Applied Intelligence (2022) 52:9650–9663
https://doi.org/10.1007/s10489-021-02607-9

Abstractive document summarization via multi-template decoding

Yuxin Huang[1,2] · Zhengtao Yu[1,2] · Junjun Guo[1,2] · Yan Xiang[1,2] · Zhiqiang Yu[1,2] · Yantuan Xian[1,2]

Accepted: 9 June 2021 / Published online: 8 January 2022

Abstract

Most previous abstractive summarization models generate the summary in a left-to-right manner without making the most use of *target-side global information*. Recently, many researchers seek to alleviate this issue by retrieving target-side templates from large-scale training corpus, yet have limitations in template quality. To overcome the problem of template selection bias, one promising direction is to get better target-side global information from multiple high-quality templates. Hence, this paper extends the encoder-decoder framework by introducing a *multi-template decoding mechanism*, which can utilize multiple templates retrieved from the training corpus based on the semantic distance. In addition, we introduce a *multi-granular attention mechanism* by simultaneously taking into account the importance of words in templates and the importance of different templates. Extensive experiment results on CNN/Daily mail and English Gigaword show that our proposed model significantly outperforms several state-of-the-art abstractive and extractive baseline models.

Keywords Abstractive document summarization · Multiple templates · Target-side global information · Multi-granular attention

1 Introduction

Recently, sequence-to-sequence (seq2seq) models have achieved impressed improvement in the abstractive text summarization task [17, 45]. Generally, the seq2seq model falls into an encoding-decoding paradigm, which first encodes the source document $\boldsymbol{x}$ to high-level abstract representation, then generates summary $\boldsymbol{y}$ word by word from left to right. Intuitively, the t-th target word y_t is generated merely conditioning on the source document $\boldsymbol{x}$ and previously generated words $y_{<t}$, rather than considering the global target-side information. Recently, different methods are proposed to overcome this limitation by using a pseudo sequence $\mathbf{y}'$ as the target-side global information. A few recent approaches seek to obtain $\mathbf{y}'$ by retrieving a soft template from the large-scale training corpus, which also is known as template-based models. The motivation behind these methods is that *documents with similar semantics have similar summaries*, and the retrieved soft template (the gold summary of the retrieved document) can be regarded as a high-quality pseudo summary of the current document, which can provide complete target-side global information for the decoding process. In this direction, Cao et al. [2] presented a Re^{3}Sum model, which first retrieved several candidate templates by Lucene[1] (a search engine based on word matching), then re-rank module was used to obtain the 1-best template as the soft template, finally, the summary was generated according to the source document and the soft template. More recently, Wang et al. [27] extended the Re^{3}Sum model by introducing a bi-directional selective layer to overcome the problem of insufficient use of the template and achieved state-of-the-art performance on English Gigaword corpus [18]. Although template-based approaches have achieved competitive performance in sentence-level summarization task, there are still two weaknesses in document-level summarization scenario: 1) generally, the 1-best retrieved

✉ Zhengtao Yu
ztyu@hotmail.com

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China

[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China

[1] https://lucene.apache.org/

template is far from satisfactory for the decoder, namely *template selection bias problem*; 2) the retrieval methods based on word matching can not provide suitable template for document-level summarization task. Therefore, it is a promising direction to redesign the process of template-based summarization method to adapt the long document and learn target-side global information from multiple templates instead of 1-best template.

In this paper, we propose a novel framework to leverage the multiple retrieved templates as target-side global information. Specifically, the pre-trained language model such as BERT [3] is firstly utilized to produce high-level semantic representation for each source document, then several templates are retrieved according to semantic distance from the training corpus. Finally, a multi-granular attention mechanism is introduced into the decoding process, which not only takes advantage of the *word-level attention* to learn the importance weights of each word in templates, but also considers the importance of different template at *sentence level*. Briefly, our approach searches multi-templates from training corpus by estimating the semantic distance of the samples, which can not only ensure the similarity at word-level (e.g. named entity) but also reflect the semantic similarity. Besides, using multi templates can alleviate the problem of template selection bias caused by 1-best template. Extensive experiments are conducted on CNN/Daily Mail dataset [17, 22] and English Gigaword dataset [18]. Experimental results show that the performance of our proposed model achieves significant increment compared with several strong baseline models.

The major contributions of this work are summarized as follows:

- We propose a novel abstractive summarization framework, which efficiently learns target-side global information from multiple templates by introducing a multi-granular attention mechanism.
- We introduce a semantic-based retrieval approach to generate templates.
- Experimental results on CNN/Daily Mail and English Gigaword dataset demonstrate that the proposed model achieves significant improvements compared with extractive and abstractive baseline systems.

2 Related works

Our work is mainly to make use of the target-side information in the seq2seq framework, the previous works in this direction are mainly in the line of *generation-based models*, *memory-augmented models* and *retrieval-based models*. We introduce them in turn below.

2.1 Generation-based models

A few recent approaches seek to obtain target-side information by introducing a bi-directional decoder into the sequence-to-sequence framework. Xia et al. [30] developed a second-pass decoding paradigm known as deliberation network, which first performed first-pass decoder to generate a raw sentence and then polished it by introducing a second-pass decoder. Similar to the deliberation network, Zhang et al. [40] presented an asynchronous bidirectional decoding network by utilizing a backward decoder to generate the target-side sequence from right to left, then the forward decoding generated the output simultaneously constrained by the source-side and target-side information. Zhou et al. [43] further proposed a synchronous bidirectional network to predict the output using left-to-right and right-to-left decoding simultaneously and interactively. Recently, in order to improve the efficiency and quality of bi-directional decoding architecture, Wang et al. [28] introduced a soft prototype sequence into the encoder-decoder framework to leverage the target-side global information. Especially, the soft prototype sequence is generated by mapping each word in the input sequence into a distribution over the target vocabulary. In other words, each word of the source input can be regarded as the weighted average of target words. Although these approaches achieved competitive performance in different sequence generation tasks, yet limitations in the quality of global information generated by two-pass decoding.

2.2 Memory-augmented models

Memory-augmented models aim to utilize global information via an additional memory network, which has been proved useful for question answering task [15] and machine reading comprehension task [31]. Tu et al. [25] proposed a memory-augmented sequence-to-sequence model for neural machine translation task (NMT), in which a key-value memory network was utilized to store the global information specific to the longer translation history. Xia et al. [29] introduced a graph-based memory module to replace the key-value memory network, which packed redundant global information into a compact graph. An alternative approach to obtain memory is based on retrieval. Gu et al. [8] proposed a search engine guided NMT model with key-value memory, in which the translation memory was generated by a set sentence pair retrieved from the training corpus. In analogy to [8], Zhang et al. [39] also retrieved sentence pairs from the training corpus, but only collected n-grams both in the retrieved target sentence and source sentence to generate translation pieces, which were used as extra rewards to update the generation probabilities in the testing stage. Although, these approaches are different in memory

generation and utilization, share the same intrinsic idea, which is to introduce a memory cache to the encoder-decoder framework. In this paradigm, the model needs to design complex strategies for memory generation, updating, and matching, and also more parameters and high computation will be introduced.

