradient calculation. In order to eliminate these limitations, the Self-Critical Policy Gradient Training Algorithm could be adopted to generate two independent output sequences in the training iteration. A greedy search can be performed at each time step from the probability distribution y^s of $p\left(y_i^s|y_1^s,\cdots y_{i-1}^s,x\right)$ at each decoder's time step and the baseline output $y^\wedge$ can be obtained by maximizing the output probability distribution. By defining $r(y)$ as the reward function of the output sequence y and comparing it with the real sequence y^*, one can get the following formula:

$$L_{rl} = \left(r\left(y^\wedge\right) - r\left(y^s\right)\right)\sum\nolimits_{i=1}^{n} log\left(p\left(y_i^s|y_1^s,\cdots,y_{i-1}^s,x\right)\right) \tag{12}$$

where x denotes the input vector, $y^\wedge$ denotes the baseline output obtained by maximizing the output probability distribution, and $r(y)$ denotes the reward function of the output sequence y. The minimized L_{rl} is equivalent to the conditional likelihood of the maximized sampling sequence y^s. If it gets a higher return than the baseline y^*, one can expect more return from the model. After improvement, the global reward is used, as depicted in Figure 7.

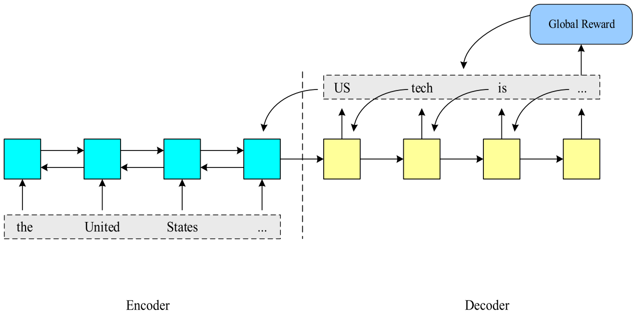


FIGURE 7. The training process of the proposed 'DA-PN + Cover + MLO' model.

IV. EXPERIMENTS

A. DATASETS

Two datasets were used in the experiments conducted with the proposed ATS models: (i) the LCSTS short-text corpus [20], created from the Chinese micro-blogging website Sina Weibo; and (ii) the TTnews long-text corpus [33], created for the NLPCC 2017&2018 shared summarization tasks.

The LCSTS corpus consists of over 2.4 million 'text–summary' pairs of Chinese short texts along with corresponding short summaries provided by their authors. It is divided into three parts: (i) 2,400,591 'text–summary' pairs, used as a training set; (ii) 10,666 human-labeled 'text–summary' pairs with scores, ranging from 1 to 5, indicating the relevance between the short text and the corresponding summary; this part was used as a validation set; and (iii) 1,106 'text–summary' pairs, with scores, ranging from 1 to 5, used as a test set in the experiments. In the validation set and test set, we only use pairs with scores no less than 3.

The TTNews corpus is the largest corpus for single document summarization in Chinese [33]. It contains: (i) 50,000 news articles, browsed on Toutiao app, and their corresponding summaries, written by experts, and 50,000 news articles without summary – in the training set; and (ii) 2,000 news articles – in the test set. In addition, we randomly extracted 2000 news articles from the training set and used these as a validation set for the purposes of providing an unbiased evaluation of the models' fit to the training set during the hyperparameters' tuning by early stopping of training when the error on the validation set increases, which is a sign of overfitting. The TTNews articles came from different fields, such as sport, food, entertainment, politics, technology, finance, etc.

B. EVALUATION METRICS

We adopted the ROUGE [32] – a set of metrics and a software package used for evaluating automatic summarization in NLP, based on recall and accuracy. The most popular ROUGE evaluation metrics include: (i) ROUGE-N, evaluating the overlap of N-grams (e.g., ROUGE-1 for unigrams, ROUGE-2 for bigrams, etc.), (ii) ROUGE-L (the longest common subsequence-based statistics), and (iii) ROUGE-S (a skip-bigram based co-occurrence statistic). The advantages of ROUGE-N come from its intuitiveness, conciseness, and reflection of the word order. But when $N > 3$, the ROUGE-N value is usually very small. The advantage of ROUGE-L is that it does not require continuous matching of words, and only requires matching the order of their appearance. However, it only calculates the longest subsequence, and the final value ignores the influence of other candidate subsequences, both longer and shorter. The advantage of ROUGE-S is that it considers all word pairs arranged in order, which reflects sentence-level word order more deeply than the N-gram model. However, if it does not set the maximum number of jumping words, many meaningless word pairs will appear. In the conducted experiments, for performance evaluation and comparison of different models, we used the calculated ROUGE-1, ROUGE-2, and ROUGE-L scores. The greater the score, the better the performance of the corresponding model.

