Text summarization based on multi-head self-attention mechanism and pointer network

Dong Qiu^{1,2} · Bing Yang¹

Received: 4 February 2021 / Accepted: 5 September 2021 / Published online: 2 October 2021
© The Author(s) 2021

Abstract

Existing text summarization methods mainly rely on the mapping between manually labeled standard summaries and the original text for feature extraction, often ignoring the internal structure and semantic feature information of the original document. Therefore, the text summary extracted by the existing model has the problems of grammatical structure errors and semantic deviation from the original text. This paper attempts to enhance the model's attention to the inherent feature information of the source text so that the model can more accurately identify the grammatical structure and semantic information of the document. Therefore, this paper proposes a model based on the multi-head self-attention mechanism and the soft attention mechanism. By introducing an improved multi-head self-attention mechanism in the model coding stage, the training model enables the correct summary syntax and semantic information to obtain higher weight, thereby making the generated summary more coherent and accurate. At the same time, the pointer network model is adopted, and the coverage mechanism is improved to solve out-of-vocabulary and repetitive problems when generating abstracts. This article uses CNN/DailyMail dataset to verify the model proposed in this article and uses the ROUGE indicator to evaluate the model. The experimental results show that the model in this article improves the quality of the generated summary compared with other models.

Keywords Automatic text summarization · Pointer generation network · Multi-head self-attention mechanism

Introduction

Currently, the internet generates a large quantity of text data at all times, and the problem of text information overload is becoming increasingly serious. It is necessary to reduce dimensionality for various types of texts, so the task of automatic text summarization has emerged. Automatic text summarization [1] is an important research field of natural language processing. Automatic text summarization extracts a paragraph of content from the original text or generates a paragraph of new content to summarize the main information of the original text. That is, compressing and refining the orig-

inal text provides users with a concise content description. It helps users save considerable time, and users can obtain all the key information points of the text from the abstract without having to read the entire document. Recently, text summaries have been used in many aspects of life, such as generating report summaries and automatically generating news headlines [2–5]. Currently, the research directions of automatic text summarization can be divided into extractive text summarization [6,7] and abstractive text summarization [8]. Extractive text summarization extract key sentences and keywords from the original text to generate summaries. This method may generate incoherent abstracts and poor control of the number of words. The effect depends entirely on the type of original text. Abstractive text summarization are generated by rewriting the content of the original text, allowing the generated summaries to contain new lexicons or phrases, which is more flexible. Therefore, this article focuses on Abstractive text summarization.

Recently, some researchers use the seq2seq (sequence-to-sequence) model to conduct abstractive text summarization studies [9–11], simulating the thinking process of people reading documents, first extracting the original text infor-

✉ Dong Qiu
qiudong@cqupt.edu.cn ; dongqiumath@163.com
Bing Yang
s190231195@stu.cqupt.edu.cn

¹ School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Nan'an, Chongqing 400065, China

² College of Science, Chongqing University of Posts and Telecommunications, Nan'an, Chongqing 400065, China

mation globally, and then summing up the main ideas of the original text. Cho et al. [12] and Sutskever et al. [13] proposed a sequence-to-sequence model composed of an encoder and a decoder to solve the machine translation problem. Bahdanau et al. [14] gradually applied the sequence-to-sequence model to abstractive text summarization. Rush et al. [15] proposed to use sequence-to-sequence model and attention mechanism to generate abstractive text summaries on DUC-2004 [16] and Gigaword, and use CNN to encode the original text. The model has the ability to extract text information and generate text summaries. However, the generated abstract has some deviating main information from the original text and out of vocabulary. Then, Chopra et al. [17] used the same encoder to encode the original document, but they used a recurrent neural network as a decoder and this model greatly improved the effect of abstract generation. Nallapati et al. [18] proposed to use keywords and attention mechanisms to extract the original text and encode the original text. Zhou et al. [19] proposed the use of selective encoding to encode the original text, so that the information extraction of the original text was enhanced.

In the process of using the sequence-to-sequence model, the researchers found that the model can extract information from the original text, but the text summary generated by the model has out-of-vocabulary and word repetition problems. To solve the out-of-vocabulary problem in the generated abstract, Gu et al. [20] used the copy mechanism to directly copy the important words in the original text instead of the out-of-vocabulary to generate the abstract. See et al. [21] used the coverage mechanism and pointer generation network model to provide a solution for out-of-vocabulary and word repetition. The coverage mechanism can avoid giving multiple high attention weight values at the same position when generating a summary and can effectively reduce the problem of repeated words in abstract generation. The pointer generation network introduces the pointer network on the basis of the sequence-to-sequence model [22]. In the pointer generation network, two probabilities are generated when generating the summary. That is, the probability that the current word is generated from the vocabulary and the probability that it is copied at the pointer. The network is used to learn the weights between the two probabilities. The pointer generation network can effectively adapt to the extractor and abstract network [23] and has been used in many abstract generation tasks [24,25]. Zhou et al. [26] continued to propose SeqCopyNet to enhance the readability of the generated abstract and copy the complete input sentence sequence to the abstract. [27] used word sense ambiguity to improve query-based text summarization. Wang et al. [28] and Liang et al. [29] use reinforcement learning to introduce to text summarization. [30] use a dual encoding model including the primary and the secondary encoders.

