Unleashing the Potential of Attention Model for News Headline Generation

1st Yong Liao, 2nd Kui Meng, 3rd Jianshen Zhang, 4th Gongshen Liu*
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
{liaoyong, mengkui, zjs_007, lgshen}@sjtu.edu.cn

***Abstract*—Headline generation is a special summarization generation task and the difficulty lies in requiring the generated headline to be concise, fluent and informative. Limited by the ability of commonly used encoder and decoder modules to capture long-term dependencies in seq2seq tasks, previous work rarely researched headline generation by end-to-end methods. However, the recent success of Transformer model and its subsequent improved versions demonstrate their remarkable performance on seq2seq tasks, which provide us with a feasible solution. In this paper, we propose a novel model Transformer(XL)-CC to generate headline from the perspective of understanding the whole text, the segment-level recurrence mechanism and relative positional encoding make our model learn ultra-long-term dependencies. In addition, we combine the copy and coverage mechanisms to generate more readable titles. Experimental results on the NYT and Chinese LSCC news datasets also confirm that our method significantly achieves better performance on the headline generation task.**

***Index Terms*—Headline generation, seq2seq tasks, end-to-end, Transformer(XL)-CC, NYT, LSCC news**

I. Introduction

Text summarization is an important task in Natural Language Processing. It can effectively solve the problem of information overflow and information overload. However, summarization of the whole document in many cases does not meet our needs, we desire a more compact summary task specialized to headline generation [1]. For example, in mobile news client scenarios [2], users often only decide whether the news meets his preferences based on the content of the short title in the push message, and then decide whether to read or not, which directly affects the reading quantity of the news in the mobile news client. In addition, there are many application scenarios and potential applications for headline generation, such as machine writing, text compression, etc.

The abstract system can be roughly categorized into two types of models, one is the extractive model, and the other is the abstractive model [3]. The extractive model cuts out important segments from the original text and reassembles them into a coherent paragraph, which is the final summarization. The abstractive model is a generalization of the document which exploit key words, events or phrases to form a piece of text as result, some words of the summarization are not necessarily presented in the original document. Because the fatal restriction of the extractive model is the lack of text comprehension, some important information may be lost in the generated abstract and the text may not be fully generalized, or the information may be redundant when the abstract is constructed by keywords. Therefore, abstractive model has natural advantages similar to human activities and is more favored by scholars. With the development of natural language generation technology in recent years, the potential of abstractive model is gradually being fully unleashed, the summarization task has risen to a heyday of research. There is no doubt that our headline generation task also adopts the more reasonable abstractive model. The difficulty of the headline generation task is not only to generate sentences that can cover the key information of the entire document, but also a higher requirement is to make the generated sentences short, fluent, and novel.

As the Seq2Seq model [4] achieves significant results in machine translation and text generation, the attention model that is subsequently proposed pushed Seq2Seq to the peak. However, the huge success of the Transformer model [5] subsequently provides us with an effective method that only relies on the self-attention mechanism to achieve amazing results on the sequence to sequence task. Reference [6] applies the Encoder-Decoder model to headline generation where the lead sentences are fed to the encoder as input, his approach assumes that the main information of the news articles is mostly the lead sentences, and ignores those cases where the lead sentences do not overlap the main purpose of the news. To solve such a problem, [7] comes up with a coarse-to-fine approach which well avoids the problem of limited long-term dependencies learned by the Seq2Seq model, they propose a sum-hieratt model which first utilize several statistical summarization approaches to obtain summary sentences instead of full text as input to the encoder, and then combine hierarchical attention to decode. It is not an end-to-end method, and requires additional summarization generation tools, the final quality of headline generation also depends on the summarization generation method. Our research aims to make further improvements based on reference [7] and proposes a nearly end-to-end approach to achieve automatic headline generation task based on understanding whole source document, and solve the OOV (out of vocabulary) and generating duplicate fragments problems.

*Gongshen Liu is corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Our contributions of this paper are as followings: we present a novel model Transformer(XL)-CC based on the Transformer-XL architecture [8] and first combine the BPE (byte-pair encoding) method [9] to solve the headline generation task from end to end. Our model learns longer dependencies than RNN and Vanilla Transformer [10], and gets better results in both long and short sequences. In addition, we adopt the CopyNet [11] to solve the common OOV problem in summarization generation task, and the coverage mechanism [12] to avoid the problem of duplicate fragments in the generated headline during inference stage. We conduct experiments on both the New York Times news corpus and the Chinese LSCC news corpus. Compared with various baselines, the experimental results demonstrate that our model can significantly improve the performance of the previous models on the headline generation task.

In Sect.II, we introduce the related work of headline generation. In Sect.III, we explain the preliminaries about the architecture and idea of Transformer model. In Sect.IV, we describe our model framework and explain our approach. We present the experimental results in Sect.V and finally we conclude this paper in Sect.VI.