However, as ROUGE is used for evaluating the automatic summarizations of English texts mainly, we adopted a special method for applying it to Chinese texts. According to this method, the Chinese text summarizations are first converted into a dictionary and then, by following the corresponding ID in the dictionary, summarizations are translated into sequences composed of IDs.

C. COMPARED MODELS

Having incorporated enhancements to the structure of 'Bi-LSTM + Attention' and 'Seq2Seq + Attention' models, the four proposed ATS models were primarily compared to these two baselines. In addition, a performance comparison was made also with the following state-of-the-art models:

1) 'RNN (word)' and 'RNN-Context (word)' [20] – two Seq2Seq models using gated recurrent units (GRUs) for the encoder and decoder. The main difference between these two models is that there is no attention mechanism in 'RNN (word)'.
2) 'SRB + Attention' [21], a model with a gated attention encoder, which achieves high semantic similarity between the source text and generated summary by optimizing cosine similarity loss.
3) KESG [16], which adds a keyword network based on Seq2Seq + adversarial learning.
4) Dual [17], which employs dual methods to create a summary-aware attention weight, considering both the source text and generated summary.
5) Filtering Window (FW) model [19], which directly uses the maximum value of the projection vector as the alignment position to simplify the calculation.
6) HAM [18], based on a hierarchical attention network and a replication mechanism added to it.

D. EXPERIMENTAL SETTINGS FOR PROPOSED MODELS

Global parameters, remaining unchanged during the experiments, are mainly composed of network-layer parameters and feature-representation parameters. In the former, the LSTM hidden layer adopts 256 units and the word vector uses 256 dimensions. The pointer network at the decoder can handle the characteristics of unregistered words, so the first 50,000 words were selected as a basic vocabulary source according to the word frequency after comparison. The consistency of the vocabulary was ensured at both ends (encoder and decoder). For the LSTM network layer, the weight and matrix distribution of the computing unit were generated by a uniformly distributed random initializer. The variables were iterated starting from 0, the standard deviation was generated from a 0.0001 Gaussian distribution random initializer, and the bias term was also iterated from 0. Moreover, at the LSTM dropout, the probability was set to 0.4, the Adam algorithm was adopted for gradient optimization [34], [35], the *batch_ size* (determining the number of initial hidden-state vectors in LSTM) and *beam_ size* (the number of possible results in the current state) were set to 64 and 4, respectively, and the extra loss was set to 0.8.

E. RESULTS AND DISCUSSIONS

The first group of experiments, conducted with the proposed ATS models, was based on the short-text corpus LCSTS only, whereas the second group of experiments involved also the long-text corpus TTnews. The obtained results are shown in Tables 1 and 2, respectively (**bolded** for the proposed ATS models). The results for the baselines and state-of-the-art models are taken from the corresponding literature sources. As there were no results, involving the TTnews corpus, reported in the literature for the state-of-the-art models considered, these were omitted from Table 2.

TABLE 1. Performance comparison of models (on LCSTS).

Model	ROUGE-1	ROUGE-2	ROUGE-L
'RNN (word)' [20]	17.70	8.50	15.80
'Bi-LSTM + Attention' (*baseline*)	25.47	8.09	29.03
'RNN-Context (word)' [20]	26.80	16.10	24.10
'Seq2Seq + Attention' (*baseline*)	27.14	9.01	30.89
ESN	**29.91**	**11.54**	**30.91**
'DA-PN'	**32.16**	**13.49**	**31.28**
'SRB + Attention' [21]	33.30	20.00	30.10
'DA-PN + Cover'	**33.40**	**13.87**	**32.88**
'DA-PN + Cover + MLO'	**33.78**	**14.02**	**32.76**
FW [19]	35.00	16.91	32.65
HAM [18]	35.40	11.90	33.30
Dual [17]	38.20	25.00	35.20
KESG [16]	39.40	28.40	35.30

TABLE 2. Performance comparison of models (on LCSTS and TTnews).