Some researchers use the traditional soft attention mechanism to extract the key information of the original text [31], but it cannot extract the various semantic and grammatical information inside the original text, resulting in the problems of grammatical structure errors and semantic deviations from the original text. Vaswani et al. [32] proposed a transformer model, in which a new self-attention mechanism model was used to extract the information of the input text. Experiments show that the self-attention mechanism can quickly extract important features of sparse data. It has been widely used in natural language processing tasks [33–35]. The self-attention mechanism is an improvement to the attention mechanism, which reduces the dependence on external information and better captures the internal correlation of data or features. The self-attention mechanism uses the attention mechanism to calculate the association between each word and all other words so that the model can learn relevant information in different representation subspaces. Text summaries have experienced how to extract original text information, how to generate summaries, and at the same time alleviate the problems of unregistered words and repetition in the process of generating summaries. However, in the process of extracting original text information, learning the semantics, grammatical structure and determining key information of the original text itself still need to be studied. And further consider reducing the word repetition rate. Thus, this is a research gaps in an imperfect Seq2Seq model [36]. In this model, this shortage of research was fulfilled along with solve problems that arise in the sequence-to-sequence model. Contributions of previous author(s) are elaborated in Table 1.

The rest of this article is organized as follows. The second section introduces the problem definition and assumptions of automatic text summarization. The third section describes the method proposed in the text in detail. The fourth section introduces the baseline, data set, experiment and results analysis of this article. Finally, the article ends in the fifth section.

Problem definition and assumptions for solving problem

Problem definition

With the rapid development of deep learning in recent years, sequence-to-sequence models have been widely used in sequence problems such as machine translation, question answering, and text summarization. However, sequence-to-sequence models have many problems, such as out-of-vocabulary problems, which generate repetitive problems [15]. To solve these problems, See et al. [21] proposed a pointer generation network, which adds a pointer network

Table 1 Contributions of previous authors

Author(s)	Basic model	Encoder	Decoder	SSIEM	CM	OOV
Cho and Gulcehre (2014)	Seq2seq	RNN	RNN	RNN	NA	NA
Sutskever and Vinyals (2014)	Seq2seq	LSTM	LSTM	LSTM	NA	NA
Vinyals and Fortunato (2015)	Seq2seq	RNN & attention	Pointer networks	Attention	NA	NA
Rush and Chopra (2015)	Seq2seq	Bag-of-words & convolutional & attention	RNN	Convolutional & attention	NA	NA
Chopra and Auli (2016)	Seq2seq	Attentive encoder	Recurrent decoder	Attention	NA	NA
Nallapati and Zhou (2016)	Seq2seq	LSTM & attention	Generator-pointer	Capturing keywords & attention	NA	NA
Gu and Lu (2016)	Seq2seq	LSTM & attention	copynet	Attention	NA	Yes
Zhou and Yang (2017)	Seq2seq	LSTM	LSTM	Selective encoding & attention	NA	NA
See and Liu (2017)	Seq2seq	LSTM	GGP	Attention	Yes	Yes
Vaswani and Shazeer (2017)	Seq2seq	MHA & FFNN	MMHA & MHA & FFNN	MHA	NA	NA
Zhou and Yang (2018)	Seq2seq	GRU & attention	SeqCopyNet	Attention	NA	Yes
Yao and Zhang (2018)	Seq2seq	Dual encoder & GRU	GRU	Primary encoder & Secondary encoder & attention	NA	NA
Lei and Luan (2019)	Seq2seq	LSTM & attention	LSTM & RL	Attention	NA	NA
Liang and Du (2020)	Seq2seq	LSTM	LSTM & RL	Attention & RL	NA	NA
Fan and Gong (2021)	Seq2seq	LSTM & Mask Attention	LSTM	Mask Attention	NA	NA
This paper	Seq2seq	LSTM & MHA & Attention	LSTM & attention & Gate mechanism	MHA & attention	Yes	Yes

SSIEM Syntactic semantic information extraction method, *CM* coverage mechanism, *OOV* out-of-vocabulary, *NA* not applicable, *RNN* recurrent neural network, *GGP* generator generation pointer, *LSTM* long short-term memory, *GRU* gated recurrent unit, *MHA* multi-head attention, *MMHA* masked multi-head attention, *FFNN* feed forward neural network, *SeqCopyNet* sequential copying networks, *RL* reinforcement learning

and a coverage mechanism to the seq2seq model based on the attention mechanism to alleviate the problems caused by the use of seq2seq. The pointer generation network uses the traditional soft attention mechanism, which cannot extract the various semantic and grammatical information inside the original text, resulting in grammatical structure error and semantic deviation problems from the original text in the generated abstract. Therefore, this article believe that it is necessary to learn the semantics and grammatical structure of the original text and to determine the key information. Due to the particularity and diversity of long text summaries, the input article is usually multiple sentences with multiple subspaces. If the model only focuses on one space, some information will be lost. At the same time, it is noticed that the self-attention mechanism is insufficient in extracting key features of the original text. When reproducing the pointer generation network, it was found that the effect of the coverage mechanism gradually weakened as the length of the forecast summary increased. Therefore, this research attempts to enhance the extraction and coding of original text. In addition, further reduce the repetition rate of words.

