# A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization

**Li Wang[1], Junlin Yao[2], Yunzhe Tao[3], Li Zhong[1], Wei Liu[4], Qiang Du[3]**
[1] Tencent Data Center of SNG
[2] ETH Zürich
[3] Columbia University
[4] Tencent AI Lab
lilianwang@tencent.com, jyao@student.ethz.ch, y.tao@columbia.edu,
reggiezhong@tencent.com, wl2223@columbia.edu, qd2125@columbia.edu

## Abstract

In this paper, we propose a deep learning approach to tackle the automatic summarization tasks by incorporating topic information into the convolutional sequence-to-sequence (ConvS2S) model and using self-critical sequence training (SCST) for optimization. Through jointly attending to topics and word-level alignment, our approach can improve coherence, diversity, and informativeness of generated summaries via a biased probability generation mechanism. On the other hand, reinforcement training, like SCST, directly optimizes the proposed model with respect to the non-differentiable metric ROUGE, which also avoids the exposure bias during inference. We carry out the experimental evaluation with state-of-the-art methods over the Gigaword, DUC-2004, and LCSTS datasets. The empirical results demonstrate the superiority of our proposed method in the abstractive summarization.

## 1 Introduction

Automatic text summarization has played an important role in a variety of natural language processing (NLP) applications, such as news headlines generation [Kraaij *et al.*, 2002] and feeds stream digests [Barzilay and McKeown, 2005]. It is of interest to generate informative and representative natural language summaries which are capable of retaining the main ideas of source articles. The key challenges in automatic text summarization are correctly evaluating and selecting important information, efficiently filtering redundant contents, and properly aggregating related segments and making human-readable summaries. Compared to other NLP tasks, the automatic summarization has its own difficulties. For example, unlike machine translation tasks where input and output sequences often share similar lengths, summarization tasks are more likely to have input and output sequences greatly imbalanced. Besides, machine translation tasks usually have some direct word-level alignment between input and output sequences, which is less obvious in summarization.

主要挑战是生成的摘要能准确的评估和选择重要的信息，
并有效的过滤冗余的内容，恰当的组织相关短语，使其具有可读性。

There are two genres of automatic summarization techniques, namely extraction and abstraction. The goal of extractive summarization [Neto *et al.*, 2002] is to produce a summary by selecting important pieces of the source document and concatenating them verbatim, while abstractive summarization [Chopra *et al.*, 2016] generates summaries based on the core ideas of the document, therefore the summaries could be paraphrased in more general terms. Other than extraction, abstractive methods should be able to properly rewrite the core ideas of the source document and assure that the generated summaries are grammatically correct and human readable, which is close to the way how humans do summarization and thus is of interest to us in this paper.

Recently, deep neural network models have been widely used for NLP tasks such as machine translation [Bahdanau *et al.*, 2014], and text summarization [Nallapati *et al.*, 2016b]. In particular, the attention based sequence-to-sequence framework [Bahdanau *et al.*, 2014] with recurrent neural networks (RNNs) [Sutskever *et al.*, 2014] prevails in the NLP tasks. However, RNN-based models are more prone to gradient vanishing due to their chain structure of non-linearities compared to the hierarchical structure of CNN-based models [Dauphin *et al.*, 2016]. In addition, the temporal dependence among the hidden states of RNNs prevents parallelization over the elements of a sequence, which makes the training inefficient.

In this paper, we propose a new approach based on the convolutional sequence-to-sequence (ConvS2S) framework [Gehring *et al.*, 2017] jointly with a topic-aware attention mechanism. To the best of our knowledge, this is the first work for automatic abstractive summarization that incorporates the topic information, which can provide themed and contextual alignment information into deep learning architectures. In addition, we also optimize our proposed model by employing the reinforcement training [Paulus *et al.*, 2017]. The main contributions of this paper include:

- We propose a joint attention and biased probability generation mechanism to incorporate the topic information into an automatic summarization model, which introduces contextual information to help the model generate more coherent summaries with increased diversity and informativeness.
- We employ the self-critical sequence training technique in ConvS2S to directly optimize the model with respect

to the non-differentiable summarization metric ROUGE, which also remedies the exposure bias issue.

