Global Encoding for Abstractive Summarization

Junyang Lin, Xu Sun, Shuming Ma, Qi Su
MOE Key Lab of Computational Linguistics, School of EECS, Peking University
School of Foreign Languages, Peking University
{linjunyang, xusun, shumingma, sukia}@pku.edu.cn

Abstract

In neural abstractive summarization, the conventional sequence-to-sequence (seq2seq) model often suffers from repetition and semantic irrelevance. To tackle the problem, we propose a global encoding framework, which controls the information flow from the encoder to the decoder based on the global information of the source context. It consists of a convolutional gated unit to perform global encoding to improve the representations of the source-side information. Evaluations on the LCSTS and the English Gigaword both demonstrate that our model outperforms the baseline models, and the analysis shows that our model is capable of generating summary of higher quality and reducing repetition[1].

1 Introduction

Abstractive summarization can be regarded as a sequence mapping task that the source text should be mapped to the target summary. Therefore, sequence-to-sequence learning can be applied to neural abstractive summarization (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014), whose model consists of an encoder and a decoder. Attention mechanism has been broadly used in seq2seq models where the decoder extracts information from the encoder based on the attention scores on the source-side information (Bahdanau et al., 2014; Luong et al., 2015). Many attention-based seq2seq models have been proposed for abstractive summarization (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016), which outperformed the conventional statistical methods.

[1]The code is available at https://www.github.com/lancopku/Global-Encoding

Text: the mainstream fatah movement on monday officially chose mahmoud abbas, chairman of the palestine liberation organization (plo), as its candidate to run for the presidential election due on jan. #, ####, the official wafa news agency reported.
seq2seq: fatah officially officially elects abbas as candidate for candidate .
Gold: fatah officially elects abbas as candidate for presidential election

Table 1: An example of the summary of the conventional attention-based seq2seq model on the Gigaword dataset. The text highlighted indicates repetition, "#" refers to masked number.

However, recent studies show that there are salient problems in the attention mechanism. Zhou et al. (2017) pointed out that there is no obvious alignment relationship between the source text and the target summary, and the encoder outputs contain noise for the attention. For example, in the summary generated by the seq2seq in Table 1, "officially" is followed by the same word, as the attention mechanism still attends to the word with high attention score. Attention-based seq2seq model for abstractive summarization can suffer from repetition and semantic irrelevance, causing grammatical errors and insufficient reflection of the main idea of the source text.

To tackle this problem, we propose a model of global encoding for abstractive summarization. We set a convolutional gated unit to perform global encoding on the source context. The gate based on convolutional neural network (CNN) filters each encoder output based on the global context due to the parameter sharing, so that the representations at each time step are refined with consideration of the global context. We conduct experiments on LCSTS and Gigaword, two benchmark datasets for sentence summarization, which shows that our model outperforms the state-of-the-art methods with ROUGE-2 F1 score 26.8 and 17.8 respectively. Moreover, the analysis shows

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 163–169
Melbourne, Australia, July 15 - 20, 2018.

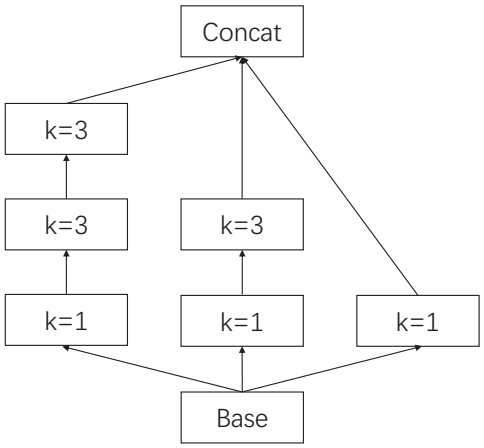

Figure 1: **Structure of our proposed Convolutional Gated Unit.** We implement 1-dimensional convolution with a structure similar to the Inception (Szegedy et al., 2015) over the outputs of the RNN encoder, where k refers to the kernel size.

that our model is capable of reducing repetition compared with the seq2seq model.

2 Global Encoding

Our model is based on the seq2seq model with attention. For the encoder, we set a convolutional gated unit for global encoding. Based on the outputs from the RNN encoder, the global encoding refines the representation of the source context with a CNN to improve the connection of the word representation with the global context. In the following, the techniques are introduced in detail .

2.1 Attention-based seq2seq

The RNN encoder receives the word embedding of each word from the source text sequentially. The final hidden state with the information of the whole source text becomes the initial hidden state of the decoder. Here our encoder is a bidirectional LSTM encoder, where the encoder outputs from both directions at each time step are concatenated ($h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$).