Assumptions for solving problem

When extracting information from the input article. If the model has different weights for the combination of original text information, according to training, increase the weights of the grammatical structure and semantic information required by the abstract to enhance the expressive ability of the model. Figure 1 shows a case of a text summarization task. In the reference, the key information rowing team at Washington University and attacked by flying carp are from S1 and S2, respectively, showing that the information is from different parts of the original text when the summary is generated. When different weights are set for S1, S2, S3, S4, and S5, different summary effects can be obtained. Since the multi-head self-attention mechanism has insufficient ability to extract the key features of the original text, the soft attention mechanism can be used to focus on the key features of the text. To obtain the key features and the contextual semantic and grammatical information of the text simultaneously, this paper use the gate mechanism to fuse the key features and the contextual semantic and grammatical information of the text. To solve the above problems, this paper proposes two models to verify the assumptions. They are based on a multi-headed self-attentive pointer network (MSAPN) and a multi-headed dual attention pointer network (MDAPN). At the same time, to reduce the repetition rate of words in long texts, the coverage mechanism is improved.

Model

In this section, the attention mechanism and the two proposed models are introduced, namely the MSAPN model and the MDAPN model. The MSAPN model is a sequence-to-sequence model based on the multi-head self-attention mechanism, as shown in Fig. 2. In this model, a new multi-headed self-attention mechanism is established, which can effectively extract the complex semantic information of the original text. To avoid generating duplication and out-of-vocabulary problems, a pointer network [22] and a coverage mechanism [21] are introduced. The MDAPN model is an improvement of the MSAPN model. To simultaneously extract the semantic information and important features of the original text, a gate mechanism is introduced, as shown in Fig. 3.

Self-attention mechanism

In [32], an attention mechanism that is different from the traditional soft attention mechanism is proposed, which reduces the introduction of external information and can effectively extract the semantic information of the original text. In the text, the original text converts the input information into a word embedding $X = \{x_1, x_2, \dots, x_l\}$ $X \in \mathbb{R}^{l \times e}$, where l is the length of the sentence, and e is the word embedding dimension for each word. X uses a BiLSTM encoder to encode the input information and obtain the sequence of encoded hidden states of the input information $X_{lstm} \in \mathbb{R}^{l \times dim}$, where dim is twice the dimensionality of the LSTM hidden layer. In the self-attention mechanism, the three matrices W_q , W_k , and W_v are defined and use these three matrices to perform three linear transformations on all word vectors, where W_q , W_k , $W_v \in \mathbb{R}^{dim \times d}$, and d is a hyper-parameter this paper set. Each word vector obtains the corresponding q_t , k_t , v_t . All the q_t vectors are concatenated into a large matrix and record it as the query matrix $Q \in \mathbb{R}^{l \times d}$. Join all k_t vectors into a large matrix to obtain the key matrix $K \in \mathbb{R}^{l \times d}$, and the value matrix $V \in \mathbb{R}^{l \times d}$ can be obtained in the same way:

$$Q = X_{lstm} W_q \quad (1)$$

$$K = X_{lstm} W_k \quad (2)$$

$$V = X_{lstm} W_v \quad (3)$$

Among them, W_q , W_k , W_v are all parameters that can be learned. The input of the self-attention mechanism comes from three different vectors of the same word, namely, the above three matrices, Q, K and V. The similarity of the input word vector is expressed by the multiplication of Q and K, and $\sqrt{d}$ is used for scaling; softmax is used for normalization to obtain the similar weight value between all words, and then

Fig. 1 S1, S2, S3, S4, and S5 is the original text, reference and PGN is summary

S1	It was a typical practice day for the Washington University of rowing team , but then danger came from beneath .
S2	The scene was Creve Coeur Lake outside of St. Louis early Friday morning . The team 's boat got near the dock , when suddenly a swarm of Asian carp emerged from the water and went on the attack , some even going into the boat .
S3	Team member Devin Patel described the moment of terror : `` The fish was flopping on my legs . It was so slippery that I could n't get a grip on it . "Patel screamed at teammate Yoni David , `` Yoni , get it off me ! "
S4	Thankfully , no rowers were injured during the ordeal , but the strong smell of fish lingered in the moments afterward .
S5	Watch iReporter Benjamin Rosenbaum 's video above .
reference :	
rowing team at washington university[S1] attacked by flying carp[S2] . member of the team caught the attack on video[S3] .	
PGN :	
the scene was creve coeur lake outside of st. louis. the team 's boat got near the dock , when suddenly a swarm of asian carp [S2]. no rowers were injured during the ordeal [S4].

Fig. 2 Multi-head self-attention pointer network model (MSAPN)

Fig. 3 Multi-head dual-attention pointer network (MDAPN)

it is multiplied with V to obtain the self-attention mechanism score, the formula is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (4)$$

MSAPN model

The self-attention mechanism can extract only the semantic information of the original text from one dimension. To enhance the extraction of multiple dimensions of the semantic information of the original text, the improved multi-head self-attention mechanism is used in the MSAPN model. The model can understand the input information from different subspace dimensions, as shown in Fig. 4.