2.3 Retrieval-based models

The starting point of retrieval-based approaches is to design informative templates manually or automatically. Zhou et al. [42] proposed a hard template-based summarization method, in which the templates were manually created by domain experts based on the linguistic features such as POS (Part-of-Speech tags), NER (Named Entities Recognition), etc. However, it is unrealistic to create all the templates manually, especially for large-scale corpus. Hence, automatic retrieval of similar samples from large-scale corpus as soft templates has become the research highlights in recent years. Pandey et al. [20] presented an exemplar encoder-decoder network (EED) to solve the closed domain neural conversation generation task. In detail, closely related context-response pairs were retrieved and encoded as the exemplar context to guide the generation of response. For the summarization task, Cao et al. [2] proposed a retrieve, rerank, and rewriting framework, aka, Re^{3}Sum, which used existing closely related summaries as soft templates to enhance the seq2seq model. This is also the first attempt to utilize the retrieved soft template for the abstractive text summarization task. Under this framework, Wang et al. [27] developed a bi-directional selective layer to mutually select key information from the source sequence and corresponding template. However, these approaches tend to obtain 1-best soft template from short sentence summarization corpus via word matching method, yet have limitations on the diversity and adaptability of the template.

3 Materials and methods

For abstractive document summarization, the encoder reads the sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and yields the context representation $\mathbf{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$, from which the decoder generates the summary sequence $\boldsymbol{y} = (y_1, y_2, \ldots y_M)$ word by word. The encoder and decoder can be realized by recurrence neural network (RNN) [24], convolution neural network (CNN) [6] or self-attention [26].

In this paper, we introduce a multi-template decoding mechanism into the encoder-decoder framework to take advantage of the target-side global information. Concretely, given the source document $\boldsymbol{x}$, the semantic retrieval module first searches several high-quality templates $\mathcal{D} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_D\}$ from training set based on semantic distance. Then the multi-template encoder-decoder model is designed to jointly model the templates and source document.

3.1 Semantic retrieval module

The start point of this module is that the source document with similar semantic representation should hold a similar summary. Therefore, we first introduce a widely-used pre-trained language model BERT [3] to map the source document and corresponding summary into the high-dimensional representations. Then each source document is considered as the query to retrieve several similar documents from the training set based on semantic distance, the target summaries of which are then be adopted as the soft templates for the current query. In this work, we use L2 distance as semantic similarity evaluation metrics. Note that, due to the time complexity of similarity comparison algorithm on large-scale dense vectors is extraordinarily high, we leverage FAISS tool[2] to improve the retrieval efficiency. The detailed settings for BERT and FAISS will be introduced in Section 3.3.2.

3.2 Multi-Template encoder-decoder model

As shown in Fig. 1, the multi-template encoder-decoder model consists of three components, which are the conventional encoder-decoder model, the multi-template encoder, and the multi-granular attention mechanism.

3.2.1 Conventional encoder-decoder model

The conventional encoder-decoder model includes encoder, decoder and decoder-to-source document attention (D2S Attention). Our model is built based on Transformer [26], which achieved state-of-the-art performance in many text generation task such as text summarization [12, 41], neural machine translation [4, 5] etc. The core component of the Transformer framework is a multi-head attention mechanism, which enables the model to jointly attend to different representation subspaces at different positions. To this end, the multi-head attention operates on queries Q, keys K, and values V. Specifically, the multi-head attention first yields k-th distinct representations of (Q_k, K_k, V_k) through projecting the hidden state matrix into separate subspace as shown in (1). Then, the Concat operation is performed to obtain the multi-head attention representation.

$$\text{head}_k = \text{Attention}\left(Q\mathbf{W}_k^Q, K\mathbf{W}_k^K, V\mathbf{W}_k^V\right) \tag{1}$$

$$\text{MHAtt}\,(Q, K, V) = \text{Concat}\,(\text{head}_1, \ldots, \text{head}_k) \tag{2}$$

[2] https://github.com/facebookresearch/faiss

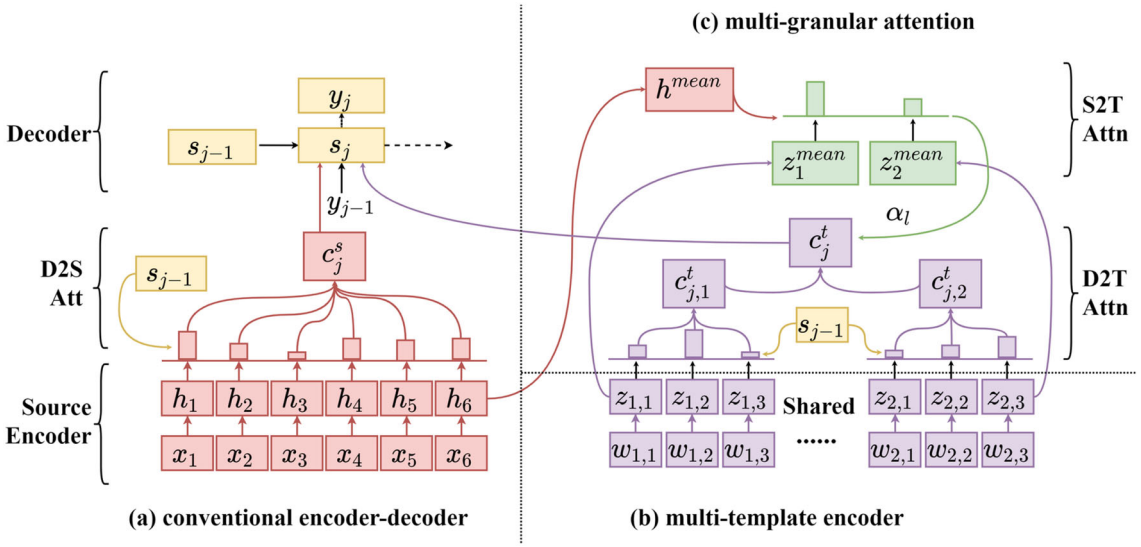


Fig. 1 Overview of our proposed model. The model is divided into three components: 1) the conventional encoder-decoder model; 2) the multi-template encoder to yield the high-level representation of each template; 3) the multi-granular attention mechanism including sentence-level S2T attention and word-level D2T attention

in which, $head_k$ denotes the k-th attention head, Concat is the concatenate operation, and $\mathbf{W}_k^Q$, $\mathbf{W}_k^K$, and $\mathbf{W}_k^V$ are parameter of projection matrices. Note that Attention is a scaled dot-product operation as shown in (3):

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where d_k denotes the dimension of the key.