- Extensive experimental results on three benchmark datasets demonstrate that by fully exploiting the power of the ConvS2S architecture enhanced by topic embedding and SCST, our proposed model yields high accuracy for abstractive summarization, advancing the state-of-the-art methods.

## 2 Related Work

Automatic text summarization has been widely investigated. Many approaches have been proposed to address this challenging task. Various methods [Neto *et al.*, 2002] focus on the extractive summarization, which select important contents of text and combine them verbatim to produce a summary. On the other hand, abstractive summarization models are able to produce a grammatical summary with a novel expression, most of which [Rush *et al.*, 2015; Chopra *et al.*, 2016; Nallapati *et al.*, 2016a] are built upon the neural attention-based sequence-to-sequence framework [Sutskever *et al.*, 2014].

The predominant models are based on the RNNs [Nallapati *et al.*, 2016b; Shen *et al.*, 2016; Paulus *et al.*, 2017], where the encoder and decoder are constructed using either Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] or Gated Recurrent Unit (GRU) [Cho *et al.*, 2014]. However, very few methods have explored the performance of convolutional structure in summarization tasks. Compared to RNNs, convolutional neural networks (CNNs) enjoy several advantages, including the efficient training by leveraging parallel computing, and mitigating the gradient vanishing problem due to fewer non-linearities [Dauphin *et al.*, 2016]. Notably, the recently proposed gated convolutional network [Dauphin *et al.*, 2016; Gehring *et al.*, 2017] outperforms state-of-the-art RNN-based models in the language modeling and machine translation tasks.

While the ConvS2S model is also evaluated on the abstractive summarization [Gehring *et al.*, 2017], there are several limitations. First, the model is trained by minimizing a maximum-likelihood loss which is sometimes inconsistent with the quality of a summary and the metric that is evaluated from the whole sentences, such as ROUGE [Lin, 2004] In addition, the exposure bias [Ranzato *et al.*, 2015] occurs due to only exposing the model to the training data distribution instead of its own predictions. More importantly, the ConvS2S model utilizes only word-level alignment which may be insufficient for summarization and prone to incoherent generalized summaries. Therefore, the higher level alignment could be a potential assist. For example, the topic information has been introduced to a RNN-based sequence-to-sequence model [Xing *et al.*, 2017] for chatbots to generate more informative responses.

## 3 Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model

In this section, we propose the Reinforced Topic-Aware Convolutional Sequence-to-Sequence model, which consists of a

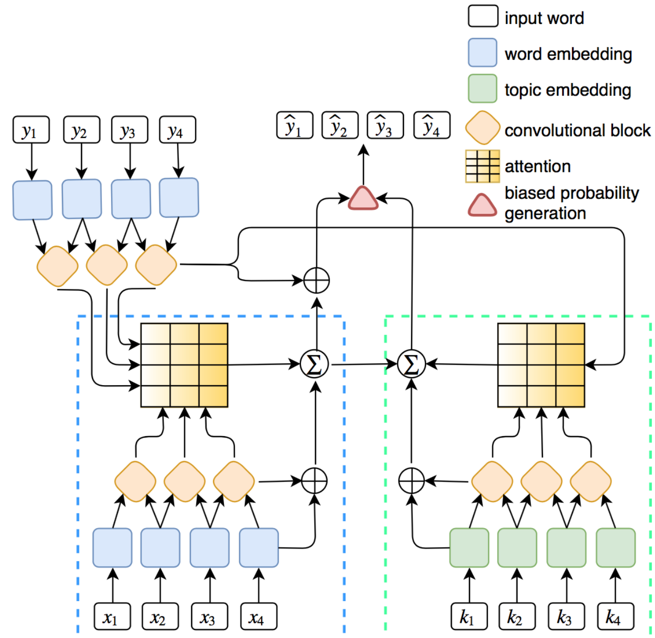


Figure 1: A graphical illustration of the topic-aware convolutional architecture. Word and topic embeddings of the source sequence are encoded by the associated convolutional blocks (bottom left and bottom right). Then we jointly attend to words and topics by computing dot products of decoder representations (top left) and word/topic encoder representations. Finally, we produce the target sequence through a biased probability generation mechanism.

convolutional architecture with both input words and topics, a joint multi-step attention mechanism, a biased generation structure, and a reinforcement learning procedure. The graphical illustration of the topic-aware convolutional architecture can be found in Figure 1.