Based on the self-attention mechanism, the linear matrix is changed from one set (W_q, W_k, W_v) to multiple sets ($W_0^q, W_0^k, W_0^v, \dots, W_i^q, W_i^k, W_i^v$). It can be understood that different randomly initialized linear matrices (W_q, W_k, W_v) can map the input vector to different subspaces, allowing the model to understand the input information from different spatial dimensions. It can effectively extract the semantic information of the original text.

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (5)$$

Fig. 4 Improved multi-head self-attention mechanism

$$H = (\text{head}_1, \dots, \text{head}_i, \dots, \text{head}_h) \quad (6)$$

$\text{head}_i \in \mathbb{R}^{l \times d}, H \in \mathbb{R}^{l \times d \times n}$. n is the number of multi-heads. The reshape function is used to reduce the dimensionality of $H_m \in \mathbb{R}^{l \times \text{tem}}$. $\text{tem} = d * n$. Linear transformation is used to transform it into a vector $M \in \mathbb{R}^{l \times \text{dim}}$ and the SUM function is used to merge its multi-heads to form a vector, and compress the word information of length l to one dimension. $Y \in \mathbb{R}^{\text{dim}}$. The formula is as follows:

$$H_m = \text{reshape}(H) \quad (7)$$

$$M = H_m W_m \quad (8)$$

$$Y = SUM(M) \quad (9)$$

where $W_m \in \mathbb{R}^{tem \times dim}$ is a parameter that can be learned, and the probability distribution of the dictionary can be obtained by using the softmax function:

$$P_{Self-vocab} = softmax(V'(Y) + b) \quad (10)$$

where V' and b are the parameters that can be learned. According to the dictionary probability distribution, the predicted words generated at the current time are obtained:

$$P(w) = P_{Self-vocab}(w) \quad (11)$$

In the training phase, using the cross-entropy function, the loss of w_t^* is predicted at time t :

$$loss_t = -\log P(w_t^*) \quad (12)$$

Then, the overall loss of the input sentence sequence is

$$loss = \frac{1}{T} \sum_{t=0}^T loss_t \quad (13)$$

To solve the out-of-vocabulary problem and the duplication problem of abstract generation in text summarization, the pointer generation network [21] is introduced. In the model, the hidden layer state and encoder state of each decoding step are used to calculate the weight. After the input text information passes through the bidirectional LSTM, the encoded hidden state h_i can be obtained. On the decoding side, the decoder is a one-way LSTM, and the decoding state is s_t at time t . h_i and s_t are used to calculate the attention weight of the i -th word in the original text at decoding time t :

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{attn}) \quad (14)$$

$$a^t = softmax(e^t) \quad (15)$$

$$C_t = \sum_i a_i^t h_i \quad (16)$$

where V , W_h , W_s , and b_{attn} are learnable parameters, and a^t is the attention distribution at the current moment. C_t is the context vector at the current moment. h_i is the encoder hidden state, and a_i^t is the attention weight of the i -th word in the original text at decoding time t .

In the pointer network, to solve the out-of-vocabulary problem, this model choose to extract some words from the

original text to expand the dictionary. To determine whether the word generated by the decoder at the current time step is generated or copied, a generation probability P_{gen} is introduced. When P_{gen} is 1, it means that only words from the dictionary can be used, not words in the original text. When P_{gen} is 0, it means that only words from the original text can be used, not words in the dictionary:

$$P_{gen} = \sigma(w_C^T C_t + w_s^T s_t + w_x^T x_t + b_{ptr}) \quad (17)$$

Among them, w_C^T , w_s^T , w_x^T and b_{ptr} are all parameters that can be learned. C_t is the context vector from the soft attention mechanism, s_t is the decoded hidden layer state value at the current time t of the decoding layer, and x_t is the input of the decoder at the current time. Then, the probability distribution of the predicted word at time t is

$$P(w) = P_{gen} P_{Self-vocab} + (1 - P_{gen}) \sum_{i:w_i=w} a_i^t \quad (18)$$

To solve the repetition problem caused by seq2seq, the attention weights of the previous time steps are added together to obtain the coverage vector c^t (coverage vector), and the previous attention weight decisions are used to influence the current attention weight decisions, avoiding the same. The position is repeated to avoid repeated text generation. In terms of calculation, first, calculate the coverage vector c^t :

$$c^t = \sum_{t'=0}^{t-1} a^{t'} \quad (19)$$

In the calculation process of adding the cover vector to the attention weight, c^t is used to calculate e_i^t :

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{attn}) \quad (20)$$

To avoid repetition, a loss function is needed to punish repeated attention. The coverage loss calculation method is

$$covloss_t = \sum_i \min(a_i^t, c_i^t) \quad (21)$$

Since c^t is accumulated by $a^{t'}$ when the accumulation reaches a certain number, the value in c^t will be greater than 1, and $a^{t'}$ is the value obtained by softmax, which will be less than 1. At this time, the min in formula 18 will be useless, so that as the length of the generated summary increases, the coverage mechanism is weakened. For this reason, this article modified the c^t calculation method to reduce the problem of excessive focus given to attention in the accumulation process. In addition, the large attention is reduced and the

small attention is increased. The formula is as follows:

$$c^t = uc^{t-1} + (1 - u) \left(\frac{1}{1 + e^{at-1}} \right). \quad (22)$$

The u is a hyper-parameter and $0 < u < 1$. coverage loss is a bounded quantity $\text{covloss}_t \leq \sum_i a_i^t = 1$, the final loss function is

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t) \quad (23)$$

where λ is the hyper-parameter.