II. Related Work

There have been numerous studies on the task of headline generation. The traditional methods usually utilize text compression technology to produce the headline, or extract important information such as keywords, phrases, concepts and then make use of natural language generation techniques to integrate them. Reference [1] proposes a statistical model for content selection and surface realization, which is capable of generating summaries shorter than sentences by building so-called count-based noisy-channel model. Reference [13] presents Hedge Trimmer system using linguistically-motivated heuristics for sentence compression. In [14], they explore whether the extracted key words and phrases are suitable candidates for inclusion in final headline by Singular Value Decomposition (SVD) algorithm. Reference [15] adopts novel word features for keyword extraction by Wikipedia and then employ keyword clustering to construct the target headline. Reference [16] applies methods of Noisy-OR Bayesian network to their multi-sentence compression model named HEADY based on the generalization of syntactic patterns. Reference [17] proposes an event-driven model which extract a set of structural events to identify a key event chain and fuse them by a novel multi-sentence compression algorithm. These methods are roughly first to extract the key information and then synthesize the title.

In recent years, with the proposal of the Seq2Seq model and the Transformer model centered on the attention mechanism, the fully abstractive method based on these new models has gradually surpassed the traditional extractive method. Reference [6] first adopts the neural encoder-decoder framework combined with attention mechanism for sentence summarization. Reference [18] adds the position information of words on the basis of using RNN as an encoder, which demonstrates progressive improvement. Reference [7] considers that lead sentences do not always conclude the entire document, they propose a sum-hieratt model which adopts a coarse-to-fine method, first utilizing several statistical summarization approaches to obtain summary sentences instead of full text as input fed to the encoder, and then combining hierarchical attention to decode. Based on the work of [7], our paper further proposes an end-to-end processing method to solve the headline generation task. Reference [22] presents the encoder-decoder model based on LSTM [19] which also simply takes lead sentences as input to reformulate a headline, and it is worth mentioning that they find that reversing the input sentences usually produce better results.

Almost none of the methods mentioned above solve the natural language generation task of headline generation from an end-to-end perspective, so our research work is different from them and quite meaningful.

III. Preliminaries

In this section, we introduce the preliminaries about Transformer framework [5], which is the basis of the model adopted by our headline generation approach. Although the traditional RNN structure is quite prevalent in NLP processing tasks, it remains two problems: (a) the circular network structure leads to limited parallel computing and low computing efficiency; (b) the long-term dependencies still cannot be well solved. The most prominent features of Transformer are its excellent parallel computing ability and the advantage to solve long-term dependencies problem. It is essentially an attention structure that can directly capture global information without the procedure to gradually recurse like RNN.

A. Encoder of Transformer Framework

The Encoder module of the Transformer structure consists of $N = 6$ identical layers, and each Layer consists of two sublayers, namely multi-head self-attention mechanism and fully connected feed-forward network. Each of these sublayers has a residual connection and normalization, therefore the output of the sub-layer can be expressed as:

$$sub\_layer\_output = LayerNorm(x + (SubLayer(x))) \quad (1)$$

These two sublayers are explained as follows.

1) Multi-head Self-attention: In the encoder module, each input word is multiplied by three weight matrices after word embedding to create the Query vector($\mathbf{Q}$), the Key vector($\mathbf{K}$) and the Value vector($\mathbf{V}$), attention mechanism is simply expressed in the following form:

$$attention\_output = Attention(Q, K, V) \quad (2)$$

Multi-head attention is to project $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ through $\mathbf{h}$ different linear transformations, and finally stitch the different attention results together, self-attention is to take the same $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$:

$$MultiHead(Q, K, V) = Concat\left(head_1, \ldots, head_h\right) W^O \quad (3)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$.

In addition, the calculation of attention in Transformer model adopts scaled dot-product which makes the gradient more stable, $\mathbf{d_k}$ is the number of dimensions of the Key vector:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

2) Position-wise Feed-forward Networks: The second sub-layer is a fully connected layer whose main function is to provide non-linear transformations. The dimension of the attention output is $[bsz * seq\_len, num\_heads * head\_size]$ and the reason for it is position-wise because the processed attention output is the attention output of a certain position i.

B. Decoder Module

The structure of decoder module is similar to encoder, and the decoder module is also composed of $N = 6$ identical layers, but the layer here contains three sub-layers, including a self-attention layer, an encoder-decoder attention layer and a fully connected layer. The first two sub-layers are based on a multi-head attention layer, one special point here is the mask, which can ensure that the future information will not be contacted when predicting the word of the $i-\mathbf{th}$ position considering that the output during training stage is ground truth. In addition, the input can be calculated in parallel during the encoding process, but in the decoding process the output is decoded one by one like RNN, because the input of the previous position is treated as the attention query.