Typically, in order to get the representation of input sequence $\boldsymbol{x}$, the word embedding matrices is used to map it into continuous space:

$$\begin{aligned}&\mathbf{E}_i = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N)\\ &\text{where}\quad \mathbf{e}_n = e\,(x_n) + \mathbf{p}_n\end{aligned} \tag{4}$$

where $e\,(x_n)$ and $\mathbf{p}_n$ are the word and positional embedding of x_n, respectively. Note that the positional embeddings are calculated by sin-cos functions of different frequencies. The positional embedding $\mathbf{p}_n$ for the n-th token in the sequence is calculated as:

$$\mathbf{p}_n[i] = \sin\left(n/10000^{2i/d}\right) \tag{5}$$

$$\mathbf{p}_n[2i+1] = \cos\left(n/10000^{2i/d}\right) \tag{6}$$

where $\mathbf{p}_n[i]$ denotes the i-th dimension of the embedding vector. d is the dimension of positional embedding. Due to the dimension of the positional embedding corresponds to a sinusoid, any fixed offset o, $\mathbf{p}_{n+o}$ can be denoted as a linear function of $\mathbf{p}_n$, which enables the model to distinguish the relative positions of input tokens.

Then a transformer layer [26] is applied to map $\mathbf{E}$ into a high level hidden representations $\mathbf{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$, which composed of two sub-layers:

$$\mathbf{E}' = \text{LayerNorm}\,(\mathbf{E} + \text{MHAtt}(\mathbf{E})) \tag{7}$$

$$\mathbf{H} = \text{LayerNorm}\,\left(\mathbf{E}' + \text{FFN}\left(\mathbf{E}'\right)\right) \tag{8}$$

where LayerNorm is layer normalization function introduced by [1]; MHAtt denotes the multi-head attention mechanism proposed by [26], as shown in (2); FNN is a feed-forward network with ReLU as activation function.

Finally, at each decoding step j, a multi-head attention mechanism from decoder to source document (D2S) is adopted to calculate the attention context vector as:

$$\boldsymbol{c}_j^s = \text{MHAtt}\,\left(Q = \boldsymbol{s}_{j-1}, K = \mathbf{H}, V = \mathbf{H}\right) \tag{9}$$

where s_{j-1} is the $j-1$ step hidden state of the decoder, and K, V pairs come from the hidden state $\mathbf{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N)$.

3.2.2 Multi-Template encoder

For the templates $\mathcal{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_L\}$, where $\boldsymbol{w}_l = \left(w_{l,1}, w_{l,2}, \ldots, w_{l,R}\right)$ is the l-th template, and $w_{l,r}$ denotes the r-th word in l-th template. In analogy to the source document encoder, the templates encoder is another transformer encoder but shared the parameters between different templates. We obtain the hidden states of l-th template $\mathbf{Z}_l = \left(\boldsymbol{z}_{l,1}, \boldsymbol{z}_{l,2}, \ldots, \boldsymbol{z}_{l,R}\right)$ by running a transformer encoder layer as shown in (1)-(3). Note that different templates should be projected in the same semantic

space, hence we share the parameters between different templates.

3.2.3 Multi-Granular attention mechanism

The motivation behind multi-granular attention is to simultaneously take into account the importance of words in templates and the importance of different templates. Specifically, the multi-granular attention mechanism involves the following two steps: source-to-templates attention (S2T) at sentence-level and decoder-to-templates attention (D2S) at word-level. The starting point for S2T attention is an observation that although multiple templates can obtain sufficient target-side global information, they also bring noise into the decoding process. we argue that different templates should be assigned different importance scores. To this end, we calculate the similarity between the source document and templates to yield the weights distribution over different templates. Mean-Pooling operation [9] is first applied to obtain the representation of source document and template, as shown in (10), (11):

$$\boldsymbol{h}^{mean} = \text{Mean-Pooling}\ (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N) \tag{10}$$

$$\boldsymbol{z}_l^{mean} = \text{Mean-Pooling}\ \left(z_{l,1}, z_{l,2}, \ldots, z_{l,R}\right) \tag{11}$$

where $\boldsymbol{h}^{mean}$ denotes the representation of source document and $\boldsymbol{z}_l^{mean}$ indicates the representation of l-th template.

Then $\boldsymbol{h}^{mean}$ is taken as the query to calculate the weight distribution over each template as follows:

$$e_l = \mathbf{v}_a^{\top} \tanh\left(\mathbf{W}_a \left[\boldsymbol{z}_l^{mean}; \boldsymbol{h}^{mean}\right]\right) \tag{12}$$

where $\mathbf{v}_a$ and $\mathbf{W}_a$ are learning parameters, tanh is activation function, [;] denotes the concatenate operation. Then the normalization operation is performed to limit the attention weights to [0, 1]:

$$\alpha_l = \frac{\exp\left(e_l\right)}{\sum_{l=1}^{L} \exp\left(e_l\right)} \tag{13}$$

where α_l denotes the final distribution of l-th template.

On the other hand, D2T attention is responsible to obtain word-level global information at decoding time step j. The intuition behind this mechanism is the fact that since the encoder-decoder model generates summary in a word-by-word manner, the target-side context information is generally stored at the word-level. To this end, at the decoding step j, we treat each template as independent context and perform the multi-head attention operation over each template as follows:

$$\boldsymbol{c}_{j,l}^t = \text{MHAtt}\ \left(Q = \boldsymbol{s}_{j-1}, K = \boldsymbol{z}_l, V = \boldsymbol{z}_l\right) \tag{14}$$

where $\boldsymbol{c}_{j,l}^t$ denotes the attention context representation of l-th template at the decoding step j. Then, the final multi-template context representation is calculated as the weighted sum of each $\boldsymbol{c}_{j,l}^t$, in which the weighted value α_l is considered as the sentence level distribution of each template:

$$\boldsymbol{c}_j^t = \sum_{l=1}^{L} \alpha_l \boldsymbol{c}_{j,l}^t \tag{15}$$

here $\boldsymbol{c}_j^t$ is regarded as the final target-side global information produced by the multiple templates.

At decoding step j, a simple concatenation operation is used to combine the hidden state $\boldsymbol{s}_j$, the D2S context representation $\boldsymbol{c}_j^s$ and the target-side global context $\boldsymbol{c}_j^t$ into a new hidden state $\boldsymbol{o}_j$:

$$\boldsymbol{o}_j = \tanh\left(\mathbf{W}_j \left[\boldsymbol{s}_j; \boldsymbol{c}_j^s; \boldsymbol{c}_j^t\right]\right) \tag{16}$$

where $\mathbf{W}_j$ is the learning parameter, tanh denotes activation function. Finally, $\boldsymbol{o}_j$ is mapped to an output distribution over target vocabulary and then feed through a softmax function to get the final output distribution of y_j:

$$p\left(y_j | y_1, \ldots, y_{j-1}\right) = \text{softmax}\left(\mathbf{W}_p \boldsymbol{o}_j\right) \tag{17}$$

We optimize our model with the negative log-likelihood loss function between the generated summary $\boldsymbol{y}$ and the ground-truth $\hat{\boldsymbol{y}}$:

$$\mathcal{L} = -\sum \log p_\theta\left(\boldsymbol{y} | \boldsymbol{x}, \mathcal{D}\right) \tag{18}$$

where the loss function is equivalent to maximizing the conditional probability of summary $\boldsymbol{y}$ given parameters θ, source document $\boldsymbol{x}$, and the corresponding templates $\mathcal{D}$.