MDAPT model

After using the multi-head self-attention mechanism, it is found that the semantic extraction effect of the original text is better than other models, and the soft attention mechanism can focus on the key features of the text. This paper attempted to use the two mechanisms together, as shown in Fig. 2. To build this fusion, the gate mechanism is introduced into the network to generate the probability value $P_{gate} \in [0, 1]$ required for each decoding step according to the encoded hidden layer state and the decoded hidden layer state. The formula is as follows:

$$P_{gate} = \sigma(W_y Y + W_{ec} C_t + W_{es} S_{es} + W_{ds} S_t + B) \quad (24)$$

where W_y , W_{ec} , W_{es} , W_{ds} and B are all parameters that can be learned, and σ is the sigmoid function. Y is the context vector from the multi-head self-attention mechanism, C_t is the context vector from the soft attention mechanism, S_{es} is the hidden layer state value output by the bidirectional LSTM of the coding layer, S_t is the decoded hidden layer state value at time t , and P_{gate} is used to determine whether the current word generation should focus on the semantic information of the original text or the key feature information. Therefore, the probability of current vocabulary generation is

$$P_{vocab} = \text{softmax}(V''(V^{temp}[s_t, C_t] + b) + b') \quad (25)$$

$$P_{wvocab} = P_{gate} P_{vocab} + (1 - P_{gate}) P_{Self-vocab} \quad (26)$$

where V'' , V^{temp} , b' and b are the parameters that can be learned, s_t is from the decoded hidden layer state value at time t , and $P_{Self-vocab}$ is from formula 10. Introducing the pointer network, then the generation probability of the current vocabulary is

$$P_w = P_{gen} P_{wvocab} + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t \quad (27)$$

In the training phase, to prevent repetitive problems, the same coverage mechanism as the MSAPN model are used.

Experiment

Datasets and evaluation indicators

In this experiment, the dataset CNN/DailyMail [37] is used. which is a multi-sentence summary dataset. In the CNN/DailyMail data, CNN has a total of 92,579 files, and DailyMail has a total of 219,506 files. Stanford-coreNLP is used for word segmentation, and a dictionary file is created. The size of the dictionary is 50,000. In the English news dataset, the core key features of the article generally appear in the previous part of the article, and because in the seq2seq model, the performance of the input sequence will decrease. Therefore, in the data preprocessing stage, this paper limit the maximum length of the original text to 400 and the maximum abstract length to 100. After preprocessing the dataset, it is divided into three parts: training set, validation set and test set There were 287,227 text pairs in the training set, 13,368 text pairs in the validation set, and 11,490 text pairs in the test set.

This paper use the ROUGE evaluation mechanism to evaluate this model. ROUGE-N calculates its score based on the n-gram repetition rate between the generated abstract and the real abstract. ROUGE-L is the longest common subsequence for generating digests and real digests for calculation. This paper use ROUGE in the pyROUGE package to calculate ROUGE-1, ROUGE-2 and ROUGE-L, which can intuitively and concisely reflect the effect of generating abstracts.

Experimental details

Table 1 shows the parameter definition and assignment of the model. Due to the equipment limitations, this paper set the batchSize for training and evaluation to 16. Through experiments, this paper found that the number of heads and the dimension of d in the multi-head self-attention mechanism are not as large as possible. Therefore, the number of heads is set to 4, and the dimension of d is 32. λ is from the [21] setting. u prevents the value of c^t from being greater than 1. After many experiments, $u = 0.85$, and the effect of the model is better. Figures 5 and 6 show the MDPN and PGN loss function declines. From a comparison of the two graphs, MDPAN has a smaller loss function. Note that due to the different PGNs and MDAPNs entering the batch size resulting in different steps, two models are trained in the same epoch.

Table 2 Parameter setting

Parameter	Definition	Values
e	e is the word embedding dimension for each word	128
l	Maximum length of input sentence	-
$hidden_size$	The number of features in the hidden state h	256
d	d is a hyper-parameter, $d = d_v = d_q = d_k$	32
n	n is the number of multi-head	4
p	Probability of an element to be zeroed in Dropout	0.5
$epoch$	The number of times the model was trained with the training set	10
λ	λ is a hyper-parameter	1
max_grad_norm	Clips gradient norm of an iterable of parameters	2
$batchSize$	The number of samples entered in each batch	16
$beam_size$	Optimal number of each selection in beam search	4
$optimizer$	Optimizer selected in backpropagation	Adam
lr	Optimizer learning rate	0.001
u	u is a hyper-parameter	0.85

Fig. 5 MDAPT loss **Fig. 6** PGN loss

Baselines

The model this paper propose is compared with the following models:

1. TextRank [38]: an unsupervised algorithm based on weighted graphs, which can effectively extract keywords and sentences from the original text.
2. Lead-baseline [39]: select the first three sentences in the original text as the summary, and the news dataset is used, and the important information is placed in the first paragraph, so the effect is better.
3. Seq2seq-att [15]: a bidirectional LSTM encoder and one-way LSTM decoder structure that uses the attention mechanism.