We implement a unidirectional LSTM decoder to read the input words and generate summary word by word, with a fixed target vocabulary embedded in a high-dimensional space $Y \in R^{|Y| \times dim}$. At each time step, the decoder generates a summary word y_t by sampling from a distribution of the target vocabulary P_{vocab} until sampling the token representing the end of sentence. The hidden state of the decoder s_t and the encoders output h_i at each time step i of the encoding process are computed with a weight matrix W_a to obtain the global attention $\alpha_{t,i}$ and the context vector c_t. It is described below:

$$P_{vocab} = softmax(g([c_t; s_t])) \quad (1)$$

$$s_t = LSTM(y_{t-1}, s_{t-1}, C_{t-1}) \quad (2)$$

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i \quad (3)$$

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{n} exp(e_{t,j})} \quad (4)$$

$$e_{t,i} = s_{t-1}^{\top} W_a h_i \quad (5)$$

where C refers to the cell state in the LSTM, and $g(\cdot)$ refers to a non-linear function.

2.2 Convolutional Gated Unit

Abstractive summarization requires the core information at each encoding time step. To reach this goal, we implement a gated unit on top of the encoder outputs at each time step, which is a CNN that convolves all the encoder outputs. The parameter sharing of the convolutional kernels enables the model to extract certain types of features, specifically n-gram features. Similar to image, language also contains local correlation, such as the internal correlation of phrase structure. The convolutional units can extract these common features in the sentence and indicate the correlation among the source annotations. Moreover, to further strengthen the global information, we implement self-attention (Vaswani et al., 2017) to mine the relationship of the annotation at a certain time step with other annotations. Therefore, the gated unit is able to find out both common n-gram features and global correlation. Based on the convolution and self-attention, the gated unit sets a gate to filter the source annotations from the RNN encoder, in order to select information relevant to the global semantic meaning. The global encoding allows the encoder output at each time step to become new representation vector with further connection to the global source side information. For convolution, we implement a structure similar to inception (Szegedy et al., 2015). We use 1-dimension convolution to extract n-gram features. Following the design principle of inception, we did not use kernel where $k = 5$ but instead used two kernels where $k = 3$ to avoid large kernel size. The details of convolution block is described be-

low:

$$g_i = ReLU(W[h_{i-k/2}, ..., h_{i+k/2}] + b) \quad (6)$$

where $ReLU$ refers to the non-linear activation function Rectified Linear Unit (Nair and Hinton, 2010). Based on the convolution block, we implement a structure similar to inception, as shown in Figure 1.

On top of the new representations generated by the CNN module, we further implement self-attention upon these representations so as to dig out the global correlations. Vaswani et al. (2017) pointed out that self-attention encourages the model to learn long-term dependencies and does not create much computational complexity, so we implement its scaled dot-product attention for the connection between the annotation at each time step and the global information:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (7)$$

where the representations, are computed through the attention mechanism with itself and packed into a matrix. To be specific, we refer Q and V to the representation matrix generated by the CNN module, while $K = W_{att}V$ where W_{att} is a learnable matrix.

A further step is to set a gate based on the generation from the CNN and self-attention module g for the source representations h' from the RNN encoder, where:

$$\tilde{h} = h \odot \sigma(g) \quad (8)$$

Since the CNN module can extract n-gram features of the whole source text and self-attention learns the long-term dependencies among the components of the input source text, the gate can perform global encoding on the encoder outputs. Based on the output of the CNN and self-attention, the logistic sigmoid function outputs a vector of value between 0 and 1 at each dimension. If the value is close to 0, the gate removes most of the information at the corresponding dimension of the source representation, and if it is close to 1, it reserves most of the information.

2.3 Training

In the following, we introduce the datasets that we conduct experiments on as well as our experimental settings.

Given the parameters θ and source text x, the models generates a summary $\tilde{y}$. The learning process is to minimize the negative log-likelihood between the generated summary $\tilde{y}$ and reference y:

$$\mathcal{L} = -\frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} \sum_{t=1}^{\mathrm{T}} p(y_t^{(n)} | \tilde{y}_{<t}^{(n)}, x^{(n)}, \theta) \quad (9)$$

where the loss function is equivalent to maximizing the conditional probability of summary y given parameters θ and source sequence x.

3 Experiment Setup

In the following, we introduce the datasets that we conduct experiments on and our experiment settings as well as the baseline models that we compare with.

3.1 Datasets

LCSTS is a large-scale Chinese short text summarization dataset collected from Sina Weibo, a famous Chinese social media website (Hu et al., 2015), consisting of more than 2.4 million text-summary pairs. The original texts are shorter than 140 Chinese characters, and the summaries are created manually. We follow the previous research (Hu et al., 2015) to split the dataset for training, validation and testing, with 2.4M sentence pairs for training, 8K for validation and 0.7K for testing.