### 3.1 ConvS2S Architecture

We exploit the ConvS2S architecture [Gehring *et al.*, 2017] as the basic infrastructure of our model. In this paper, two convolutional blocks are employed, associated with the word-level and topic-level embeddings, respectively. We introduce the former in this section and the latter in next, along with the new joint attention and the biased generation mechanism.

#### Position Embeddings

Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ denote the input sentence. We first embed the input elements (words) in a distributional space as $\boldsymbol{w} = (w_1, \ldots, w_m)$, where $w_i \in \mathbb{R}^d$ are rows of a randomly initialized matrix $\mathcal{D}_{\text{word}} \in \mathbb{R}^{V \times d}$ with $V$ being the size of vocabulary. We also add a positional embedding, $\boldsymbol{p} = (p_1, \ldots, p_m)$ with $p_i \in \mathbb{R}^d$, to retain the order information. Thus, the final embedding for the input is $\boldsymbol{e} = (w_1 + p_1, \ldots, w_m + p_m)$. Similarly, let $\boldsymbol{q} = (q_1, \ldots, q_n)$ denote the embedding for output elements that were already generated by the decoder and being fed back to the next step.

#### Convolutional Layer

Both encoder and decoder networks are built by stacking several convolutional layers. Suppose that the kernel has width of $k$ and the input embedding dimension is $d$. The convolution takes a concatenation of $k$ input elements $X \in \mathbb{R}^{kd}$ and maps it to an output element $Y \in \mathbb{R}^{2d}$, namely,

$$Y = f_{\text{conv}}(X) \doteq W_Y X + b_Y \,, \tag{1}$$

where the kernel matrix $W_Y \in \mathbb{R}^{2d \times kd}$ and the bias term $b_Y \in \mathbb{R}^{2d}$ are the parameters to be learned.

Rewrite the output as $Y = [A; B]$, where $A, B \in \mathbb{R}^d$. Then the gated linear unit (GLU) [Dauphin *et al.*, 2016] is given by

$$g([A; B]) = A \otimes \sigma(B) \,, \tag{2}$$

where $\sigma$ is the sigmoid function, $\otimes$ is the point-wise multiplication, and the output of GLU is in $\mathbb{R}^d$.

We denote the outputs of the $l$-th layer as $\boldsymbol{h}^l = (h_1^l, \ldots, h_n^l)$ for the decoder, and $\boldsymbol{z}^l = (z_1^l, \ldots, z_m^l)$ for the encoder. Take the decoder for illustration. The convolution unit $i$ on the $l$-th layer is computed by residual connections as

$$h_i^l = g \circ f_{\text{conv}} \left( \left[ h_{i-k/2}^{l-1}; \cdots ; h_{i+k/2}^{l-1} \right] \right) + h_i^{l-1} \,, \tag{3}$$

where $h_i^l \in \mathbb{R}^d$ and $\circ$ is the function composition operator.

### Multi-step Attention

The attention mechanism is introduced to make the model access historical information. To compute the attention, we first embed the current decoder state $h_i^l$ as

$$d_i^l = W_d^l h_i^l + b_d^l + q_i \,, \tag{4}$$

where $q_i \in \mathbb{R}^d$ is the embedding of the previous decoded element. Weight matrix $W_d^l \in \mathbb{R}^{d \times d}$ and bias $b_d^l \in \mathbb{R}^d$ are the parameters to be learned.