C. Positional Encoding

The Transformer model not only makes advancements in the main encoder and decoder modules, but also in the data preprocessing part. Transformer architecture abandons the traditional RNN structure whose most prominent advantage is the representation of data in time steps. To achieve an alternative for understanding the order of input words, the approach taken by [5] is to add an additional positional encoding vector for each input word after embedding. These vectors follow a specific pattern learned by the model, which helps determine the position of each word, or the distance between different words in the sequence. Reference [5] uses sine and cosine functions of different frequencies to construct the position information:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \quad (6)$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \quad (7)$$

Where pos is the position and i is the dimension, $PE_{(pos+k)}$ can be represented as a linear function of $PE_{(pos+k)}$ for any fixed offset k.

IV. Proposed Method

A. Overview

The purpose of this research is to propose a better solution which aims to solve the natural language generation task of headline generation from end to end. The headline generation

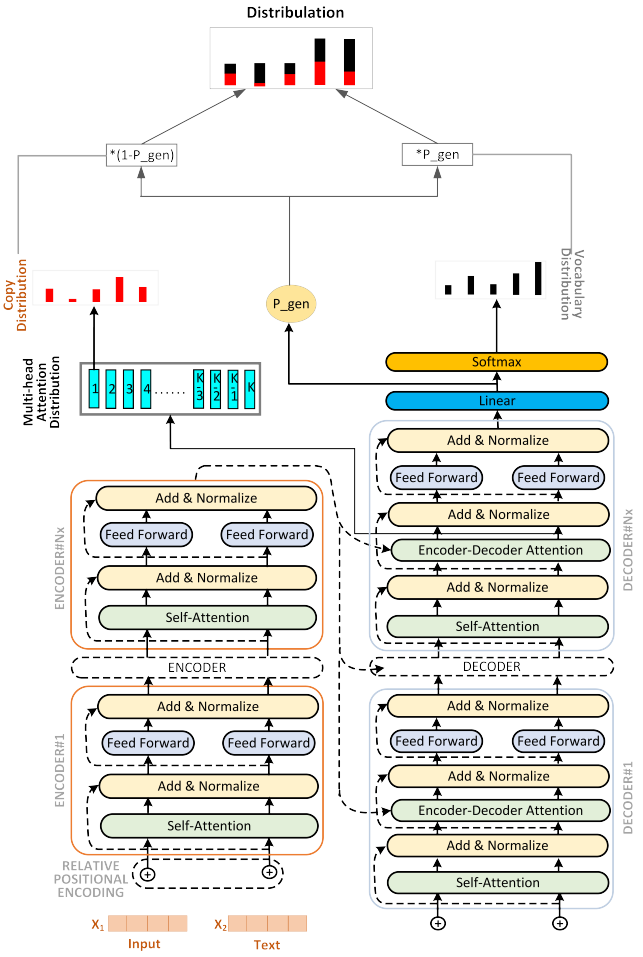

Fig. 1. The framework of our model for headline generation.

can be regarded as a special genre of abstract generation task because the target of generation is a more concise, fluent and informative sentence. Previous related studies are based on Seq2Seq framework, LSTM [19], GRU [20] and their variant units are often employed to model the text now that text is usually a variable-length sequence. Limited by the learning capability of the above neural network models, for document-level sequence, the long-term dependencies captured by these models are undoubtedly limited. Therefore, for the document summarization task which requires document-level input, its ultra-long-term dependencies keep a huge challenge.

Inspired by relative researches and considering that previous studies almost avoid addressing the summarization task of headline generation from the perspective of understanding the entire news document, we propose an end-to-end method that combines the Transformer-XL model which is capable of learning ultra-long-term context dependencies for extracting features from whole news at the document-level with byte-pair encoding technique , and utilize the copy mechanism commonly used when solving OOV problems to ensure the fidelity of those substantive nouns and proper nouns. In addition, the coverage mechanism is added to the model during

