Improving Abstractive Summarization via Dilated Convolution

Dawei Jin¹, Ruizhi Kang¹, Hongjun Zhang¹, Wening Hao¹, Gang Chen¹

¹ Army Engineering University of PLA, Nanjing 210007, China

*Corresponding author: Dawei Jin (e-mail: dave_nj@163.com).

Abstract: In this paper, a sequence-to-sequence based hybrid neural network model is proposed for abstractive summarization. Our method utilizes Bi-directional Long Short-Term Memory (Bi-LSTM) and multi-level dilated convolutions (MDC) to capture the global semantic information and semantic-unit level information, respectively. In decoding phrase, our model generates words according to summary relevant information captured by attention mechanism. Experiment shows that this proposed model outperforms several strong baselines on both of Gigawords corpus and DUC-2004 task.

1. Introduction
Text summarization is a technology to help quickly grasp the text information. It compresses the length of original text while keeping the semantics of original text as far as possible, with the aim of making the length of summary less that 50 percent of that of the original text [1]. Text summarization is challenging due to that the model should obtain the semantic of source text first, and then generate a readable short text retaining the key points.

There are two prominent types of summarization algorithms: extractive summarization models extract sentences from source text by measuring their importance to make up the summary; Abstractive summarization, which is similar to manual summary, that is to understand the semantics of the source text and then reorganize words to generate a new summary. For extractive summarization models, since sentences are extracted from the original text, it does not need to pay attention to some problems such as grammar and tense. Importance of sentence is based on linguistic and statistical features. Abstractive approaches are more difficult, because the model generate summary using new words and phrases on the basis of understanding the source text.

Sequence to sequence model, which maps an input sequence into another output sequence, has been used in abstractive summarization and achieves state-of-the-art performance. The overall structure of Seq2Seq models consist of one encoder and one decoder. The encoder is mainly used to get semantic information of the input sequence and represent the input into a hidden state. The decoder learns the relationship between the hidden state and the output sequence. The basic model of encoder and decoder can be RNNs or CNN. RNNs based Seq2Seq model [2][3], such as LSTM [4] based encoder, can capture long-term dependency pretty well and the information in the starting of the sequence is able to traverse down the line. The training efficiency of CNN based Seq2Seq model is faster[5], therefore, more and more CNN based models combined with attention mechanism are used in NLP tasks and achieve considerable results [6][7].

The key point of abstractive summarization model is how to capture the semantic information of the input sequence and retain the key points containing the salient meanings of the source text for the
output sequence. We believe the core meanings of text are consisted of several core semantic-units. However, standard CNN extracts semantic information at word-level, which is not enough to represent core semantic-units. For RNNs, they capture semantic information of the entire source text, which will submerge the core semantic-units. Dilated convolutions [8] systematically aggregate multi scale contextual information without losing resolution, which are used in image semantic segmentation tasks and achieve a better accuracy than traditional CNN. Dilated convolution refers to convolution inserted with “holes” so that it is able to remove the negative effects such as information loss caused by common down-sampling methods, such as max-pooling and strided convolution. Inspired by the success of image semantic segmentation, we propose a dilated CNN to encode input sequence aiming at achieving a better semantic hidden state for Seq2Seq based abstractive summarization model.

In this paper, we introduce a Seq2Seq architecture based on hybrid networks. For this architecture, the encoder consists of one Bi-LSTM layer and three dilated convolution layers, the decoder is a standard LSTM. In the encoding phrase, we use dilated convolution to capture local correlation between words and generate representations of higher levels information. LSTM is used to capture the global semantic information while addressing the long-term dependency issue. On this basis, we use attention mechanism to establish relation between local correlation of words with global semantic information, form the final hidden state, and make a better semantic modeling of the source text.

2. Related Work

In the past research work on text summary, most of them are extractive. Abstractive summarization is data-driven, and it arises with the improvement of computing power and the development of deep learning technology.

Rush et al. [15] proposed a neural network based model and achieved state-of-the-art results on Gigaword and DUC datasets. Nallapati et al. [10] used attentional encoder-decoder recurrent neural networks with attention, and they used POS, NER tags, TF, IDF together with word embeddings as the input of encoder; for the decoder, switching generator-pointer is introduced to decide generating a word from the target vocabulary or generating a pointer to one of the word-positions in the source. Gehring et al. [17] introduced an architecture based entirely on convolutional neural networks. They outperformed many strong recurrent models on very large benchmark datasets at an order of magnitude faster speed. Liu et al. [18] proposed an adversarial process for abstractive text summarization. Coverage mechanism [19] and intra-attention mechanism [20] were used in abstractive summarization task, which was designed to help avoid repeating particular word and has been proved to be effective. Besides intra-attention mechanism, Paulus et al. [20] came up with a reinforcement learning based training method that gave huge improvements.