3.3 Experimental setup

3.3.1 Datasets

We experiment on two popular summarization datasets: CNN/Daily mail and English Gigaword. The former is a standard document-level summarization corpus. In this paper, we use the same script as [17, 22] to yield the same non-anonymous version corpus[3], which contains 287,227 training samples, 13,368 validation samples, and 11,490 test samples. English Gigaword is a sentence-level summarization corpus [18], which is generated by pairing the first sentence of the news article and the headline. To obtain comparable experimental results, we use the same preprocessing script[4] to yield the standard training, testing, and validation sets. The statistical information of these two datasets are summarized in Table 1.

[3] https://github.com/bisee/cnn-dailymail

[4] https://github.com/facebook/NAMAS

Table 1 Statistics for CNN/Daily mail (Top) and Engish Gigaword (Bottom) corpus

	Train	Validation	Test
#samples	286,227	13,368	11,490
AvgDocumentLen	751	769	778
AvgSummaryLen	55	61	58
#samples	3.8M	8,000	1,951
AvgDocumentLen	31.35	30.68	31.45
AvgSummaryLen	8.23	8.46	8.27

#samples denotes the number of samples. AvgArticleLen and AvgSummaryLen represent the average length of document and summary, respectively

3.3.2 Training configuration

Our model is built based on Open-NMT [10] framework. In detail, the source document is truncated to 400 tokens and the target summary is limited to 100 tokens. For all experiments below, we employ a 4-layer Transformer with 8 attention heads and 512 hidden dimensions for both encoder and decoder. We train our model on NVIDIA V100 for 20000 iterations with a batch size of 4096 tokens and also limit the vocabulary size of input and target to the 50K most frequent tokens in the training set. To reduce the parameters, we share the embedding matrix between encoder and decoder. The model is optimized using Adam with learning rate of $1e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, learning rate warms up 8,000 steps and uses the decay strategies in [26]. The dropout rate in all layers is 0.2, and the label smoothing regularization strategy is applied with 0.1. During Inference, the standard beam search strategy is used with beam size 10. The decoding process will stop when the *eos* token is predicted or the length of the generated summary exceeds 100. The setting of hyper-parameters can be found on the Open-NMT website[5].

For the semantic retrieval model, the pre-trained BERT-BASE model with 12-layer, 768-hidden, 12-heads, and 110M parameters[6] is firstly employed to transform the source document and target summary into a fixed-length 768-dimensional vector. In detail, we perform an average pooling operation over all of the tokens of the second-to-last hidden layers of BERT to yield the fixed-length vector. Similar to Open-NMT, the source document is truncated to 400 tokens and the target summary is limited to 100 tokens. Then, an open-source toolkit FAISS for dense vectors similarity search is used to generate the top k vectors with the highest similarity for each document vector. Note that the semantic distance in this paper is L2 distance.

[5] https://opennmt.net/OpenNMT-py/Summarization.html

[6] The BERT-BASE model can be downloaded at https://github.com/google-research/bert#pre-trained-models

Table 2 Full-length ROUGE F1 results of different models on the CNN/Daily mail test set

Models	RG-1	RG-2	RG-L
Pointer-Generator [22]	36.44	15.66	33.42
Pointer-Generator+Coverage [22]	39.53	17.28	36.38
Open-NMT-Transformer [10]	39.25	17.54	36.45
Open-NMT-BRNN [10]	39.12	17.35	36.12
Bottom-Up [7]	41.22	18.68	**38.34**
KIGN [32]	40.34	17.70	36.57
DEATS [38]	40.85	18.08	37.13
Lead-3 [22]	40.34	17.70	36.57
SummaRuNNer [16]	39.60	16.20	35.30
NeuSum [44]	41.59	19.01	37.98
HER [14]	41.70	18.30	37.10
Multi-Template+1-best template	40.36	18.24	36.63
Multi-Template+3-best template	**41.85**	**19.46**	37.56

Bold entries are the best results

4 Experimental results

In this paper, we assess the efficiency of our model on two summarization datasets: CNN/Daily mail and English Gigaword. The quality of the generated summary is estimated by ROUGE scores [11], which are calculated by counting the number of overlapping units between the generated summary and the golden summary. We report the ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest common sequence) score in this paper using the pyrouge script[7]. For simplicity, we refer them as RG-1, RG-2, and RG-L in the rest of this paper, respectively.

4.1 Results on CNN/Daily mail

We compare our model with several strong abstractive and extractive baseline models in terms of ROUGE, and the results are presented in Table 2. The first block in the table shows the results of abstractive approaches: (1) **Pointer-Generator** is proposed by See et al. [22], in which the model can copy words from the source document to solve the out of vocabulary problem (OOV) and build a solid state-of-the-art baseline; (2) **Pointer-Generator+Coverage** [22] is improved from the Pointer Generator model by introducing a coverage mechanism to alleviate the problem of repetition word problem; (3) **Open-NMT-Transformer** [10] is a basic benchmark model based on the standard hyper-parameters of Open-NMT toolkit; (4) **Open-NMT-BRNN** [10] takes Bi-direction Long Short Term Memory network as the encoder and decoder instead of multi-head self-attention mechanism. In detail, 2-layer BILSTM is adopted and the dimension of embedding and hidden

[7] https://github.com/bheinzerling/pyrouge

state are set to 256 and 1024 respectively; (5) **Bottom-Up** [7] modifies the Open-NMT-Transformer by introducing a content selector and achieve significant improvement compared with the basic Transformer; (6) **KIGN** [32] extracts key information including key sentences and keywords from the source document and fuse the key information into the decoding process by utilizing a multi-view attention mechanism; (7) **DEATS** [38] employs a dual encoding mechanism to produce different level representations of the source document. The second block shows the results of several strong extractive methods: (1) **Lead-3** [22] is a simple and strong extractive baseline, which takes the first three sentences of the source document as the gold summary; (2) **SummaRuNNer** [16] regards the extractive summarization task as sentence-level sequence labeling processing, which utilizes hierarchical encoder to yield the representation of sentence and optimizes the model with pseudo label generated by ROUGE score; (3) **NeuSum** [44] jointly models the sentence scoring and selection progress and extracts the summary in auto-regression paradigm. (4) **HER** [14] divides the process of summary extraction as rough reading and careful reading, which firstly captures the feature of paragraphs through a hierarchical encoder, then the summaries are selected by a decision making policy mechanism. The last block presents the results of our models. Note that we just list the results of 3-best template model and 1-best template model, and the performance of other settings can be seen in Section 5.4. The reason for choosing these two models is that 1-best template model is the basic model, while 3-best model achieves the best performance. The results show that our model substantially achieves improvements by introducing the target-side template compared with the vanilla Transformer model. Furthermore, compared with both abstractive and extractive methods, our model with 3-best template achieves significant improvements on RG-1 and RG-2 metrics, which further demonstrates the effectiveness of our model.