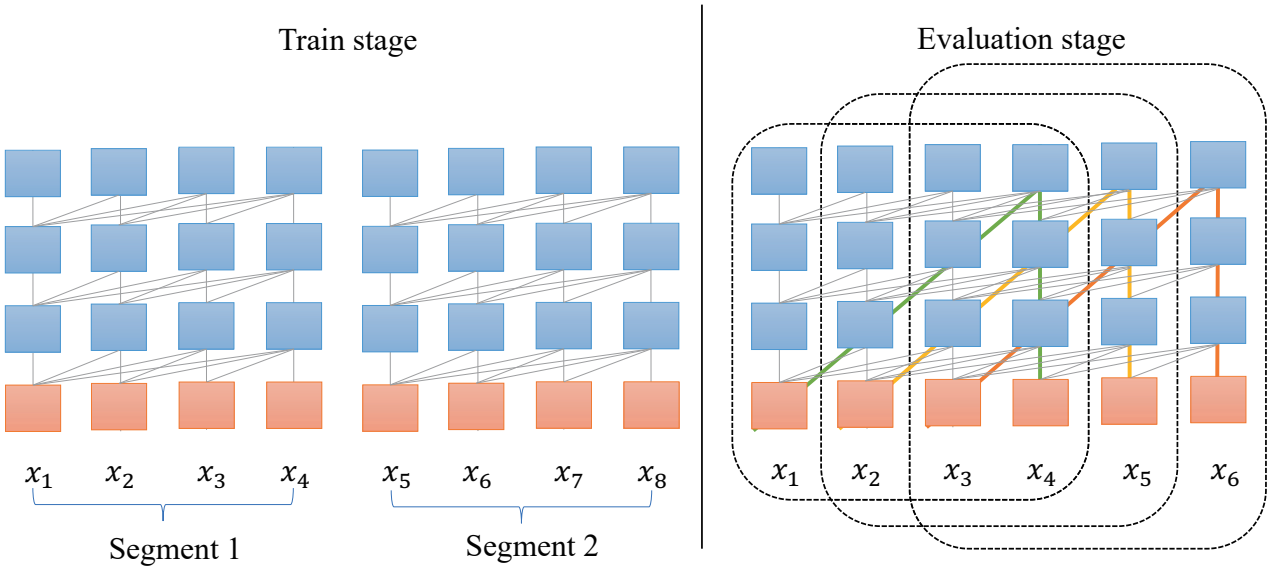

Fig. 2. Process of standard Transformer model with a segment length 4.

inference stage to avoid the problem of generating duplicate fragments.

B. Our Model Architecture

In this paper we propose a model Transformer(XL)-CC inspired by [8], which attempts to improve the performence on model design, in order to solve the problem of limited long-term dependencies learning caused by fixed-length input constraints of Transformer. Our proposed model for headline generation is shown in Fig. 1. Inspired by the ideas of Transformer-XL, we also introduce segment-level recurrence mechanism in the traditional Transformer model to solve the problem of context fragmentation for achieving ultra-long-term dependencies, and employ the novel relative position encoding proposed by [8] at the same time. We construct $N = 4$ identical layers for both the encoder and decoder module of the proposed model. Afterwards, considering that some essential substantive nouns and proper nouns in news articles are usually indispensable information for the headline, we add copy mechanism on the top of our model paired with coverage mechanism, where the coverage mechanism is employed to avoid generating duplicate fragments.

1) Segment-level Recurrence Mechanism: Standard Transformer model fixes the length of input sequences, and there is no connection between different sequences(segments), the training and evaluation stages of standard Transformer are shown in Fig. 2, the entire corpus is divided into shorter segments of manageable size for training, ignoring all contextual information from previous segments during training phase. Therefore, the ability of such a Transformer model to capture long-term dependencies is limited, and it causes the tricky problem of context fragmentation. We divide the corpus into segments of equal fixed length in advance, and each segment is separately invested in computing self-attention during training stage like [10], but the hidden state of each layer's output is stored in memory as an additional input when training the next segment, which represents the previous context information, as shown in Fig. 3. In this way the model can capture longer-term dependencies.

Mathematically, suppose that the two adjacent segments are $\mathbf{s}_\tau = [x_{\tau,1}, \ldots, x_{\tau,L}]$ and $\mathbf{s}_{\tau+1} = [x_{\tau+1,1}, \ldots, x_{\tau+1,L}]$ where L is the length of segment and the hidden state of $\mathbf{s}_\tau$ in the $n-\mathbf{th}$ layer is $\mathbf{h}_\tau^n \in \mathbb{R}^{L\times d}$, where $\mathbf{d}$ is the hidden dimension. Then, the hidden state of the $n-\mathbf{th}$ layer for segment $\mathbf{s}_{\tau+1}$ is generated as follows:

$$\widetilde{\mathbf{h}}_{\tau+1}^{n-1} = \left[\mathrm{SG}\left(\mathbf{h}_\tau^{n-1}\right) \circ \mathbf{h}_{\tau+1}^{n-1}\right] \tag{8}$$

$$\mathbf{q}_{\tau+1}^n, \mathbf{k}_{\tau+1}^n, \mathbf{v}_{\tau+1}^n = \mathbf{h}_{\tau+1}^{n-1}\mathbf{W}_q^\top, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_k^\top, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_v^\top \tag{9}$$

$$\mathbf{h}_{\tau+1}^n = \text{Transformer-Layer}\left(\mathbf{q}_{\tau+1}^n, \mathbf{k}_{\tau+1}^n, \mathbf{v}_{\tau+1}^n\right) \tag{10}$$

Where the function $\mathbf{SG}(\cdot)$ is the abbreviation of stop-gradient, which determines whether to cut off gradient update, $[\mathbf{h}_u \circ \mathbf{h}_v]$ represents the concatenation of two hidden sequences along the length dimension, and $\mathbf{W}$. is the model parameters. After the above adjustments, the critical differences from the standard Transformer are that the key $\mathbf{k}_{\tau+1}^n$ and value $\mathbf{v}_{\tau+1}^n$ are conditioned on the extended context $\widetilde{\mathbf{h}}_{\tau+1}^{n-1}$ and $\widetilde{\mathbf{h}}_\tau^{n-1}$ carries the information from the previous segment.