Table 3 ROUGE recall score

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-Baseline [41]	38.40	34.31	37.32
Pointer generator (ours)	37.06	16.03	34.76
MSAPN (ours)	47.58	20.24	43.14
MDAPT (ours)	43.51	18.88	39.98

4. Pointer generator network [21]: a seq2seq model based on a pointer network and coverage mechanism, which can alleviate vocabulary and repetitive problems.
5. Dual Encoding [30]: in this paper, abstractive text summarization using a dual encoding model including the primary and the secondary encoders.
6. Mask attention network [42]: this is an improved transformer model that uses a masked attention network and an improved feedforward network and self-attention mechanism.

Results and analysis

Evaluation of the MSAPN model and MDAPT model: to evaluate the superiority of this paper model, the public dataset CNN/DailyMail is used. The recall scores in ROUGE-1, ROUGE-2, and ROUGE-L scores are used, as shown in Table 3.

From Table 3, the MSAPN and MDAPN models have better ROUGE recall scores than the current model. The processing of the MSAPN and MDAPN models on the multi-head self-attention mechanism is different from other multi-head self-attention models. This models synthesize all the multi-head information. The multi-subspace information is fully used, which makes this model have a greater improvement in the ROUGE recall score. The semantic and grammatical extraction of this paper model is more effective. This paper model also uses a pointer network and an improved coverage mechanism, which can greatly reduce out-of-vocabulary problems and generate summary duplication problems, thereby improving the quality of the generated summary.

It can be seen in Table 4 that this two models have their advantages and disadvantages in the F1 value of ROUGE-1, ROUGE-2 and ROUGE-L compared with the latest model. MDAPN is better than the pointer generation network. This paper model has better abstract generation and uses improved multi-head self-attention to extract contextual semantic and grammatical information from the original text.

In Tables 3 and 5, the lead-baseline model and the MSAPN model have high ROUGE recall scores, but the ROUGE accuracy scores of these two models are very low. This indicates that with the introduction of a large amount of redundant information in these two models to improve their

Table 4 ROUGE F1 score

Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank [40]	35.65	12.59	31.65
Lead-Baseline [41]	34.91	16.32	38.61
Seq2Seq-att [40]	31.31	11.82	28.87
Pointer generator [21]	36.44	15.66	33.42
Dual Encoding [30]	40.85	18.08	37.13
Mask Attention Network [42]	40.98	18.82	37.88
MSAPN (ours)	37.41	15.92	36.08
MDAPT (ours)	39.21	17.34	37.31

Table 5 ROUGE Precision score

Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank [40]	34.94	12.42	31.08
Lead-Baseline [41]	31.92	10.71	39.90
Seq2Seq-att [40]	35.10	14.58	32.44
Pointer generator (ours)	37.13	16.29	36.25
MSAPN (ours)	32.40	13.82	32.26
MDAPT (ours)	37.33	16.80	36.67

Table 6 MDAPN uses coverage and imp-coverage, respectively. ROUGE F1 score

Model	ROUGE-1	ROUGE-2	ROUGE-L
Coverage	38.86	17.02	36.82
Imp-coverage	39.21	17.34	37.31

ROUGE recall scores, the model also learns more grammar and semantic information. This also explains why the ROUGE recall score of the MSAPN model is high. The MDAPN model has a similar scores in the ROUGE accuracy score and pointer generation network, indicating that its model is more stable in generating abstracts.

From Tables 3, 4 and 5, the MSAPN model has a higher ROUGE recall score, but the accuracy is the lowest among all comparison models. This shows that if only the multi-head self-attention model is used, it can improve the model's understanding of the semantic information of the original text. However, in regard to generating summaries, it introduces considerable redundant information, as with models such as lead-baseline. This paper combine the MDAPN model with the fusion of a soft attention mechanism and multi-headed self-attention mechanism. All three scores in ROUGE are relatively stable, and their scores are improved, indicating that this fusion of two attention mechanisms is effective and can understand the semantic information of the original sentences more accurately and can effectively improve the effectiveness of generating summaries.

Fig. 7 Text summaries generated by different models on the CNN/Daily Mail dataset

<p>Source article: Struggling Hamburg have appointed Bruno Labbadia as their new coach with immediate effect in a surprise decision with the former European champions anchored in last place with six matches left. Despite widespread speculation about a pending deal with former Mainz coach Thomas Tuchel, Hamburg said they had agreed on a 15-month contract with Labbadia, who had coached them in 2009-10. The coach 's contract is also valid for the second division should Hamburg be relegated. He takes over from interim coach Peter Knaebel who replaced sacked Joe Zinnbauer in March and will now return to his post as sports director. Hamburg have appointed Bruno Labbadia as their new coach on a 15-month contract. We decided to take this step because we consider the coach question the most important , said Hamburg CEO Dietmar Beiersdorfer in a statement. With Bruno Labbadia we now have a strong coach who is also experienced in relegation battles. Beiersdorfer said the club had ended talks with Tuchel after an agreement could not be reached. Hamburg are the only club with a perfect Bundesliga record, having played every season in the top division since the league 's creation in 1963. Labbadia returns to Hamburg for a second spell after coaching the club for one season in 2009-10 The 49-year-old Labbadia, a former Hamburg player, has also coached Bayer Leverkusen and VfB Stuttgart among other. We have no time to waste, Labbadia said. Every point matters for us and I am happy that Hamburg has such support from the fans. Because we have to be a unit until the last matchday.</p>
<p>Reference Summary: Hamburg sacked Joe Zinnbauer in March with Peter Knaebel taking over Knaebel has now returned to his original post as sports director , Club has been in talks with Thomas Tuchel but could not reach agreement , Hamburg appointed former coach Bruno Labbadia on 15-month contract.</p>
<p>Pointer Generation Network with Coverage: hamburg have appointed bruno labbadia as their new coach on a 15-month contract. the coach 's contract is also valid for the second division should hamburg be relegated . hamburg are the only club with a perfect bundesliga record.</p>
<p>MSAPT: hamburg have appointed bruno labbadia as their new coach on a 15-month contract. the coach 's contract is also valid for the second division should hamburg be relegated. hamburg are the only club with a perfect bundesliga record, having played every season in the top division since the league 's creation in 1963.</p>
<p>MDAPT: struggling hamburg have appointed bruno labbadia as their new coach on a 15-month contract. despite widespread speculation about a pending deal with former mainz coach thomas tuchel, hamburg said they had agreed on a 15-month contract with labbadia, who had coached them in 2009-10.</p>