4.2 Results on english Gigaword

Table 3 summarizes the results on English Gigaword corpus. The first block shows several strong abstractive methods: (1) ABS [21], the first abstractive summarization methods based on sequence-to-sequence framework; (2) SEASS [45] constructs the second-level representations of source document by introducing selective gate mechanism; (3) SunAttn [19] introduces a special attention mechanism, which adds decoder input into the processing of generating attention distributions; (4) MASS [23] is a pre-trained sequence generation model, and achieves competitive performance compared with previous works; (5) Open-NMT-Transformer [10] is the standard encoder-decoder model based on Transformer. The second block presents two template-based methods: R^3Sum [2] and BISET [27], which utilize the 1-best template retrieved by Lucene tool to guide the decoding process. As shown in Table 3, compared with Open-NMT-Transformer model, our model achieves significant improvements by introducing target-side templates. The results further show that our proposed model consistently outperforms the abstractive baselines, even the state-of-the-art pre-trained model MASS. We also compare our model with the template-based model, which further demonstrates the effectiveness of our proposed semantic retrieval method.

Table 3 Full-length ROUGE F1 evaluation results on the English Gigaword test set

Models	RG-1	RG-2	RG-L
ABS [21]	29.55	11.32	26.42
SEASS [45]	36.15	17.54	33.63
SunAttn [19]	38.27	16.45	36.08
MASS [23]	38.73	19.71	35.96
Open-NMT-Transformer [10]	35.51	17.35	33.17
R^3Sum [2]	37.04	19.03	34.46
BISET [27]	39.11	**19.78**	36.87
Multi-Template+1-best template	38.27	18.02	35.13
Multi-Template+3-best template	**39.71**	19.65	**37.41**

Best model models are bold

5 Discussion

5.1 Analysis on CNN/Daily mail dataset

In this part, we give a further analysis of our model on the CNN/Daily mail dataset. As presented in Table 2, we first notice that our model with 3-best templates significantly outperforms other baseline systems both in ROUGE-1 and ROUGE-2, while weaker on ROUGE-L than the Bottom-UP model [7]. We hypothesize that the superior performance of our model stems from the target-side global information learned from multiple high-quality templates. Another interesting point that can be observed is that 1-best template model achieves better performance than several strong baseline models. For the popular ROUGE-1 metric, compared with the Pointer-Generator network, our model with 1-best template acquires a notable improvement (+3.92 in terms of ROUGE-1). This also indicates that the model can learn target-side global information from the templates to guide the decoding process. Especially, compared with the Open-NMT-Transformer model, which is our baseline model, the performance is improved by introducing only 1-best template, achieves +1.11, +0.70 and +0.18 points

increment in terms of ROUGE-1, ROUGE-2 and ROUGE-L, respectively. Furthermore, it is clear that extractive methods acquire competitive performance compared with the abstractive baseline models. We find that 1-best template model achieves similar performance with the Lead-3 model. However, the 3-best template model outperforms the Lead-3 model by a large margin (+1.51 ROUGE-1, +1.76 ROUGE-2 and +0.99 ROUGE-L). Therefore, it is safe to be concluded that the performance of our model can be significantly improved by introducing target-side templates into the encoder-decoder framework and build a new state-of-the-art baseline for document-level summarization task.

Moreover, we also realize that introducing multiple templates into the model may bring more parameters, resulting in low-efficient decoding. There are several promising directions to deal with this problem. Spiking neural networks (SNNs) is the third generation of artificial neural network, which can realize large-scale calculation in a low-power dissipation environment [13, 34]. Hence, we can explore the application of SNNs to produce fast and energy-efficient text summarization model. In addition, to deploy the SNNs-based model on resource-critical devices, we need to further explore how to design an effective digital neural computing platform [33, 35–37].

5.2 Analysis on english Gigaword dataset

Taking a step further into analyzing the performance of our proposed model on the sentence-level dataset. First of all, we observe that the Open-NMT-Transformer model achieves poor performance. This indicates that the Transformer model has no obvious advantage compared with RNN model (e.g. LSTM, GRU) in sentence-level summarization task. But it can be seen that our model achieves 2.76 improvements in terms of ROUGE-1 by introducing only 1-best template into the Open-NMT-Transformer model. The second point can be concluded from Table 3 is that our model with 3-best template model gains better performance compared with the abstractive models, even the pre-trained model MASS. This also indicates the target-side templates play a crucial role in the decoding process. Furthermore, the performance of our proposed model with 1-best template is superior to the template-based model R^3Sum. This also can testify our hypothesis that the semantic-based retrieval method can yield high-quality target-side templates and is also effective in the sentence level summarization scenario. Moreover, in contrast to R^3Sum and BISET, our model with 3-best template exhibits much better performance. These results verify our hypothesis that multiple high-quality templates can provide more informative information for the decoder and achieve better performance.

5.3 Performance of the semantic retrieval model

The role of the semantic retrieval model is to produce high-quality target-side templates via semantic distance. To assess the quality of retrieved templates, we first report the ROUGE scores of different templates with the reference summary as the ground truth. We introduce three retrieval methods: **Random method** is to search a summary completely randomly from the training set. **Word matching method** means to take the source document as query to search the template by Lucene, which is an open-source search tool based on word matching. The summary of the searched document is treated as the target-side template. Note that the searched documents are ordered by the default document scores, which are produced by Lucene to weight the similarity of documents and queries. **Semantic matching** means to retrieve the target-templates by L2 distance from the vector space of training set, which is generated by the BERT model.

As shown in Table 4, the template obtained by Random is completely irrelevant with the gold summary and can not provide any useful information for the decoder. Moreover, although the word matching method has achieved competitive performance in sentence-level summarization task [2, 27], it can not generate informative templates for long document. It can be seen in Table 4 that the ROUGE score of the retrieved template is terrible (only 10.46 in terms of ROUGE-1) based on word matching. Furthermore, the template quality generated by the semantic matching method largely outperforms the word matching method (25.46% increase in terms of ROUGE-1 score). It indicates that the BERT model can map the long document into high-level vector space and ensure the semantic similarity. It is also evident from the table that the semantic-based retrieval method is an effective approach to search semantic similarity templates from vector space by semantic distance.