Dilated convolution [8] was introduced to greatly increase the effective receptive field size without increasing the computational cost and achieved state-of-the-art accuracy in semantic segmentation systems. Yang et al. [21] used a dilated CNN to improve variational autoencoders for text modeling. Lin et al. [22] proposed a model that generates higher-level semantic-unit representations with multi-level dilated convolution for multi-label text classification, which significantly outperformed the baseline models.

3. Model

The abstractive summarization task considered in this paper is defined as follows. Given a source text with \( m \) words \( S = \{s_1, s_2, \ldots, s_m\} \), a target output text with \( n \) words \( T = \{t_1, t_2, \ldots, t_n\} \), must contain the main meaning of the source text and have good readability.
3rd International Symposium on Big Data and Applied Statistics
Journal of Physics: Conference Series 1616 (2020) 012078 doi:10.1088/1742-6596/1616/1/012078

Figure 1 The overall structure of our model, we call it MDCABS. MDCABS contains two main parts: encoder and decoder. The major differences between our approach and other Seq2Seq based abstractive summarization are as follows: in the encoding phrase, we use Bi-LSTM and multi-level dilated CNN (MDC) to model source text semantically; for decoding phrase, the decoder first focuses on the coding representation of MDC to obtain information at the semantic-unit level rather than at word level. Then generates a new representation based on the attention. Finally, the new representation pays attention to the representation from the Bi-LSTM encoder to extract sentence level information. The combination of dilated CNN and Bi-LSTM in abstractive summarization task is novel and produce the remarkable results.

In detail, our model consists the following six layers:

1) Input embedding layer. Inspired by QANet[11], we use word embedding and character embedding in our model. The word embedding p1 = 384 dimensional which is pre-trained by GloVe [9], the character embedding is obtained as [11]: Each character is represented as a trainable vector of dimension p2 = 128, meaning each word can be viewed as the concatenation of the embedding vectors for each of its characters. The length of each word is either truncated or padded to 16. We also adopt a two-layer highway network [12] on top of this representation. The output dimension d = 512.

2) Bi-LSTM encoding layer. As many papers doing, we use Bi-LSTM to make semantic modeling and obtain semantic information in two directions of the input text. In this way, we can solve the problem of long-distance dependence of text semantics. The dimension of this layer is 512.

3) MDC encoding layer. Figure 2 shows the structure of Multi-level Dilated Convolution [14] encoding layer in our model, as we can see, many “holes” are inserted into convolution, which expanding its receptive files without increasing the number of parameters. Compared with down-sampling methods, such as max-pooling, dilated convolution can avoid the information loss without discarding any neural units. A position encoding layer is added to the input of MDC encoding layer consisting of sin and cos functions at varying wavelengths, as defined in [13]. The function of position encoding coding is to add some relative position information to MDC to improve the disadvantage of convolutional neural network in obtaining position information. In our model, the dimension of MDC encoding layer is 512.
4) Semantic-unit level attention layer. The decoder of our model is a Stacked-LSTM. In the semantic-unit level attention layer, the Stacked-LSTM pays attention to MDC and generates a new representation based on the attention, which figures out the most relevant semantic-units level information, making it possible to focus on repetitive elaboration at semantic-units level when generating abstracts.

5) Global attention layer. The attention weights at the global level are computed by the new representation from semantic-unit level attention layer and Bi-LSTM encoding layer. The function of global level attention layer is to get global level information, it enables the model to learn information with semantic proximity and repetition between words.

6) Output layer. At each step of decoding, we use a linear layer as the output layer to predict the probability of the current output word. The predicted words sequence is obtained at the end of decoding phrase, and we use cross-entropy as the loss function between the predicted value and the target.

4. Experiment and Results

4.1. Datasets and Evaluation

Gigaword Corpus1. We use Gigaword corpus used in [15] to train our model and use the open scripts2 to preprocess the corpus. It roughly contains 3.8M training pairs, 190K validation pairs, and 2,000 test pairs.

DUC-20043. Because the amount of DUC-2004 dataset is small, there are only 500 pairs, so it is only used for testing in our experiments.