The English Gigaword is a sentence summarization dataset based on Annotated Gigaword (Napoles et al., 2012), a dataset consisting of sentence pairs, which are the first sentence of the collected news articles and the corresponding headlines. We use the data preprocessed by Rush et al. (2015) with 3.8M sentence pairs for training, 8K for validation and 2K for testing.

3.2 Experiment Settings

We implement our experiments in PyTorch on an NVIDIA 1080Ti GPU. The word embedding dimension and the number of hidden units are both 512. In both experiments, the batch size is set to 64. We use Adam optimizer (Kingma and Ba, 2014) with the default setting $\alpha = 0.001, \beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. The learning rate is halved every epoch. Gradient clipping is applied with range [-10, 10].

Following the previous studies, we choose ROUGE score to evaluate the performance of our model (Lin and Hovy, 2003). ROUGE score is to

Model	R-1	R-2	R-L
RNN	21.5	8.9	18.6
RNN-context	29.9	17.4	27.2
CopyNet	34.4	21.6	31.3
SRB	33.3	20.0	30.1
DRGD	37.0	24.2	34.2
seq2seq (Our impl.)	33.8	23.1	32.5
+CGU	**39.4**	**26.9**	**36.5**

Table 2: **F-Score of ROUGE on LCSTS.**

Model	R-1	R-2	R-L
ABS	29.6	11.3	26.4
ABS+	29.8	11.9	27.0
Feats	32.7	15.6	30.6
RAS-LSTM	32.6	14.7	30.0
RAS-Elman	33.8	16.0	31.2
SEASS	36.2	17.5	33.6
DRGD	**36.3**	17.6	33.6
seq2seq (Our impl.)	33.6	16.3	31.3
+CGU	**36.3**	**18.0**	**33.8**

Table 3: **F-Score of ROUGE on Gigaword.**

calculate the degree of overlapping between generated summary and reference, including the number of n-grams. F1 scores of ROUGE-1, ROUGE-2 and ROUGE-L are used as the evaluation metrics.

3.3 Baseline Models

As we compare our results with the results of the baseline models reported in their original papers, the evaluation on the two datasets has different baselines. In the following, we introduce the baselines for LCSTS and Gigaword respectively.

Baselines for LCSTS are introduced in the following. **RNN** and **RNN-context** are the RNN-based seq2seq models (Hu et al., 2015), without and with attention mechanism respectively. **CopyNet** is the attention-based seq2seq model with the copy mechanism (Gu et al., 2016). **SRB** is a model that improves semantic relevance between source text and summary (Ma et al., 2017). **DRGD** is the conventional seq2seq with a deep recurrent generative decoder (Li et al., 2017).

As to the baselines for Gigaword, **ABS** and **ABS+** are the models with local attention and handcrafted features (Rush et al., 2015). **Feats** is a fully RNN seq2seq model with some specific methods to control the vocabulary size. **RAS-LSTM** and **RAS-Elman** are seq2seq models with a convolutional encoder and an LSTM decoder and an Elman RNN decoder respectively. **SEASS** is a seq2seq model with a selective gate mechanism. **DRGD** is also a baseline for Gigaword.

Results of our implementation of the conventional seq2seq model on both datasets are also used for the evaluation of the improvement of our proposed convolutional gated unit (CGU).

4 Analysis

In the following sections, we report the results of our experiments and analyze the performance of our model on the evaluation of repetition. Also, we provide an example to demonstrate that our model can generate summary that is more semantically consistent with the source text.

4.1 Results

In the experiments on the two datasets, our model achieves advantages of ROUGE score over the baselines, and the advantages of ROUGE score on the LCSTS are significant. Table 2 presents the results of our model and the baselines on the LCSTS, and Table 2 shows the results of models on the Gigaword. We compare the F1 scores of our model with those of the baseline models (reported in their original articles) and our own implementation of the attention-based seq2seq. Compared with the conventional seq2seq model, our model owns an advantage of ROUGE-2 score 3.7 and 1.5 on the LCSTS and Gigaword respectively.

4.2 Discussion

We show a summary generated by our model, compared with that of the baseline seq2seq model and the reference. The source text introduces a phenomenon that Starbucks, an ordinary coffee brand in the United States, becomes a brand of high class and sells coffee in a much higher price. It is apparent that the main idea of the text is about the high price of Starbucks coffee in China. However, the seq2seq model generates a summary which only contains the information of the brand and the country. In addition, it has committed a mistake of redundant repetition of the word “China”. It is not semantically relevant to the source text and it is not coherent and adequate. Compared with it, the summary of our model is more coherent and more semantically relevant to the source text. Our model focuses on the information about price instead of country, and points

Source: 较早进入中国市场的星巴克，是不少小资钟情的品牌。相比在美国的平民形象，星巴克在中国就显得“高端”得多。用料并无差别的一杯中杯美式咖啡，在美国仅约合人民币12元，国内要卖21元，相当于贵了75%。第一财经日报 Starbucks, which entered Chinese market early, is a brand appealing to young people of petit bourgeoisie. Compared with its ordinary image in the United States, Starbucks seems to be of higher class in China. A Tall Americano sells about 12RMB in the United States, but 21RMB in China, which means it is 75% more expensive.
Reference: 媒体称星巴克美式咖啡售价中国比美国贵75%。 Media report that the price of Starbucks Americano in China is 75% more expensive than that in the United States.
seq2seq: 星巴克中国美式咖啡在中国。 Starbucks China Americano in China.
+CGU: 星巴克美式咖啡中国贵75%。 Starbucks Americano is 75% more expensive in China.