2) Relative Positional Encoding: With the addition of the recurrence mechanism, the absolute position encoding inherited from [5] loses its original role because Transformer abandons the autoregressive computing method of RNN. We adopt the solution of Transformer-XL to eliminate the position encoding during the model input phase, and instead encode Query and Key vectors before each attention layer.

Assuming that the absolute position encoding matrix is $\mathbf{U} \in \mathbb{R}^{L_{max}\times d}$, where $\mathbf{U}_i$ represents the encoding vector of the $i-\mathbf{th}$ absolute position, L_{max} prescribes the maximum

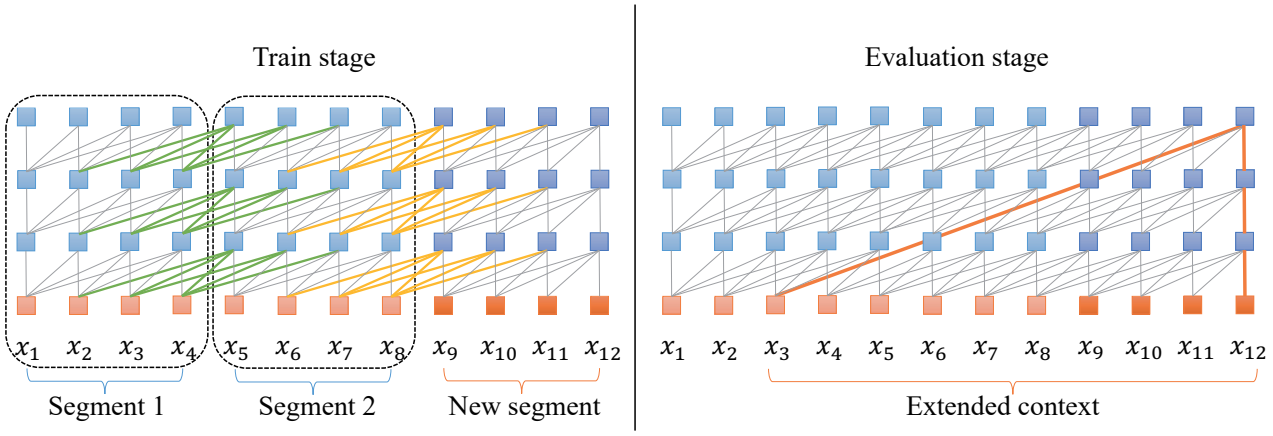

Fig. 3. Process of modified strategy for our model with a segment length 4.

possible length of the input text, and the input pre-trained representation vector (word vector or character vector) is $\mathbf{E}$, then the attention similarity of Query $\mathbf{q}_i$ and Key $\mathbf{k}_j$ in the standard Transformer is denoted as:

$$\begin{aligned} A_{i,j}^{abs} =& E_{x_i}^T W_q^T W_k E_{x_j} + E_{x_i}^T W_q^T W_k U_j \\ &+ U_i^T W_q^T W_k E_{x_j} + U_i^T W_q^T W_k U_j \end{aligned} \tag{11}$$

After adjustment, the four terms of formula (12) respectively represent content-based addressing, content-dependent positional bias, global content bias, and global positional bias:

$$\begin{aligned} A_{i,j}^{rel} = & E_{x_i}^T W_q^T W_{k,E} E_{x_j} + E_{x_i}^T W_q^T W_{k,R} R_{i-j} \\ &+ \mathbf{u}^T W_{k,E} E_{x_j} + \mathbf{v}^T W_{k,R} R_{i-j} \end{aligned} \tag{12}$$

Relevant adjustments are organized into the following points,

- $R \in \mathbb{R}^{L_{\max} \times d}$ is a sinusoid encoding matrix corresponding to U.
- $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ are trainable parameters that limit and dominate the attentive bias caused by changes in Query position respectively.
- $W_{k,E}$ and $W_{k,R}$ are separated from W_k, which respectively generates the content-based key vectors and location-based key vectors.