Model	ROUGE-1	ROUGE-2	ROUGE-L
'Bi-LSTM + Attention' (*baseline*)	25.27	8.03	29.04
'Seq2Seq + Attention' (*baseline*)	27.07	9.03	30.56
ESN	**29.17**	**10.97**	**30.95**
'DA-PN'	**32.20**	**13.46**	**31.26**
'DA-PN + Cover'	**33.43**	**13.77**	**32.93**
'DA-PN + Cover + MLO'	**33.75**	**14.09**	**32.69**

Table 3 provides two running cases for the proposed ATS models, showing that the best performing one among them is ‘DA-PN + Cover + MLO’.

From the results, shown in Tables 1 and 2, the following observations could be made:

1) All four proposed ATS models outperform the baselines ‘Bi-LSTM + Attention’ and ‘Seq2Seq + Attention’. In addition, the baselines need more training time. As can be seen from Figure 8, in order to reach a convergence loss of 2.6, ‘Bi-LSTM + Attention’ and ‘Seq2Seq + Attention’ need to be trained for 40 and 17 epochs, respectively, whereas all proposed ATS models need less training, i.e., ESN needs 9 epochs, ‘DA-PN’ – 5 epochs, and ‘DA-PN + Cover’ and ‘DA-PN + Cover + MLO’ – 8 epochs.
2) From the two baselines, ‘Seq2Seq + Attention’ shows better performance, proving by this that, in general, it is more effective to use at the decoder a complex network than a statistical method.
3) Among the proposed ATS models, firstly, ESN is obviously superior to ‘Seq2Seq + Attention’ because it emphasizes on the strong correlation between the headword and source text, so the coherence can be correlated according to the semantic phase speed, which makes it easier to get the headword and pay attention to the low-frequency words to a certain extent. Furthermore, after adding the semantic, the model bias could be restrained, and its performance improved due to increasing the correlation between the source text and generated text summary.
4) The results for the second proposed model, ‘DA-PN’, confirmed that the use of a decoder’s pointer network can indeed improve the model performance effectively, because even if a smaller basic vocabulary is initially used, this can be expanded to help dealing with the unregistered words and, moreover, the encoder can effectively utilize the attention information of the decoder.
5) With the multi-attention coverage mechanism added, the third proposed model, ‘DA-PN + Cover’, can perfectly solve the problem of word recurrence, so that even better performance than that of ‘DA-PN’ could be achieved.
6) By utilizing a mixed learning objective (MLO) function [12], the last proposed model, ‘DA-PN + Cover + MLO’, showed that the ATS performance could be improved even further.
7) Even though the proposed ATS models do not perform so well as some of the state-of-the-art models (i.e., FW [19], HAM [18], Dual 17], KESG [16]), there is a room for improving their performance, especially for ‘DA-PN + Cover + MLO’, e.g., by further

TABLE 3. Sample summaries generated by the proposed models (Richer information, generated by a model, is shown in bold).

Document: 随着微信等的崛起，电信传统业务正面临巨大冲击。截至D年末，全国短信发送量同比仅增D%，远低于往年。据知名电信咨询公司Ovum估算，由于免费通讯软件普及，D年全球电信运营商短信营收损失将达D亿美元，D年或增至D亿美元。 **ESN:** 微信软件冲击电信业务 **'DA-PN':** 微信**崛起**，**免费的通讯**软件冲击电信业务。 **'DA-PN + Cover':** 微信**等软件**崛起，免费通讯软件的**普及给**电信业务**带来巨大**冲击。 **'DA-PN + Cover + MLO':** 微信崛起，免费通讯软件**使传统**电信业务**损失**巨大。
Document (Concise): 在赢得尼日利亚-里亚那总统宝座一天后，穆罕默杜·布哈里告诉CNN(Cable News Network)的克里斯蒂安·阿曼普尔，他计划积极打击长期困扰尼日利亚的腐败，并追查国家动荡的根源，布哈里说，他将通过与邻国乍得、喀麦隆和尼日尔合作，迅速关注“博科圣地”组织在尼日利亚东北部活动的暴力事件，他说他的政府有信心能够挫败犯罪分子和其他造成尼日利亚不稳定的人，因为根据尼日利亚全国独立选举委员会的数据，在尼日利亚历史上，反对派首次在民主选举中击败执政党，布哈里以大约200哈里说，他将通过与邻国乍得、喀麦隆和尼日尔合作，迅速关注“博科圣地”组织在尼日利亚东北部活动的暴力事件，他说他的政府有信心能够挫败犯罪分子和其他造成尼日利亚不稳定的人，因为根据尼日利亚全国独立选举委员会的数据，在尼日利亚历史上，反对派首次在民主选举中击败执政党，布哈里以大约200万票击败现任总统古德勒克·乔纳森，这场胜利是在长期的军事统治、政变和失败的民主尝试之后取得的人口大国 **ESN:** UNK[2] UNK 说，他的政府有信心它能够破坏尼日利亚经济的稳定。UNK说，他的政府有信心能够挫败犯罪分子和其他尼日利亚人。他说，尼日利亚和尼日利亚的经济由来已久。 **'DA-PN': 穆罕默杜·布哈里**说，他计划**在尼日利亚东北部地区积极抗击森林破坏。他说，他将“迅速关注”在尼日利亚东北部地区的暴力事件。**他说，他的政府有信心能够挫败罪犯。 **'DA-PN + Cover':** 穆罕默杜·布哈里说，他计划**积极打击长期困扰尼日利亚的腐败。**他说，他的政府有信心能够挫败犯罪分子，**这场胜利是在他长期的军事统治、政变和在这个非洲人口最多的国家里失败的民主尝试之后取得的。** **'DA-PN + Cover + MLO':** 穆罕默杜·布哈里说，他计划积极打击长期困扰尼日利亚的腐败。他的政府有信心能够挫败犯罪分子，**在长期的军事统治、政变和在这个非洲人口最多的国家里失败的民主尝试之后会取得这场胜利。**