The attention weights $\alpha_{ij}^l$ of state $i$ and source input element $j$ is computed as a dot product between $d_i^l$ and the output $z_j^{u_o}$ of the last encoder block $u_o$, namely,

$$\alpha_{ij}^l = \frac{\exp(d_i^l \cdot z_j^{u_o})}{\sum_{t=1}^{m} \exp(d_i^l \cdot z_t^{u_o})} \,. \tag{5}$$

The conditional input $c_i^l \in \mathbb{R}^d$ for the current decoder layer is computed as

$$c_i^l = \sum_{j=1}^{m} \alpha_{ij}^l (z_j^{u_o} + e_j) \,, \tag{6}$$

where $e_j$ is the input element embedding that can provide point information about a specific input element. Once $c_i^l$ has been computed, it is added to the output of the corresponding decoder layer $h_i^l$ and serves as a part of the input to $h_i^{l+1}$.

## 3.2 Topic-Aware Attention Mechanism

A topic model is a type of statistical model for discovering the abstract ideas or hidden semantic structures that occur in a collection of source articles. In this paper, we employ the topic model to acquire latent knowledge of documents and incorporate a topic-aware mechanism into the multi-step attention-based ConvS2S model, which is expected to bring prior knowledge for text summarization. Now we present the novel approach on how to incorporate the topic model into the basic ConvS2S framework via the joint attention mechanism and biased probability generation process.

### Topic Embeddings

The topic embeddings are obtained by classical topic models such as Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003]. During pre-training, we use LDA to assign topics to the input texts. The top $N$ non-universal words with the highest probabilities of each topic are chosen into the topic vocabulary $\boldsymbol{K}$. More details will be given in Section 4. While the vocabulary of texts is denoted as $\boldsymbol{V}$, we assume that $\boldsymbol{K} \subset \boldsymbol{V}$. Given an input sentence $\boldsymbol{x} = (x_1, \ldots, x_m)$, if a word $x_i \notin \boldsymbol{K}$, we embed it as before to attain $w_i$. However, if a word $x_i \in \boldsymbol{K}$, we can embed this topic word as $t_i \in \mathbb{R}^d$, which is a row in the topic embedding matrix $\mathcal{D}_{\text{topic}} \in \mathbb{R}^{K \times d}$, where $K$ is the size of topic vocabulary. The embedding matrix $\mathcal{D}_{\text{topic}}$ is normalized from the corresponding pre-trained topic distribution matrix, whose row is proportional to the number of times that each word is assigned to each topic. In this case, the positional embedding vectors are also added to the encoder and decoder elements, respectively, to obtain the final topic embeddings $\boldsymbol{r} = (r_1, \ldots, r_m)$ and $\boldsymbol{s} = (s_1, \ldots, s_n)$.

### Joint Attention

Again we take the decoder for illustration. Following the convolutional layer introduced before, we can obtain the convolution unit $i$ on the $l$-th layer in the decoder of topic level as $\tilde{h}_i^l \in \mathbb{R}^d$. Similar to (4), we have

$$\tilde{d}_i^l = \tilde{W}_d^l \tilde{h}_i^l + \tilde{b}_d^l + s_i \,. \tag{7}$$

We then incorporate the topic information into the model through a joint attention mechanism. During decoding, the joint attention weight $\beta_{ij}^l$ is given by

$$\beta_{ij}^l = \frac{\exp(\tilde{d}_i^l \cdot z_j^{u_o} + \tilde{d}_i^l \cdot z_j^{u_t})}{\sum_{t=1}^{m} \exp(\tilde{d}_i^l \cdot z_t^{u_o} + \tilde{d}_i^l \cdot z_t^{u_t})} \,, \tag{8}$$

where $z_j^{u_t}$ is the output of the last topic-level encoder block $u_t$. Then the conditional input $\tilde{c}_i^l \in \mathbb{R}^d$ is computed as

$$\tilde{c}_i^l = \sum_{j=1}^{m} \beta_{ij}^l (z_j^{u_t} + r_j) \,. \tag{9}$$

In the joint attention mechanism, both $\tilde{c}_i^l$ and $c_i^l$ are added to the output of the corresponding decoder layer $\tilde{h}_i^l$ and are a part of the input to $\tilde{h}_i^{l+1}$.