3) Copy and Coverage Mechanism: For low-frequency words, just adopting the generative method is actually quite unreliable. Therefore, we employ the method of CopyNet [11] to copy the words from the original text through the probability distribution of attention when generating some proper nouns of the headline:

$$\begin{aligned} p\left(y_t | s_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}\right) =& p\left(y_t, c | s_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}\right) \\ &+ p\left(y_t, g | s_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}\right) \end{aligned} \tag{13}$$

Where $\mathbf{M}$ is the set of input hidden layer states, $\mathbf{c}_t$ is the attention score, s_t is the hidden state of the output, g stands for generate mode and c denotes copy mode. Generation or copy is selected based on the maximum probability. Such a method has basically solved the OOV problem in our experiments.

As for the coverage mechanism [12] added in the inference phase, the overall structure of the model is almost unchanged, and only the calculation method of attention needs to be adjusted:

$$e_i^t = v^T \tanh\left(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}}\right) \tag{14}$$

$$a^t = \text{softmax}\left(e^t\right) \tag{15}$$

where v, W_h, W_s and b_{attn} are learnable parameters, w_c is also a learnable parameter vector with the same length as v, c_t is not a semantic vector but a newly defined parameter: $c^t = \sum_{t'=0}^{t-1} a^{t'}$, it is a unnormalized distribution over the words in the source document, and it indicates the degree of coverage those words have received from the attention mechanism. The purpose of adding this parameter is to transmit the information of the previously generated words to attention calculation. If these words have been generated before, they should be suppressed later and suppression is achieved by adding penalty term to loss function:

$$\text{loss}_t = -\log P\left(w_t^*\right) + \lambda \sum_i \min\left(a_i^t, c_i^t\right) \tag{16}$$

The above formula denotes that during training, the loss of timestep $\mathbf{t}$ is updated to the negative log likelihood of the target word w_t^* at that timestep plus the coverage loss reweighted by some hyperparameter λ .

C. Byte-pair Encoding

In previous headline generation studies, the preprocessing of corpus adopts a classic and relatively simple approach. Usually, a vocabulary is generated from corpus first and only words with frequency greater than a certain threshold will be thrown into the vocabulary in order to avoid the size of vocabulary being too large, and all the remaining words will be uniformly encoded as <*UNK*>, this method is seriously affected by low-frequency words. We utilize byte-pair encoding for the machine translation task proposed by [9] instead, which is essentially a data compression technique. Byte-pair encoding combine the advantages of word-level and char-level model and allow us to represent more words with fixed vocabulary, including those unseen in training phase.

V. EXPERIMENTS

A. Datasets

We conduct our experiments on Chinese and English datasets, respectively. The English dataset we adopt is the New York Times Annotated Corpus (NYT)[1] from the Linguistic Data Consortium and the Chinese dataset is the public dataset Chinese LSCC news corpus[2].

a) NYT dataset.: This dataset contains 1.8 million news articles written and published by the New York Times between 1987 and 2007, which is widely used in summarization generation and headline generation tasks. We filter out news articles with titles shorter than 3 words or longer than 15 words, and articles with text less than 20 words or more than 2000 words. After filtering, we obtain 1.4 million news articles, the average length of the headline is 8.2 words, and the mean text length is 752.6 words. We randomly retain 20000 articles as the test set. Furthermore, we train the BPE tokenization on NYT dataset and it is tokenized with a vocabulary size of nearly 40000 and we limit the text length within 2000 BPE tokens.

b) LSCC news dataset.: The Chinese LSCC news dataset contains 2.5 million news articles, which covers 63 thousand media, including titles, keywords, descriptions, and text. This dataset has been divided into three part. The training set is consist of 2.43 million news and other 77 thousand news of validation set is used as the test set in our experiments. We exclude articles with less than 100 words and more than 2,500 words, titles with less than 3 words and more than 30 words, and finally we leave 1.8 million articles for training and nearly 70 thousand articles as the test set. In addition, we use jieba word segmentation[3] instead of BPE for Chinese vocabulary.

B. Evaluation and Results

We compare our model called **Transformer(XL)-CC** with the following competitive models, which take different inputs or model frameworks from our approach:

- **First sentence** is simply to take the first sentence of the news as its headline. It is a simple but straightforward baseline for headline generation.
- **Moses+** [21] is usually employed in phrase-based statistical machine translation tasks. We set an infinite deformation limit in order to improve the baseline for headline generation task.
- **En-decoder-LSTM** [22] is a typical model for seq2seq task which takes the LSTM units for both the encoder and decoder module, we follow [22] to use the lead sentences only as input to generate the headline.
- **ABS** [6] is a sentence summarization model based on the attention bag-of-words encoder.
- **Summ-hieratt** [7] is a coarse-to-fine approach, which utilizes several statistical summarization approaches to obtain summary sentences instead of full text or the lead sentences as input and unite hierarchical attention.