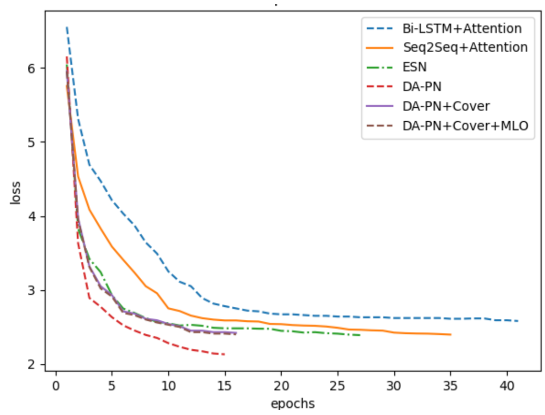


FIGURE 8. Comparison between the proposed models and baselines in terms of convergence.

optimization of the evaluation indexes in order to solve the problem of insufficient attention at the decoder. This will be a subject of our future research.

There are some limitations and deficiencies in the proposed ATS models to overcome in the future, as follows:

1) Although some semantic similarity is used, further mining of features could be utilized, e.g., part of speech, location, semantic understanding, etc.
2) Both ends (encoder and decoder), used in the proposed ATS models, utilize a small number of network layers, whereas further deepening of the neural network would allow the models to learn more information.
3) A relatively simple reward function is used for the hybrid learning of the ROUGE index. Further optimization of the reward function will be explored as a research direction in the future.

V. CONCLUSION

This paper has put forward enhancements to the structure of the attention-based bi-directional LSTM model ('Bi-LSTM + Attention') and the attention-based sequence model ('Seq2Seq + Attention') in order to improve the Automatic Text Summarization (ATS). Firstly, a novel enhanced semantic network (ESN) model has been proposed, which works out the semantic similarity between the encoder and decoder, and maximizes the semantic relevance during training, which increases the probability of generating more accurate text summaries, thus also improving the correlation with the source text. Secondly, aiming at solving the problem of unregistered words, a novel 'DA-PN' model has been proposed, which utilizes decoder attention (DA) based on a pointer network (PN). In addition, simultaneous attention is paid to both the encoder and decoder, resulting in more

[2] *UNK*, in the ESN model, represents the unregistered words.

accurate text summaries. Thirdly, it has been proposed to combine the elaborated 'DA-PN' model with a coverage mechanism integrating multi-attention. In the resultant 'DA-PN + Cover' model, by using the attention distribution of the encoder and decoder in all the previous time steps, the attention of the current time step is affected in a positive way, leading to finding more accurate words for inclusion in the text summaries by avoiding repeated words. Lastly, in order to prevent the spread of cumulative errors in generated text summaries, it has been proposed to add a mixed learning objective (MLO) function [12] to the 'DA-PN + Cover' model. The resultant 'DA-PN + Cover + MLO' model is the best performing one among the ATS models proposed in this paper. It considers the evaluation as a part of the model iteration along with a global reward, thus further increasing the readability of generated text summaries.