Table 4: An example of our summarization, compared with that of the seq2seq model and the reference.

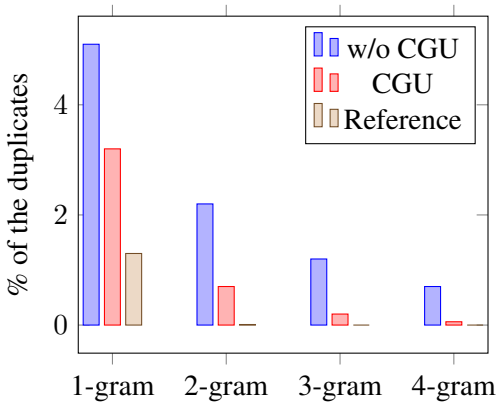

Figure 2: **Percentage of the duplicates at sentence level.** Evaluated on the Gigaword.

out the price gap in its generated summary. As “China” appears twice in the source text and it is hard for the baseline model to put it in a less significant place, but for our model with CGU, it is able to filter the trivial details that are irrelevant to the core meaning of the source text and just focuses on the information that contributes most to the main idea.

As our CGU is responsible for selecting important information of the outputs from the RNN encoder to improve the quality of the attention score, it should be able to reduce repetition in the generated summary. We evaluate the degree of repetition by calculating the percentage of the duplicates at the sentence level. The evaluations on the Gigaword for duplicates of 1-gram to 4 gram prove that our model significantly reduces repetition compared to the conventional seq2seq and its repetition rate is similar to the reference’s. This also shows that our model is able to generate summaries of higher diversity with less repetition.

5 Related Work

Researchers developed many statistical methods and linguistic-rule-based methods to study automatic summarization (Banko et al., 2000; Dorr et al., 2003; Zajic et al., 2004; Cohn and Lapata, 2008). With the development of Neural Network in NLP, more and more researches have appeared in abstractive summarization since it seems possible that Neural Network can help achieve the two goals. Rush et al. (2015) first applied sequence-to-sequence model with attention mechanism to abstractive summarization and realized significant achievements. Chopra et al. (2016) changed the ABS model with an RNN decoder and Nallapati et al. (2016) changed the system to a fully-RNN sequence-to-sequence model and achieved outstanding performance. Zhou et al. (2017) proposed a selective gate mechanism to filter secondary information. Li et al. (2017) proposed a deep recurrent generative decoder to learn latent structure information. Ma et al. (2018) proposed a model that generates words by querying word embeddings.

6 Conclusion

In this paper, we propose a new model for abstractive summarization. The convolutional gated unit performs global encoding on the source side information so that the core information can be reserved and the secondary information can be filtered. Experiments on the LCSTS and Gigaword show that our model outperforms the baselines, and the analysis shows that it is able to reduce repetition in the generated summaries, and it is more robust to inputs of different lengths, compared with the conventional seq2seq model.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 61673028), National High Technology Research and Development Program of China (863 Program, No. 2015AA015404), and the National Thousand Young Talents Program. Xu Sun is the corresponding author of this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014*, pages 1724–1734.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 137–144. Association for Computational Linguistics.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. 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.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL 2016*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *EMNLP 2015*, pages 1967–1972.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP 2013*, pages 1700–1709.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *EMNLP 2017,*, pages 2091–2100.

Chin Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, pages 1412–1421.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *NAACL 2018*.

Shuming Ma, Xu Sun, Jingjing Xu, Houfeng Wang, Wenjie Li, and Qi Su. 2017. Improving semantic relevance for sequence-to-sequence learning of chinese social media text summarization. In *ACL 2017*, pages 635–640.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML 2010*, pages 807–814.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL 2016*, pages 280–290.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP 2015*, pages 379–389.

Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. 2017. Label embedding network: Learning label representation for soft training of deep networks. *CoRR*, abs/1710.10393.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *EMNLP 2016*, pages 1054–1059.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS 2017*, pages 6000–6010.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2004. Bbn/umd at duc-2004: Topiary. In *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pages 112–119.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *ACL 2017*, pages 1095–1104.