[1]https://catalog.ldc.upenn.edu/LDC2008T19

[2]https://github.com/brightmart/nlp_chinese_corpus

[3]https://github.com/fxsjy/jieba

TABLE I
ROUGH-1,2,L SCORES ON NYT AND LSCC NEWS DATASETS.

Model	Rough-1	Rough-2	Rough-L
New York Times dataset			
First sentence	19.43	5.35	17.72
Moses+	18.39	7.96	15.31
En-decoder-LSTM	23.81	11.09	22.13
ABS	25.82	7.03	23.07
Summ-hieratt	29.60	8.17	26.05
Transformer	26.32	12.58	25.86
Transformer(XL)-CC	**30.73**	**13.46**	**27.19**
LSCC news dataset			
First sentence	34.39	18.72	30.28
Moses+	21.53	12.25	27.18
ABS	33.76	21.39	29.26
En-decoder-LSTM	34.92	**24.50**	30.35
Transformer	36.57	22.83	34.61
Transformer(XL)-CC	**39.86**	23.97	**36.02**

- **Transformer** [5] is a model based entirely on the attention mechanism, which abandons traditional RNN and CNN, and is widely used in seq2seq tasks such as machine translation and automatic summarization.

We implement experiments by using the same hyper-parameters set in our models for both the NYT and Chinese LSCC news datasets, i.e. $N = 4$ identical layers are constructed for both the encoder and decoder module with 8 heads of attention. Furthermore, we add a dropout to the output of each sub-layer with $p = 0.2$ before it is added to the sub-layer input and normalized. Furthermore, we adopt the Adam optimizer to train our models using a scaled learning rate until convergence, just like the standard Transformer with the number of warmup steps equal to 4000 in both cases and $\beta_1 = 0.9$, $\beta_2 = 0.98$.

As for evaluation metrics, headline generation is usually regarded as a special task of automatic summarization, thus we take the widely used ROUGH [23] as our automatic evaluation metric whose fundament thought is to take the n-tuple co-occurrence statistics of the headline to be reviewed and the reference headline as the evaluation basis, and then grade through a series of standards considering several dimensions. We report recall results on Rough-1, Rough-2 and Rough-L here.

As shown in Table. I, it is not a sane choice to simply take the first sentence as the headline of the news directly judging from the experimental results both on the NYT dataset and Chinese LSCC news dataset, and the results of *Moses+* indicate that using traditional statistical methods for the title generation task will get poor results. For the NYT dataset, the experimental results of *En-decoder-LSTM* show that the encoder-decoder model combined with LSTM units is effective but not outstanding, because the information fed to the input is limited. Although the *summ-hieratt* model united with hierarchical attention whose input information is more abundant improve the performance, this method will be significantly affected by the summarization generation

OT: it is against the law to build a bomb, but teachers are not barred from showing students how to do it. the senate education committee passed a bill yesterday that would outlaw bomb-building instruction. the measure was spurred by an incident last year in which a vernon high school teacher allowed his students to watch a video showing fellow classmates building and detonating pipe bombs, the a.p. reported. dennis o'leary, the sussex county prosecutor, tried to charge the teacher, but discovered that what he did was legal.
OH: new jersey daily briefing; banning bomb-making guide
GH: prohibition against making bombs

OT: warning that the unstable situation in iraq could fall into "anarchy," france called today for the creation of an international military force and a provisional government there under united nations authority. the proposal, made by foreign minister dominique de villepin in a speech to the annual gathering of france's ambassadors, comes as the united states, having failed to persuade many key allies to send troops and money to iraq, has begun to discuss the possibility of accepting an international military force under a united nations mandate. … "the united nations system is not adapted to deal with the new threats, like international terrorism."
OH: france calls for international force in iraq under the u.n..
GH: france call for military force under united nations authority.

OT: 安徽新闻网讯：10月25日，韩国艺总江原道文化代表团在安徽省文化厅外事处处长周化东，宿州市文广新局局机关党委书记张志翔，埇桥区副区长扶元广及市、区文化部门负责人的陪同下到符离镇沈圩村访问。 据了解，韩国艺总江原道文化代表团是本着"魅力文化、传承共享"理念来进行这次访问的。 韩国艺总江原道文化代表团在沈圩村先后访问了沈圩村农民文化乐园、农村生活驿站、陈毅淮海战役纪念馆、村民俗馆、沈圩古井等地，参访结束后，韩国客人还邀请沈圩村村民合影留念。 据陪同访问的宿州市文广新局有关负责同志介绍，韩国艺总江原道文化代表团还将赴萧县、灵璧等地参访有文化特色的景点。
OH: 韩国艺总江原道文化代表团到符离镇沈圩村访问
GH: 韩国艺总江原道文化代表团在沈圩村访问

Fig. 4. Headlines generated by our model on NYT test set and Chinese LSCC test set. **OT** is original text, **OH** is original headline and **GH** is geneated headline by our model.