The performance of the four proposed ATS models was compared to that of two baselines, 'Bi-LSTM + Attention' and 'Seq2Seq + Attention', and seven state-of-the-art models, using short-text and long-text corpora. The obtained experimental results demonstrated the superiority of the elaborated ATS models compared to the baselines and some of the state-of-the-art models. By using the blended learning of MLO, the best performing proposed model, 'DA-PN + Cover + MLO', opens up a prospective to improve further the accuracy of automatically generated text summaries by optimization of the evaluation indexes, which could effectively solve the problem of insufficient attention at the decoder. This research direction will be explored in the future.

REFERENCES

[1] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "A decomposition-based multi-objective optimization approach for extractive multi-document text summarization," *Appl. Soft Comput.*, vol. 91, Jun. 2020, Art. no. 106231.

[2] A. Hernandez-Castaneda, R. A. Garcia-Hernandez, Y. Ledeneva, and C. E. Millan-Hernandez, "Extractive automatic text summarization based on lexical-semantic keywords," *IEEE Access*, vol. 8, pp. 49896–49907, 2020.

[3] M. Yang, X. Wang, Y. Lu, J. Lv, Y. Shen, and C. Li, "Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint," *Inf. Sci.*, vol. 521, pp. 46–61, Jun. 2020.

[4] Y. Guan, S. Guo, R. Li, X. Li, and H. Zhang, "Frame semantics guided network for abstractive sentence summarization," *Knowl.-Based Syst.*, vol. 221, Jun. 2021, Art. no. 106973.

[5] H. Zhang, Y. Lan, L. Pang, J. Guo, and X. Cheng, "Detecting the relevant contexts with self-attention for multi-turn dialogue generation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 3721–3730.

[6] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, Aug. 2020.

[7] J. Deng, L. Cheng, and Z. Wang, "Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification," *Comput. Speech Lang.*, vol. 68, Jul. 2021, Art. no. 101182.

[8] S. Song, H. Huang, and T. Ruan, "Abstractive text summarization using LSTM-CNN based deep learning," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 857–875, Jan. 2019.

[9] M. Sangiorgio and F. Dercole, "Robustness of LSTM neural networks for multi-step forecasting of chaotic time series," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110045.

[10] C.-W. Chen, S.-P. Tseng, T.-W. Kuan, and J.-F. Wang, "Outpatient text classification using attention-based bidirectional LSTM for robot-assisted servicing in hospital," *Information*, vol. 11, no. 2, p. 106, Feb. 2020.

[11] A. Abdi, S. Hasan, S. M. Shamsuddin, N. Idris, and J. Piran, "A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106658.

[12] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proc. 6th Int. Conf. Learn. Represent.*, Ottawa, ON, Canada, 2018, pp. 1–13.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, Ottawa, ON, Canada, 2014, pp. 3104–3112.

[14] C. Chootong, T. K. Shih, A. Ochirbat, W. Sommool, and Y.-Y. Zhuang, "An attention enhanced sentence feature network for subtitle extraction and summarization," *Expert Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 114946.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3th Int. Conf. Learn. Represent.*, Washington, DC, USA, 2015, pp. 1–15.

[16] Z. Deng, F. Ma, R. Lan, W. Huang, and X. Luo, "A two-stage Chinese text summarization algorithm using keyword information and adversarial learning," *Neurocomputing*, vol. 425, pp. 117–126, Feb. 2021.

[17] Q. Wang and J. Ren, "Summary-aware attention for social media short text abstractive summarization," *Neurocomputing*, vol. 425, pp. 290–299, Feb. 2021.

[18] W. Yang, Z. Tang, and X. Tang, "A hierarchical neural abstractive summarization with self-attention mechanism," in *Proc. 3rd Int. Conf. Autom., Mech. Control Comput. Eng. (AMCCE)*, Beijing, China, 2018, pp. 514–518.

[19] L. Huang, W. Chen, Y. Liu, S. Hou, and H. Qu, "Summarization with self-aware context selecting mechanism," *IEEE Trans. Cybern.*, early access, Jan. 11, 2021, doi: 10.1109/TCYB.2020.3042230.

[20] B. Hu, Q. Chen, and F. Zhu, "LCSTS: A large scale Chinese short text summarization dataset," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1967–1972.

[21] S. Ma, X. Sun, J. Xu, H. Wang, W. Li, and Q. Su, "Improving semantic relevance for sequence-to-sequence learning of Chinese social media text summarization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, 2017, pp. 635–640.