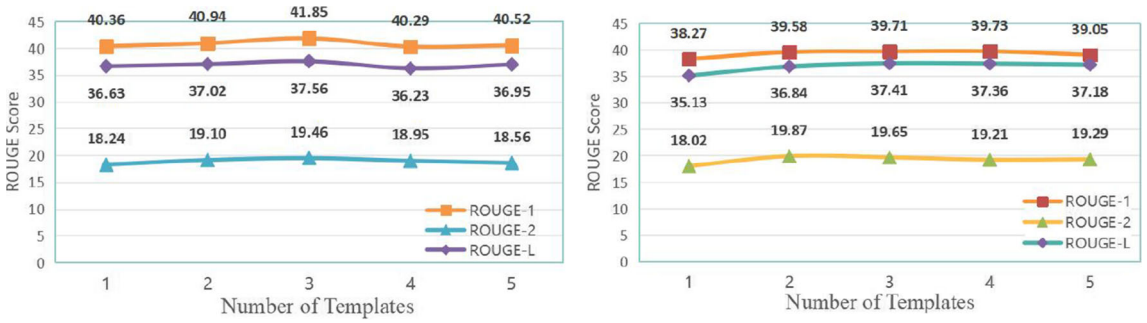
5.4 Performance of different number of templates

In this section, we investigate the impact of the number of templates on model performance. We present the ROUGE scores for the different number of templates on CNN/Daily mail and English Gigaword corpus in Fig. 2. We can find

Table 4 Full-length ROUGE F1 scores of templates generated by different retrieval methods

Methods	RG-1	RG-2	RG-L
Random	3.68	1.24	2.31
Word Matching	10.46	6.70	8.12
Semantic Matching	**35.92**	**17.93**	**29.39**

Bold entries are the best results



(a) ROUGE Score on CNN/Daily mail (b) ROUGE Score on English Gigaword

Fig. 2 The ROUGE scores of the models with different number of templates on document-level CNN/Daily mail corpus and sentence-level English Gigaword corpus

that n-best template models ($n > 1$) show improvements compared with the 1-best baseline model. This shows that multiple templates can provide richer target-side global information than only 1-best template. We also find that the 3-best template model achieves significant improvement than other settings in both CNN/Daily mail and English Gigaword corpus. We conjecture this happens because that the decoder can learn informative target-side global information from multiple templates. Also, we observe that the ROUGE scores do not increase with the number of templates. For instance, In Fig. 2a, compared with the 3-best template model, the 5-best template model achieves a serious decline in terms of ROUGE score (-1.33 ROUGE-1, -0.90 ROUGE-2,-0.61 ROUGE-L). This phenomenon also can be observed in the English Gigaword corpus. We assume the possible reason is that multiple templates will also bring noise into the model, and the model can not discriminate the truly useful information from multiple noisy templates. Due to this observation and the limitation of GPU memory, we only report the score of 5-best templates model at most.

5.5 Ablation study

In this section, we analyze the effectiveness of the sub-modules proposed in this paper. Concretely, **w/o D2T attention** denotes that the D2T attention module is detached from our model, the hidden state at the last time step of each template is directly taken as the context representation of l-th template, that is, the formula (14) is modified as $\boldsymbol{c}^t_{j,l} = z_{l,R}$. **w/o S2T attention** means to remove the S2T attention from our model, each template is considered to be equally for the decoding process, that is, the formula (15) is modified as $\boldsymbol{c}^t_j = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{c}^t_{j,l}$. **w/o D2T and S2T** means to detach both D2T and S2T attention model. In particular, the number of templates in all ablation models is set to 3.

As shown in Table 5, we find that all ablation models obtain lower scores compared with our model (full), which demonstrates that each module of multi-granular attention is effective and necessary. It is also worth noticing that when the S2T and D2T attention are removed, the model performs even worse than Open-NMT-Transformer. We conjecture this happens because the decoder can not distinguish the useful information from multiple templates without any constraint. Then, we only detach the D2T attention model from our model, which means that the same template context information is utilized at each decoding step. As a result, the w/o D2T attention model exhibits terrible performance and shows a serious decline of 1.93,1.78 and 1.29 in terms of ROUGE-1, ROUGE-2, and ROUGE-L, respectively. These results testify our hypothesis that the decoder should learn different contexts from multiple templates at each decoding time step. It is also noteworthy that the model also achieves a slight decline performance after removing the S2T attention module (-1.38 in terms of ROUGE-1), which reveals that the S2T attention can distinguish the important templates by assigning different importance weights to different templates. Another important finding is that the D2T attention module plays a more important role than S2T attention. This may suggest that the D2T attention already captures the important target-side global information at the word-level, while the S2T attention module plays a supplemental role at the sentence-level.

Table 5 Comparison of different ablation models in terms of ROUGE scores

Models	RG-1	RG-2	RG-L
w/o D2T and S2T attention	38.41	17.06	35.55
w/o D2T attention	39.92	17.68	36.27
w/o S2T attention	40.47	18.52	36.88
our model(full)	**41.85**	**19.46**	**37.56**

Bold entries are the best results

Table 6 ROUGE scores of the models with various contaminate strategies on CNN/Daily mail test set

Operation	Single-Template				Multi-Template			
	RG-1	RG-2	RG-L	Δ	RG-1	RG-2	RG-L	Δ
None	40.36	18.24	36.63	–	41.85	19.46	37.56	–
Token Masking	38.68	17.63	35.89	–1.68	40.73	18.50	37.39	–1.12
Token Deletion	38.14	16.72	34.55	–2.22	40.08	18.33	36.98	–1.77
Token Permutation	39.51	17.96	36.41	–0.85	41.54	18.93	37.17	–0.31

5.6 Robustness study

In this paper, we argue that our proposed model overcomes the template selection bias problem caused by using 1-best template by introducing multi-template decoding mechanism. Ideally, multi-templates should provide more robust performance by learning complement information from multi-templates. In order to verify this conjecture, we designed a set of experiments to verify the model's robustness by polluting the templates. As shown in Table 6, we present three types of operation to yield contaminated templates. The *Token Masking* operation replaces the tokens in the templates with Mask sign with a probability of 10%, and the *Token Deletion* operation denotes that deleting the tokens in the templates with a probability of 10%. The *Token Permutation* operation chooses 10% tokens for location exchange. Specifically, the *None* denotes using the original templates, and Δ is the deviation in terms of ROUGE-1 by comparing different contaminate strategy with the original model.