Table 7 The improved coverage mechanism uses different hyper-parameter u and scores in MDAPN. ROUGE F1 score

u	ROUGE-1	ROUGE-2	ROUGE-L
0.2	37.80	16.68	36.52
0.5	38.86	16.75	36.76
0.85	39.21	17.34	37.31

Table 6 shows that the improved coverage mechanism is more effective than the traditional coverage mechanism and improves the effect of generating abstracts. It can also be seen in Tables 7, 8 and 9 that when u is 0.85, the effect of the model is the highest.

This paper use a trained model to predict and generate abstracts. This paper randomly select one of the original text

Table 8 The improved coverage mechanism uses different hyper-parameter u and scores in MDAPN. ROUGE Precision score

u	ROUGE-1	ROUGE-2	ROUGE-L
0.2	36.33	16.38	36.39
0.5	36.73	16.59	36.65
0.85	37.33	16.80	36.67

Table 9 The improved coverage mechanism uses different hyper parameter u and scores in MDAPN. ROUGE recall score

u	ROUGE-1	ROUGE-2	ROUGE-L
0.2	39.50	17.35	37.51
0.5	42.73	18.65	39.13
0.85	43.51	18.88	39.98

and the corresponding reference abstract to compare with the abstracts generated by the three models, as shown in Fig. 7.

In Fig. 7, the original text stated that the Hamburg team will sign Bruno Rabadia as the new coach with a contract period of 15 months. Comparing the reference abstract with these three models, it can be seen that these three models can extract some key features of the original text. Comparing the abstract generated by the pointer generation network with the reference abstract, this paper find that the pointer generation network model focuses too little attention on the key information of the original text. Compared with the pointer generation network, both the MSAPN model and the MDAPN model extract more semantic information, focusing on the semantic relationship of the original text. There are no out-of-vocabulary problems or duplication problems in the above models, which shows that this paper introduction of pointer networks and coverage mechanisms is effective. Compared with the MSAPN model, the MDAPN model has more concise summary generation, and the MSAPN model has more comprehensive summary generation, both of which can improve the effect of generating summaries.

Conclusion

In this article, two models based on attention mechanisms are proposed, namely, MSAPN and MDAPN. Both models use an improved multi-head self-attention mechanism to obtain the semantic and grammatical information inside the original text so that the generated abstract text summary has similar meaning to the semantic and grammatical information of the text. In addition, both models use pointer networks and an improved coverage mechanism to reduce duplicate content and extra-lexical problems. The experimental results show that the MSAPN model is better than the traditional pointer generation network in ROUGE Recall score, but it is not effective in ROUGE Precision score. For this reason, the proposed MDAPN model, combined with the gating mechanism, makes the model's ROUGE indicators better than the pointer generation network, it can effectively contain more original text information, and the generated abstraction is more complete. However, the F1 score value of the MDAPN model is lower than that of the existing model. Therefore, in the future, the model can be extended by considering pre-training [43] models and Reinforcement Learning [29], so that the model's F1 score can be improved. This model introduces a pre-training model to improve the extraction of text information and reduces the exposure bias in training and prediction through reinforcement learning, and further improves the generation of abstracts. In addition, hope that the model proposed in the future can be used in different fields[44].

Acknowledgements The authors would like to thank the referee for his valuable comments. This work was supported by the National Natural Science Foundations of China (Grant nos. 12171065 and 11671001).