[22] J. Gu, Z. Lu, H. Li, and V. O. K. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 1631–1640.

[23] X. Zhu, C. Lyu, and D. Ji, "Keyphrase generation with CopyNet and semantic web," *IEEE Access*, vol. 8, pp. 44202–44210, 2020.

[24] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 832–847, 2019.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, Washington, DC, USA, 2017, pp. 6000–6010.

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

[27] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2017.

[28] S. Gupta and S. K. Gupta, "An approach to generate the bug report summaries using two-level feature extraction," *Expert Syst. Appl.*, vol. 176, Aug. 2021, Art. no. 114816.

[29] X. Jiang, P. Hu, L. Hou, and X. Wang, "Improving pointer-generator network with keywords information for Chinese abstractive summarization," in *Proc. Int. Conf. Natural Lang. Process. Chin. Comput.*, Beijing, China, 2018, pp. 464–474.

[30] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, Berlin, Germany, 2016, pp. 280–290.

[31] P. Li, W. Lam, L. Bing, and Z. Wang, "Deep recurrent generative decoder for abstractive text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 2091–2100.

[32] F. Liu and Y. Liu, "Exploring correlation between ROUGE and human evaluation on meeting summaries," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 187–196, Jan. 2010.

[33] L. Hua, X. Wan, and L. Li, "Overview of the NLPCC 2017 shared task: Single document summarization," in *Proc. Conf. Natural Lang. Process. Chin. Comput.*, Dalian, China, 2018, pp. 942–947.

[34] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.

[35] C. Chen, D. Han, and J. Wang, "Multimodal encoder-decoder attention networks for visual question answering," *IEEE Access*, vol. 8, pp. 35662–35671, 2020.

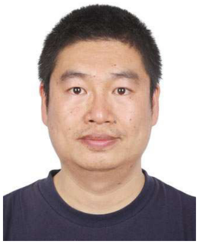
JIAWEN JIANG received the B.S. degree from the North China University of Science and Technology, China, in 2020, where he is currently pursuing the master's degree. His research interests include recommender systems, natural language processing (NLP), relevance vector machines, and neural networks.

HAIYANG ZHANG received the B.S. degree from Jilin University, China, in 2013, and the Ph.D. degree from the University of Limerick, Ireland, in 2018. She is currently a Research Associate with the Department of Computer Science, The University of Sheffield, U.K. Her current research interests include recommender systems, data mining, collaborative filtering, and NLP.

CHENXU DAI received the B.S. and M.S. degrees from the North China University of Science and Technology, China. She is currently a Research Associate with the North China University of Science and Technology. Her research interests include neural networks, intelligent optimization algorithms, and deep learning.

QINGJUAN ZHAO received the B.S. and M.S. degrees from the North China University of Science and Technology, China, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with Beihang University, China. Her research interests include NLP, support vector machines, and data mining.

HAO FENG received the B.S. and M.S. degrees from the North China University of Science and Technology, China, in 2017 and 2020, respectively. He is currently a Research Associate with the North China University of Science and Technology. His research interests include neural networks and deep learning.



ZHANLIN JI (Member, IEEE) received the M.Eng. degree from Dublin City University, Ireland, in 2006, and the Ph.D. degree from the University of Limerick, Ireland, in 2010. He is currently a Professor with the North China University of Science and Technology, China, and an Adjunct Researcher with the Telecommunications Research Centre (TRC), University of Limerick. He has authored/coauthored more than 100 research papers in refereed journals and conferences. His research interests include ubiquitous consumer wireless world (UCWW), the Internet of Things (IoT), cloud computing, big data management, and data mining.

IVAN GANCHEV (Senior Member, IEEE) received the engineering *(summa cum laude)* and Ph.D. degrees from Saint-Petersburg University of Telecommunications, in 1989 and 1995, respectively. He is currently an International Telecommunications Union (ITU-T) Invited Expert and an Institution of Engineering and Technology (IET) Invited Lecturer. He is lecturing at the University of Limerick, Ireland, and Plovdiv University "Paisii Hilendarski", Bulgaria. He was involved in more than 40 international and national research projects. He has served on the TPC for more than 330 prestigious international conferences/symposia/workshops. He has authored/coauthored one monographic book, three textbooks, four edited books, and more than 280 research papers in refereed international journals, books, and conference proceedings. He is on the editorial board of multiple international journals and has served as a guest editor for multiple international journals.