We present the ROUGE scores of the models with various contaminate strategies on CNN/Daily mail test set. Results show that the Token Deletion operation yields a significant decline compared to other operations. We argue the reason is that deleting tokens directly from the templates may result in losing some vital information, which is crucial for the decoder. In contrast, although the Token Masking operation also replaces some tokens with Mask tag, the position information is maintained, and the model can learn that there should be a word in the Mask position. Among the three operations, the Token Permutation operation has little effect on the model's performance. We think

(a) Token Masking

(b) Token Deletion

(c) Token Permutation

Fig. 3 The ROUGE-1 score deviation of different contaminate proportions

Source	andrew getty , one of the heirs to billions of oil money , **appears to have died of natural causes** , a los angeles police department spokesman said . the coroner 's preliminary assessment is there was no foul play involved in the death of getty , grandson of oil tycoon j. paul getty , said detective meghan aguilar . ..... **there is no criminal investigation underway** , he said . ..... **gordon getty is one of three living sons of j. paul getty** , the oil baron who was thought to be the richest man in the world at the time of his death in 1976 . ..... andrew getty had three brothers and three half-sisters . people we 've lost in 2015 .
Gold summary	andrew getty 's **death appears to be from natural causes** , police say , citing coroner 's early assessment . in a petition for a restraining order , getty had written he had a serious medical condition . **police say this is not a criminal matter at this time .**
Template	(1) getty 's **death appears to be natural causes or accident** , coroner 's office says . mother and father of andrew getty confirm death , asks for privacy . (2) jamie silvonek , 14 , was **charged as adult with homicide and criminal** conspiracy in her mother 's stabbing death ..... **police** found victim 's blood-soaked car submerged in a pond and traced it to silvonek family (3) bryan santana , 20 , of orlando , florida , was found **guilty of murdering his former roommate** , shelby fazio , 23 , on thursday . he was sentenced to life terms in prison . ..... they also showed photos of the messages he wrote on the wall in her dog 's blood , including one that said ` i 'm not sorry for what i did '
OpenNMT-Transformaer	**andrew getty , 47 , is one of three living sons of j. paul getty** , u.s. oil tycoon paul getty . aguilar 's parents say there is no foul .
Our model	andrew getty 's death appears to be **natural causes or an accident**. getty is a well-known of oil tycoon. The police said there is no **criminal investigation** underway.

Fig. 4 Example of generated summary on CNN/Daily mail corpus.The words highlighted in blue are the main aspects of the source document, as well as the important parts of the summary and templates

the reason is that the attention vector (Formula 15) is obtained by aggregating the word-level information, and the aggregation operation is an accumulation process without using position information. In addition, it can be seen that the multi-template model gains smaller declines compared with the single-template model under various contaminate strategies. This also proves our conjecture that the proposed multi-template decoding framework is capable of coping with the problem of template selection bias by learning complementary information from multiple templates.

In the above study, we use a 10% proportion to yield contaminated templates. Moreover, we argue the contaminate proportion is an important parameter for robustness testing. Therefore, we further explore the model's performance under different contaminating proportions. As shown in Fig. 3, we can see that the model's performance consistently degenerates with the increase of contaminate proportions, that is, the deviation compared to the original model grows continuously. We can also observe that the multi-template strategy achieves more stable performance than the single template strategy.

5.7 Case study

In this section, we investigate the effectiveness of the multi-template decoding mechanism by presenting the templates and summaries generated by different models. As shown in Fig. 4, the templates obtained by the semantic-based retrieval method are closely related to the main aspect of the source document. Especially, the first template should be a follow-up report of the 'Andrew Getty's death' event, and many entities in the template co-exist with the source document. Hence, the first template can provide critical target-side global information for the decoder. Furthermore, we can observe that although the OpenNMT-Transformer model generates a fluent summary, it focuses on the unimportant aspects of the source document due to lacking target-side global information. Meanwhile, our proposed model overcomes this limitation by introducing multi-template into the decoding process.

6 Conclusion

In this paper, we proposed a novel abstractive document summarization framework to make the best use of the target-side information, which was generated from multiple retrieved templates. Experimental results on CNN/Daily mail and English Gigaword corpus demonstrated that the semantic-based retrieval method was feasible to obtain soft-templates as target-side global information from the long document. The results further showed that the multi-template decoding strategy could improve the performance of the seq2seq model by introducing a multi-granular attention mechanism.

There are several promising future directions. First, better ways to use multiple templates can be explored. Second, the target-side and source-side information can be used simultaneously to improve the performance of the abstractive summarization model.

Acknowledgements We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (Grant Nos. 2018YFC0830105, 2018YFC0830100); the National Natural Science Foundation of China (Grant Nos. 61972186, 61762056, 61472168); the Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202002AD080001); General Projects of Basic Research in Yunnan Province (Grant Nos. 202001AT070047, 202001AT070046).

References

1. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv:1607.06450
2. Cao Z, Li W, Li S, Wei F (2018) Retrieve, rerank and rewrite: Soft template based neural summarization. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 152–161.

Association for Computational Linguistics, Melbourne, Australia. https://doi.org/10.18653/v1/P18-1015. https://www.aclweb.org/anthology/P18-1015
3. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
4. Elbayad M, Gu J, Grave E, Auli M (2019) Depth-adaptive transformer. In: ICLR 2020-Eighth international conference on learning representations
5. Fan A, Grave E, Joulin A (2019) Reducing transformer depth on demand with structured dropout. In: International conference on learning representations
6. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: Proceedings of the 34th international conference on machine learning-Volume 70, pp 1243–1252. JMLR. org
7. Gehrmann S, Deng Y, Rush A (2018) Bottom-up abstractive summarization. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 4098–4109
8. Gu J, Wang Y, Cho K, Li VO (2018) Search engine guided neural machine translation. In: Thirty-second AAAI conference on artificial intelligence
9. Iyyer M, Manjunatha V, Boyd-Graber J, Daumé HIII (2015) Deep unordered composition rivals syntactic methods for text classification. 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), pp 1681–1691. Association for Computational Linguistics, Beijing, China. https://doi.org/10.3115/v1/P15-1162. https://www.aclweb.org/anthology/P15-1162
10. Klein G, Kim Y, Deng Y, Nguyen V, Senellart J, Rush A (2018) OpenNMT: Neural machine translation toolkit. In: Proceedings of the 13th conference of the association for machine translation in the americas (Volume 1: Research Papers), pp 177–184. Association for Machine Translation in the Americas, Boston, MA. https://www.aclweb.org/anthology/W18-1817
11. Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81. Association for Computational Linguistics, Barcelona, Spain. https://www.aclweb.org/anthology/W04-1013
12. Liu Y, Lapata M (2019) Hierarchical transformers for multi-document summarization. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5070–5081
13. Lobov SA, Mikhaylov AN, Shamshin M, Makarov VA, Kazantsev VB (2020) Spatial properties of stdp in a self-learning spiking neural network enable controlling a mobile robot. Front Neurosci 0:88
14. Luo L, Ao X, Song Y, Pan F, Yang M, He Q (2019) Reading like HER: Human reading inspired extractive summarization. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 3033–3043. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1300. https://www.aclweb.org/anthology/D19-1300
15. Miller A, Fisch A, Dodge J, Karimi AH, Bordes A, Weston J (2016) Key-value memory networks for directly reading documents. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1400–1409. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1147. https://www.aclweb.org/anthology/D16-1147
16. Nallapati R, Zhai F, Zhou B (2017) Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Thirty-first AAAI conference on artificial intelligence
17. Nallapati R, Zhou B, dos Santos C, Guçehre Ç, Xiang B (2016) Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of The 20th SIGNLL conference on computational natural language learning, pp 280–290. Association for Computational Linguistics, Berlin, Germany. https://doi.org/10.18653/v1/K16-1028. https://www.aclweb.org/anthology/K16-1028
18. Napoles C, Gormley M, Van Durme B (2012) Annotated gigaword. In: Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction, pp 95–100. Association for computational linguistics
19. Niu J, Sun M, Rodrigues JJ, Liu X (2019) A novel attention mechanism considering decoder input for abstractive text summarization. In: ICC 2019-2019 IEEE International conference on communications (ICC). IEEE, pp 1–7
20. Pandey G, Contractor D, Kumar V, Joshi S (2018) Exemplar encoder-decoder for neural conversation generation. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 1329–1338. Association for Computational Linguistics, Melbourne, Australia. https://doi.org/10.18653/v1/P18-1123. https://www.aclweb.org/anthology/P18-1123
21. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 379–389. Association for Computational Linguistics, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1044. https://www.aclweb.org/anthology/D15-1044
22. See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 1073–1083. Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-1099. https://www.aclweb.org/anthology/P17-1099
23. Song K, Tan X, Qin T, Lu J, Liu TY (2019) Mass: Masked sequence to sequence pre-training for language generation. arXiv:1905.02450
24. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27, pp 3104–3112. Curran Associates, Inc. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
25. Tu Z, Liu Y, Shi S, Zhang T (2018) Learning to remember translation history with a continuous cache. Trans Assoc Computat Linguist 0:407–420. https://www.aclweb.org/anthology/Q18-1029
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Lu K, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, pp 5998–6008. Curran Associates, Inc. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