### Biased Probability Generation

Finally, we compute a distribution over all possible next target elements $y_{i+1} \in \mathbb{R}^T$, namely

$$p_\theta(y_{i+1}) := p(y_{i+1} | y_1, \ldots, y_i, \boldsymbol{x}) \in \mathbb{R}^T \,, \tag{10}$$

by transforming the top word-level decoder outputs $h^{L_o}$ and topic-level decoder outputs $\tilde{h}^{L_t}$ via a linear layer $\Psi(\cdot)$, which is computed by

$$\Psi(h) = W_o h + b_o \,, \tag{11}$$

where $W_o \in \mathbb{R}^{T \times d}$ and $b_o \in \mathbb{R}^T$ are the parameters to be learned. Then the biased generation distribution is given as

$$p_\theta(y_{i+1}) = \frac{1}{Z} \left[ \exp\left( \Psi(h_i^{L_o}) \right) + \exp\left( \Psi(\tilde{h}_i^{L_t}) \right) \otimes I_{\{w \in \boldsymbol{K}\}} \right] \,, \tag{12}$$

where $Z$ is the normalizer, $h_i^{L_o}$ and $\tilde{h}_i^{L_t}$ denote the $i$-th top decoder outputs of word and topic, respectively, and $I$ is the one-hot indicator vector of each candidate word $w$ in $y_{i+1}$. When the candidate word $w$ is a topic word, we bias the generation distribution by the topic information. Otherwise, we ignore the topic part. To some extent, the complexity of the search space is reduced by introducing the topic bias since important words are more likely to be generated directly.

### 3.3 Reinforcement Learning

The teacher forcing algorithm [Williams and Zipser, 1989] aims to minimize the maximum-likelihood loss at each decoding step, namely,

$$L_{\text{ml}} = -\sum_{i=1}^{T} \log p_\theta(y_i^* | y_1^*, y_2^*, \ldots, y_{i-1}^*, \boldsymbol{x}), \tag{13}$$

where $\boldsymbol{x}$ refers to an input sequence and $y^* = (y_1^*, y_2^*, \ldots, y_T^*)$ is the corresponding ground-truth output sequence.

Minimizing the objective in Eq. (13) often produces sub-optimal results with respect to the evaluation metrics, such as ROUGE which measures the sentence-level accuracy of the generated summaries. The sub-optimality is related to the problem called *exposure bias* [Ranzato *et al.*, 2015], which is caused by only exposing a model to the distribution of training data instead of its own distribution. During the training process, models are fed by ground-truth output sequences to predict the next word, whereas during inference they generate the next word given the predicted words as inputs. Therefore, in the test process, the error of each step accumulates and leads to the deterioration of performance.

The second reason for sub-optimality comes from the flexibility of summaries. The maximum-likelihood objective rewards models that can predict exactly the same summaries as references while penalizing those that produce different texts even though they are semantically similar. Providing multiple reference summaries is helpful yet insufficient since there are alternatives to rephrase a given summary. Therefore, minimizing the objective in Eq. (13) neglects the intrinsic property of summarization. ROUGE, on the other hand, provides more flexible evaluation, encouraging models to focus more on semantic meanings than on word-level correspondences.

In order to address such issues, we utilize self-critical sequence training (SCST) [Rennie *et al.*, 2016], a policy gradient algorithm for reinforcement learning, to directly maximize the non-differentiable ROUGE metric. During reinforcement learning, we generate two output sequences given the input sequence $\boldsymbol{x}$. The first sequence $\hat{y}$ is obtained by greedily selecting words that maximize the output probability distribution, and the other output sequence $y^s$ is generated by sampling from the distribution. After obtaining ROUGE scores of both sequences as our rewards, i.e., $r(y^s)$ and $r(\hat{y})$, we minimize the reinforcement loss

$$L_{\text{rl}} = -(r(y^s) - r(\hat{y})) \log p_\theta(y^s), \tag{14}$$

and update model parameters by gradient descent techniques.