Declarations

Conflict of interest The authors declare that they have no conflicts of interest regarding the publication of this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Silber HG, McCoy KF (2002) Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Comput Linguist* 28(4):487–496
2. Lei J, Luan Q, Song X et al (2019) Action Parsing-Driven Video Summarization Based on Reinforcement Learning. *IEEE Trans Circuits Syst Video Technol* 29(7):2126–2137
3. Hori C, Furui S (2013) A new approach to automatic speech summarization. *IEEE Trans Multimedia* 5(3):368–378
4. Raposo F, Ribeiro R, de Matos DM (2016) Using generic summarization to improve music information retrieval tasks. *IEEE/ACM Trans Audio Speech Lang Process* 24(6):1119–1128
5. Rastkar S, Murphy GC, Murray G (2014) Automatic summarization of bug reports. *IEEE Trans Software Eng* 40(4):366–380
6. Liu SH, Chen KY, Chen B et al (2015) Combining relevance language modeling and clarity measure for extractive speech summarization. *IEEE/ACM Trans Audio Speech Lang Process* 23(6):957–969
7. Bidoki M, Moosavi MR, Fakhrahmad M (2020) A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities. *Inf Process Manag* 57(6):102341
8. Carenini G, Cheung JCK, Pauls A (2013) Multi-document summarization of evaluative text. *Comput Intell* 29(4):545–576
9. Liu T, Wei B, Chang B (2017) Large-scale simple question generation by template-based seq2seq learning. In *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, Cham, pp 75–87
10. Chung E, Park JG (2017) Sentence-chain based Seq2seq model for corpus expansion. *ETRI J* 39(4):455–466
11. Chu Y, Wang T, Dodd D (2015) Intramolecular circularization increases efficiency of RNA sequencing and enables CLIP-Seq of nuclear RNA from human cells. *Nucleic Acids Res* 43(11):e75–e75
12. Cho K, Gulcehre C (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Empirical Methods in Natural Language Processing*. Stroudsburg 2:1724–1734

13. Sutskever I, Vinyals O, Le Q V (2014) Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Cambridge: MIT Press:3104–3112
14. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. *Comput Sci* 2:1–8
15. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. *Empirical Methods Nat Lang Process* 8(1):379–389
16. Over P, Dang H, Harman D (2007) DUC in context. *Inf Process Manag* 43(6):1506–1520
17. Chopra S, Auli M, and Rush A M (2016) Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*:93–98
18. Nallapati R, Zhou B, Santos C (2016) Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*:280–290
19. Zhou Q, Yang N, Wei F, Zhou M (2017) Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* 1:1095–1104
20. Gu J, Lu Z, Li H et al (2016) Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 1:1631–1640
21. See A, Liu P J et al (2017) Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* 1:1073–1083
22. Vinyals O, Fortunato M and Jaitly N (2015) Pointer networks. In *Advances in Neural Information Processing Systems*:2692–2700
23. Chen Y C, Bansal M (2018) Fast abstractive summarization with reinforce-selected sentence rewriting. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* 1:10–27
24. Guo H, Pasunuru R, and Bansal M (2018) Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* 1:687–697
25. Sun F, Jiang P, Sun H, Pei C, Ou W, and Wang X (2018) Multi-source pointer network for product title summarization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*:7–16
26. Zhou Q, Yang N, Wei F, Zhou M (2018) Sequential copying networks. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* 32(1):4987–4995
27. Rahman N, Borah B (2020) Improvement of query-based text summarization using word sense disambiguation. *Complex Intell Syst* 6:75–85
28. Wang L, Yao J, Tao Y et al (2018) A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (**Intelligence**:4453–4460)
29. Liang Z, Du J, Li C (2020) Abstractive social media text summarization using selective reinforced Seq2Seq attention model. *Neurocomputing* 410:432–440
30. Yao K, Zhang L, Du D et al (2018) Dual encoding for abstractive text summarization. *IEEE Trans Cybern* 50(3):985–996
31. Zhang J, Zhao Y, Li H et al (2018) Attention with sparsity regularization for neural machine translation and summarization. *IEEE/ACM Trans Audio Speech Lang Process* 27(3):507–518
32. Vaswani A, Shazeer N, Parmar N (2017) Attention is all you need. *Adv Neural Inf Process Syst* 2:5998–6008
33. Xiao X, Zhang D, Hu G (2020) CNN-MHSA: A convolutional neural network and multi-head self-attention combined approach for detecting phishing websites. *Neural Netw* 125:303–312
34. Zhang Y, Gong Y, Zhu H et al (2020) Multi-head enhanced self-attention network for novelty detection. *Pattern Recogn* 107:107486
35. Wei P, Zhao J, Mao W (2021) A graph-to-sequence learning framework for summarizing opinionated texts. *IEEE/ACM Trans Audio Speech Lang Process* 29:1650–1660
36. Dey BK, Pareek S, Tayyab M et al (2020) Autonomation policy to control work-in-process inventory in a smart production system. *Int J Prod Res* 59(4):1258–1280
37. Hermann K M, Kocisky T et al (2015) Teaching machines to read and comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*:1–14
38. Barrios F, López F, Argerich L, Wachenchauser R (2016) Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*
39. Grusky M, Naaman M, Artzi Y (2018) Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*:708–719
40. Li J, Zhang C, Chen X (2019) Abstractive Text Summarization with Multi-Head Attention. In *2019 International Joint Conference on Neural Networks*:1–8
41. Lebanoff L, Song K, Chang W and Liu F (2019) Scoring sentence singletons and pairs for abstractive summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*:1–15
42. Fan Z, Gong Y, Liu D (2021) Mask Attention Networks: Rethinking and Strengthen Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*:1692–1701
43. Devlin J, Chang M W, Lee K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*(1):4171–4186
44. Sarkar B, Sarkar M, Ganguly B (2021) Combined effects of carbon emission and production quality improvement for fixed lifetime products in a sustainable supply chain management. *Int J Prod Econ* 231:107867

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.