27. Wang K, Quan X, Wang R (2019) BiSET: Bi-directional selective encoding with template for abstractive summarization. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 2153–2162. Association for Computational Linguistics, Florence, Italy https://doi.org/10.18653/v1/P19-1207, https://www.aclweb.org/anthology/P19-1207
28. Wang Y, Xia Y, Tian F, Gao F, Qin T, Zhai CX, Liu TY (2019) Neural machine translation with soft prototype. In: Advances in neural information processing systems, pp 6313–6322
29. Xia M, Huang G, Liu L, Shi S (2019) Graph based translation memory for neural machine translation. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 7297–7304
30. Xia Y, Tian F, Wu L, Lin J, Qin T, Yu N, Liu TY (2017) Deliberation networks: Sequence generation beyond one-pass decoding. In: Advances in neural information processing systems, pp 1784–1794
31. Xu K, Lai Y, Feng Y, Wang Z (2019) Enhancing key-value memory neural networks for knowledge based question answering. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers), pp 2937–2947. Association for Computational Linguistics, Minneapolis, Minnesota https://doi.org/10.18653/v1/N19-1301, https://www.aclweb.org/anthology/N19-1301,
32. Xu W, Li C, Lee M, Zhang C (2020) Multi-task learning for abstractive text summarization with key information guide network. EURASIP J Adv Signal Process 0:1–11
33. Yang S, Deng B, Wang J, Li H, Lu M, Che Y, Wei X, Loparo KA (2019) Scalable digital neuromorphic architecture for large-scale biophysically meaningful neural network with multi-compartment neurons. IEEE Trans Neural Netw Learn Syst 0(1):148–162
34. Yang S, Gao T, Wang J, Deng B, Lansdell B, Linares-Barranco B (2021) Efficient spike-driven learning with dendritic event-based processing. Front Neurosci 0:97
35. Yang S, Wang J, Deng B, Liu C, Li H, Fietkiewicz C, Loparo KA (2018) Real-time neuromorphic system for large-scale conductance-based spiking neural networks. IEEE Trans Cybern 0(7):2490–2503
36. Yang S, Wang J, Hao X, Li H, Wei X, Deng B, Loparo KA (2021) Bicoss: toward large-scale cognition brain with multigranular neuromorphic architecture. IEEE Transactions on Neural Networks and Learning Systems
37. Yang S, Wang J, Zhang N, Deng B, Pang Y, Azghadi MR (2021) Cerebellumorphic: Large-scale neuromorphic model and architecture for supervised motor learning. IEEE Transactions on Neural Networks and Learning Systems
38. Yao K, Zhang L, Du D, Luo T, Tao L, Wu Y (2020) Dual encoding for abstractive text summarization. IEEE Trans Cybern 0(3):985–996
39. Zhang J, Utiyama M, Sumita E, Neubig G, Nakamura S (2018) Guiding neural machine translation with retrieved translation pieces. In: 1325–1335. Association for Computational Linguistics, New Orleans, Louisiana https://doi.org/10.18653/v1/N18-1120, https://www.aclweb.org/anthology/N18-1120,
40. Zhang X, Su J, Qin Y, Liu Y, Ji R, Wang H (2018) Asynchronous bidirectional decoding for neural machine translation. In: Thirty-second AAAI conference on artificial intelligence
41. Zhang X, Wei F, Zhou M (2019) Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5059–5069
42. Zhou L, Hovy E (2004) Template-filtered headline summarization. In: Text summarization branches out, pp 56–60. Association for Computational Linguistics, Barcelona, Spain https://www.aclweb.org/anthology/W04-1010,
43. Zhou L, Zhang J, Zong C (2019) Synchronous bidirectional neural machine translation. In: Proceedings of the 2019 association for computational linguistics, pp 91–105. Association for Computational Linguistics, Minneapolis, Minnesotahttps://www.aclweb.org/anthology/Q19-1006
44. Zhou Q, Yang N, Wei F, Huang S, Zhou M, Zhao T (2018) Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 654–663. Association for computational linguistics, Melbourne, Australia https://doi.org/10.18653/v1/P18-1061, https://www.aclweb.org/anthology/P18-1061
45. 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 (Volume 1: Long Papers), pp 1095–1104. Association for Computational Linguistics, Vancouver, Canada https://doi.org/10.18653/v1/P17-1101, https://www.aclweb.org/anthology/P17-1101

Yuxin Huang was born in 1983 year. He is a Ph.D. candidate in computer science at Kunming university of Science and Technology. His research interests includenatural language processing, text summarization, machine translation etc.

Zhengtao Yu was born in 1970. He received the Ph.D. degree in computer application technology from Beijing Institute of Technology in 2005. Now he is a professor and Ph.D. supervisor at Kunming University of Science and Technology, and the director of Yunnan Key Laboratory of Artificial Intelligence. His research interests include natural language processing, machine translation and information retrieval, etc.

Junjun Guo was born in 1988. He received the Ph.D. degree from Xi'an Jiao Tong University in 2016. Now He is a lecturer at Kunming University of Science and Technology. His research interests include natural language processing, machine translation etc.

Zhiqiang Yu was born in 1983 year. He is a Ph.D. candidate in computer science at Kunming university of Science and Technology. His research interests include natural language processing, neural machine translation etc.

Yan Xiang was born in 1979. She received the M.S. degree from Wuhan University in 2001. She is currently a Ph.D. candidate in computer science at Kunming University of Science and Technology. Her research interests include medical image processing, naturallanguage processing, sentiment classification, and text mining etc.

Yantuan Xian was born in 1981. He is currently a Ph.D. candidate in computer science at Kunming University of Science and Technology. His research interests include naturallanguage processing, information retrieval, and text mining etc.