ROUGE [16] is usually used to evaluate the result of summarization, which is to calculate the number of overlapping units between generated summaries and the reference summaries. In this paper, the evaluation is based on three variants of ROUGE: ROUGE-1, ROUGE-2, and ROUGE-L.

4.2. Baseline and Parameter Settings

We compare our model with some baselines that are also data-driven and neural network based approaches for abstractive sentence summarization.

- ABS and ABS+ [15] used local attention modeling based on neural network, which combined probabilistic model with a generation algorithm that generates each word of summary condition on the input sentence.
- RAS-Elman and RAS-LSTM [23] introduced a condition recurrent neural network provided by a convolutional attention-based encoder to ensure the decoder focus on the appropriate input words. The difference of them is that RAS-Elman use Elman RNN [24] in the decoding phrase, and RAS-LSTM uses Long Short-Term Memory [25].
- lvt2k-1sent and lvt5k-1sent [10] used large vocabulary trick described in [28] in their Encoder-Decoder RNN model.
- DRGD [26] used a recurrent latent random model to get Latent structure information and uses neural variational inference to address the intractable posterior inference for the recurrent latent variables.
- WFE [27], a word frequency estimation (WFE) sub-model, that controlled the output words based on the frequency of each target vocabulary to avoid repeated generations.
- ConvS2S [17] introduced an architecture based entirely on convolutional neural networks, which achieved a result close to the best models at an order of magnitude faster speed.

1 https://catalog.ldc.upenn.edu/ldc2012t21
2 https://github.com/facebook/NAMAS
3 http://duc.nist.gov/duc2004
We implement our model in PyTorch on a K40 GPU. We use the ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-7}$. The learning rate is initialized to 0.001, and it is halved after every epoch.

### 4.3. Results and Case Analysis

| Model       | Gigaword(F1) | DUC-2004(Recall) |
|-------------|--------------|-----------------|
|             | R-1  | R-2  | R-L  | R-1  | R-2  | R-L  |
| ABS         | 29.55 | 11.32 | 26.42 | 26.55 | 7.06 | 22.05 |
| ABS+        | 29.78 | 11.89 | 26.97 | 28.18 | 8.49 | 23.81 |
| RAS-Elman   | 33.78 | 15.97 | 31.15 | 28.97 | 8.26 | 24.06 |
| RAS-LSTM    | 32.55 | 14.70 | 30.03 | 27.41 | 7.69 | 23.06 |
| lvt2k-1sent | 32.67 | 15.59 | 30.64 | 28.35 | 9.46 | 24.59 |
| lvt5k-1sent | 35.30 | 16.64 | 32.62 | 28.61 | 9.42 | 25.24 |
| DRGD        | 36.27 | 17.57 | 33.62 | 31.79 | 10.75 | 27.48 |
| ConvS2S     | 35.88 | 17.48 | 33.29 | 30.44 | 10.84 | 26.90 |
| WFE         | 36.30 | 17.31 | 33.88 | 32.28 | 10.54 | 27.80 |
| MDCABS      | **37.09** | **17.16** | **34.19** | **32.54** | **11.98** | **28.67** |

Table 1 shows ROUGE-F1 on Gigawords and ROUGE-Recall on DUC2004. The proposed model MDCABS achieves the best summarization performance on most of the ROUGE metrics. Our model adds attention mechanism and MDC to the traditional sequence2sequence model, in which attention mechanism is the conventional practice of most neural networks including automatic summarization at this stage. Our model does not use some tricks to control the repetition of abstract words. The experimental results show that MDC is effective in automatic abstraction. We believe that this improvement is due to the fact that MDC provides a more appropriate granularity of learning semantic-unit level information. Decoders focus on different semantic-units in the decoding process, and the words generated according to different semantic-units are different, which avoiding the repetition of the generated words.

| Case 1:  |
|----------|
| **Source:** an italian factory worker who suffered years of racial abuse won compensation wednesday after an employment tribunal found he had suffered discrimination and unfair dismissal at work. |
| **Reference:** italian factory worker wins compensation after suffering racial abuse |
| **MDCABS:** italian factory worker wins compensation after suffering discrimination |

| Case 2:  |
|----------|
| **Source:** south african president thabo mbeki urged zimbabwe to solve its presidential election dispute through dialogue. |
| **Reference:** south african president urges zimbabwe to solve election dispute |
| **MDCABS:** mbeki urges zimbabwe to solve election dispute |

| Case 3:  |
|----------|
| **Source:** police in southeastern bangladesh have arrested seven suspected islamic militants, a police official said friday. |
| **Reference:** police in southeast bangladesh arrest seven suspected islamic militants |
| **MDCABS:** seven suspected islamic militants arrested in bangladesh |

| Case 4:  |
|----------|
| **Source:** some ## kite-flying teams from france, japan, the united states and ## other countries |
and regions have come to Weifang, a major city in east China’s Shandong province, to attend the ongoing #Weifang International Kite Festival.