TABLE II
HUMAN EVALUATION ON NYT AND LSCC NEWS DATASETS.

Dataset	Human(%)	Tie(%)	Machine(%)
NYT news	56.1	24.7	19.2
Chinese LSCC news	59.3	22.9	17.8

techniques. The results of *Transformer* show that the model entirely based on attention mechanism has great potential for headline generation, and our method inspired by Transform-XL combine copy and coverage mechanism achieve significant better results. For the Chinese LSCC news dataset, since there is no previous *sota* on Chinese headline generation task, we demonstrate the experimental results of our model on this dataset and the results of related competitive methods.

Considering the singularity and inflexibility of automatic evaluation metric, and numerous headlines generated by our model that seem readable are scored low by machine, we conduct additional human evaluation experiments both on the results of NYT and Chinese LSCC news datasets. We respectively invite thirty well-educated annotators to evaluate the results of 1000 randomly sampled news on each dataset. The experimental results are shown in Table. II, *Human* represents that original headlines are preferred, *Machine* means generated headlines win, *Tie* means no preference. It can be seen that our model has almost achieved competitive level as human from the tendency of experimenters' preference, 43.9% for NYT dataset and 40.7% for Chinese LSCC news dataset (Tie and Machine).

C. Case Study

We select two samples on the NYT test set and one sample on the Chinese LSCC test set, as shown in Fig. 4. The first example demonstrates that the headline generated by our model accurately refines the main content of the news. Compared with the original headline, our headline only loses the media source of the news which is also unseen in the body of news and will generate novel words "prohibition" based on understanding the whole news. From the second example, the headline generated by our model is almost the same as the headline written by humans except for the third-person grammatical error and the unintelligent generation of the UN abbreviation, it can be regarded as an excellent title. As for Chinese headline generation, the generated headline is just missing the township name as shown in the third example, this also confirms that our model is quite effective.

Although the experimental results and examples show that our model has made significant improvements on the NYT data set compared with previous studies, and has also achieved pretty good results on the Chinese LSCC news dataset, we still get many poor results on test datasets. After all, the titles written by humans are not only highly refined, but also contain the original insights of the authors about the entire news. Therefore, although our method utilizes a more powerful encoding model to learn context dependencies, it cannot be regarded as a substitute of headlines written by human. From another perspective, our model attempts to generate headlines based on understanding the full text more in line with human behavior at the same time.

VI. CONCLUSION

In this paper, we have proposed a novel and effective model Transformer(XL)-CC to tackle the headline generation task from end to end, which utilizes the ultra-long-term dependencies capture capability of Transformer-XL to solve the context fragmentation problem generated by standard Transformer. We use the entire news as input instead of the lead sentences or summary sentences on the basis of understanding the whole document. We observe that previous studies using the first

sentence or summary sentence of the news to summarize articles have respective limitations, neither accurately refine the main content of the news or rely heavily on the technique of summary sentences generation. Experimental results prove that our method significantly improves the performance on headline generation, both for English news and Chinese news.

ACKNOWLEDGMENTS

This research work has been funded by the National Natural Science Foundation of China (No.61772337) and the Eastday-SJTU Artificial Intelligence Media Joint Lab.

REFERENCES

[1] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325, Hong Kong, October 2000. Association for Computational Linguistics.

[2] Chih-Ming Chen. Intelligent location-based mobile news service system with automatic news summarization. *Expert Systems with Applications*, 37(9):6651–6662, 2010.

[3] Fei Liu and Yang Liu. From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 261–264, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[6] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[7] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, pages 4109–4115, 2017.

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[10] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166, 2019.

[11] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[12] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[13] Bonnie Dorr, David Zajic, and Richard Schwartz. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 1–8. Association for Computational Linguistics, 2003.

[14] Stephen Wan, Mark Dras, Cécile Paris, and Robert Dale. Using thematic information in statistical headline generation. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 11–20. Association for Computational Linguistics, 2003.

[15] Songhua Xu, Shaohui Yang, and Francis Lau. Keyword extraction and headline generation using novel word features. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[16] Enrique Alfonseca, Daniele Pighin, and Guillermo Garrido. Heady: News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1243–1253, 2013.

[17] Rui Sun, Yue Zhang, Meishan Zhang, and Donghong Ji. Event-driven headline generation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, 2015.

[18] Sumit Chopra, Michael Auli, and Alexander M Rush. 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*, pages 93–98, 2016.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[21] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180, 2007.

[22] Yuko Hayashi and Hidekazu Yanagimoto. Headline generation with recurrent neural network. In *New Trends in E-service and Smart Computing*, pages 81–96. Springer, 2018.

[23] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.