With SCST, we can directly optimize the discrete evaluation metric. In addition, the "self-critical" test-time estimate of the reward $r(\hat{y})$ provides a simple yet effective baseline and improves training/test time consistency. Since during learning we set the baseline of the REINFORCE algorithm as the reward obtained by the current model in the test-time inference, the SCST exposes the model to its own distribution and encourages it to produce the sequence output $\hat{y}$ with a high ROUGE score, avoiding the exposure bias issue and thus improving the test performance.

| No. | Topic Words |
|---|---|
| 1 | prime, minister, talks, leader, elections, visit |
| 2 | bird, flu, officials, opens, poultry, die |
| 3 | trade, free, EU, army, urges, ban |
| 4 | Bush, world, talks, foreign, investment, markets |
| 5 | world, Malaysia, Thailand, meet, Vietnam, U.S. |

Table 1: Examples of topic words for the Gigaword corpus.

## 4 Experimental Setup

### 4.1 Datasets

In this paper, we consider three datasets to evaluate the performance of different methods in the abstractive text summarization task. First, we consider the annotated Gigaword corpus [Graff and Cieri, 2003] preprocessed identically to [Rush *et al.*, 2015], which leads to around 3.8M training samples, 190K validation samples and 1951 test samples for evaluation. The input summary pairs consist of the headline and the first sentence of the source articles. We also evaluate various models on the DUC-2004 test set[1] [Over *et al.*, 2007]. The dataset is a standard summarization evaluation set, which consists of 500 news articles. Unlike the Gigaword corpus, each article in DUC-2004 is paired with four human-generated reference summaries, which makes the evaluation more objective. The last dataset for evaluation is a large corpus of Chinese short text summarization (LCSTS) dataset [Hu *et al.*, 2015] collected and constructed from the Chinese microblogging website Sina Weibo. Following the setting in the original paper, we use the first part of LCSTS dataset for training, which contains 2.4M text-summary pairs, and choose 725 pairs from the last part with high annotation scores as our test set.

### 4.2 Topic Information

The classical LDA with Gibbs Sampling technique is used to pre-train the corpus for topic embedding initialization and provide candidates for the biased probability generation process. The topic embedding values are normalized to a distribution with mean zero and variance of 0.1 for adaption to the neural network structure. In this paper, we pick top $N = 200$ words with the highest probabilities in each topic to obtain the topic word set. Note that the universal words are filtered out during pre-training. Randomly selected examples of topic words of the Gigaword corpus are presented in Table 1.

### 4.3 Model Parameters and Optimization

We employ six convolutional layers for both the encoder and decoder. All embeddings, including the initialized embed-

[1] `http://duc.nist.gov/data.html`

| | RG-1 (F) | RG-2 (F) | RG-L (F) |
|---|---|---|---|
| ABS [Rush *et al.*, 2015] | 29.55 | 11.32 | 26.42 |
| ABS+ [Rush *et al.*, 2015] | 29.76 | 11.88 | 26.96 |
| RAS-Elman [Chopra *et al.*, 2016] | 33.78 | 15.97 | 31.15 |
| words-lvt5k-1sent [Nallapati *et al.*, 2016b] | 35.30 | 16.64 | 32.62 |
| RNN+MLE [Shen *et al.*, 2016] | 32.67 | 15.23 | 30.56 |
| RNN+MRT [Shen *et al.*, 2016] | 36.54 | 16.59 | 33.44 |
| SEASS(beam) [Zhou *et al.*, 2017] | 36.15 | 17.54 | 33.63 |
| ConvS2S [Gehring *et al.*, 2017] | 35.88 | 17.48 | 33.29 |
| Topic-ConvS2S | 36.38 | 17.96 | 34.05 |
| Reinforced-ConvS2S | 36.30 | 17.64 | 33.90 |
| Reinforced-Topic-ConvS2S | **36.92** | **18.29** | **34.58** |

Table 2: Accuracy on the Gigaword corpus in terms of the full-length ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L). Best performance on each score is displayed in **boldface**.