**Reference:** international kite festival opens in east China

**MDCABS:** international kite festival is held in east China city

Table 2 shows the source texts, reference and the generated summaries. From the cases we can observe that MDCABS can capture semantic-unit level information indeed. Our model generates “discrimination” in the place of “racial abuse” in case 1. In case 2, “South African President Thabo Mbeki” is summarized as “mbeki”, and we think that it’s more accurate to use exact names because the phrase “south african president” used in reference is a descriptive title. From case 3 and case 4, a comparison of the source text and generated summary indicates that the active and passive voice are opposite, and the verb “held” generated in case 4 does not appear in the source texts, which proves that our model grasps the main semantic-unit of the source text and then automatically generates the summary according to the full text semantics.

5. Conclusion

In this work, we apply multi-level dilated convolution to improve the performance for the task of abstractive summarization and achieve the state-of-the-art results on two different datasets. We believe that semantic-units capture the semantic meaning of text, and MDC encoding layer helps to extract semantic-unit level information better.

Deep learning based model is data-driven, which relies heavily on the amount of datasets. In general, the larger the training set, the better the data-driven model. The proposed model is time-consuming due to the combination of LSTM and convolution. We will work on convolution and attention based abstractive summarization model to improve the training efficiency in the future.

Acknowledgements

We thank Youwei Xue for help with study design and paper revisions, and Wehui Wang for insightful feedback.

References

[1] Radev, Dragomir R., Eduard Hovy, and Kathleen McKeown. "Introduction to the special issue on summarization." Computational linguistics 28.4 (2002): 399-408.

[2] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[3] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[4] Hochreiter, Sepp, and Jürgen Schmidhuber. "LSTM can solve hard long time lag problems." Advances in neural information processing systems. 1997.

[5] Gehring, Jonas, et al. "Convolutional sequence to sequence learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

[6] Bradbury, James, et al. "Quasi-recurrent neural networks." arXiv preprint arXiv:1611.01576 (2016).

[7] Kalchbrenner, Nal, et al. "Neural machine translation in linear time." arXiv preprint arXiv:1610.10099 (2016).

[8] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[9] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[10] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).
[11] Yu, Adams Wei, et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." arXiv preprint arXiv:1804.09541 (2018).
[12] Srivastava, Rupesh Kumar, Klaus Greff, and Jürgen Schmidhuber. "Highway networks." arXiv preprint arXiv:1505.00387 (2015).
[13] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
[14] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).
[15] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).
[16] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.
[17] Gehring, Jonas, et al. "Convolutional sequence to sequence learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
[18] Liu, Linqing, et al. "Generative adversarial network for abstractive text summarization." Thirty-second AAAI conference on artificial intelligence. 2018.
[19] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368 (2017).
[20] Paulus, Romain, Caiming Xiong, and Richard Socher. "A deep reinforced model for abstractive summarization." arXiv preprint arXiv:1705.04304 (2017).
[21] Yang, Zichao, et al. "Improved variational autoencoders for text modeling using dilated convolutions." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
[22] Lin, Junyang, et al. "Semantic-unit-based dilated convolution for multi-label text classification." arXiv preprint arXiv:1808.08561 (2018).
[23] Chopra, Sumit, Michael Auli, and Alexander M. Rush. "Abstractive sentence summarization with attentive recurrent neural networks." Proceedings of the 2016 NAACL.
[24] Elman, Jeffrey L. "Finding structure in time." Cognitive science14.2 (1990): 179-211.
[25] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
[26] Li, Piji, et al. "Deep recurrent generative decoder for abstractive text summarization." arXiv preprint arXiv:1708.00625 (2017).
[27] Suzuki, Jun, and Masaaki Nagata. "Cutting-off redundant repeating generations for neural abstractive summarization." arXiv preprint arXiv:1701.00138 (2016).
[28] Jean, Sébastien, et al. "On using very large target vocabulary for neural machine translation." arXiv preprint arXiv:1412.2007 (2014).