| | RG-1 (F) | RG-2 (F) | RG-L (F) |
|---|---|---|---|
| ABS (beam) [Rush *et al.*, 2015] | 37.41 | 15.87 | 34.70 |
| s2s+att (greedy) [Zhou *et al.*, 2017] | 42.41 | 20.76 | 39.84 |
| s2s+att (beam) [Zhou *et al.*, 2017] | 43.76 | 22.28 | 41.14 |
| SEASS (greedy) [Zhou *et al.*, 2017] | 45.27 | 22.88 | 42.20 |
| SEASS (beam) [Zhou *et al.*, 2017] | 46.86 | 24.58 | 43.53 |
| Topic-ConvS2S | 46.80 | 24.74 | 43.92 |
| Reinforced-ConvS2S | 46.68 | 24.22 | 43.76 |
| Reinforced-Topic-ConvS2S | **46.92** | **24.83** | **44.04** |

Table 3: Accuracy on the internal test set of Gigaword corpus in terms of the full-length RG-1, RG-2, and RG-L. Best performance on each score is displayed in **boldface**.

ding and the output produced by the decoder before the final linear layer, have a dimensionality of 256. We also adopt the same dimensionality for the size of linear layer mapping between hidden and embedding states. We use a learning rate of 0.25 and reduce it by a decay rate of 0.1 once the validation ROUGE score stops increasing after each epoch until the learning rate falls below $10^{-5}$. We first train the basic topic-aware convolutional model with respect to a standard maximum likelihood objective, and then switch to further minimize a mixed training objective [Paulus *et al.*, 2017], incorporating the reinforcement learning objective $L_{\text{rl}}$ and the original maximum likelihood $L_{\text{ml}}$, which is given as

$$L_{\text{mixed}} = \lambda L_{\text{rl}} + (1 - \lambda) L_{\text{ml}}, \tag{15}$$

where the scaling factor $\lambda$ is set to be $0.99$ in our experiments. Moreover, we choose the ROUGE-L metric as the reinforcement reward function. Nesterov's accelerated gradient method [Sutskever *et al.*, 2013] is used for training, with the mini-batch size of 32 and the learning rate of $0.0001$. All models are implemented in PyTorch [Paszke *et al.*, 2017] and trained on a single Tesla M40 GPU.

## 5 Results and Analysis

We follow the existing work and adopt the ROUGE metric [Lin, 2004] for evaluation.

### 5.1 Gigaword Corpus

We demonstrate the effectiveness of our proposed model via a step-by-step justification. First, the basic ConvS2S structure with topic-aware model or reinforcement learning is tested, respectively. Then we combine the two to show the performance of our Reinforced-Topic-ConvS2S model. We report the full-length F-1 scores of the ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L) metrics and compare the results with various neural abstractive summarization methods, which are presented in Table 2. The ABS and ABS+ models are attention-based neural models for text summarization. The RAS-Elman model introduces a conditional RNN, in which the conditioner is provided by a convolutional attention-based encoder. The words-lvt5k-1sent model is also a RNN-based attention model which implements a large-vocabulary trick. Besides, RNN+MRT employs the minimum risk training strategy which directly optimizes model parameters in sentence level with respect to the evaluation metrics. SEASS (beam) extends the sequence-to-sequence framework with a selective encoding model. The results have demonstrated that both the topic-aware module and the reinforcement learning process can improve the accuracy on text summarization. Moreover, our proposed model exhibits best scores of RG-1, RG-2 and RG-L.

In addition, [Zhou *et al.*, 2017] further selects 2000 pairs of summaries as an internal test set of Gigaword. We also evaluate our proposed model on this set and present the results in Table 3. Again, our proposed model achieves the best performance in terms of all the three ROUGE scores.

To further demonstrate the improvement of readability and diversity by the topic information, we also present some qualitative results by randomly extracting several summaries from test. We compare the reference summaries to the summaries generated by our proposed model with or without topic-aware mechanism. The examples are presented in Table 4. We can observe that when the topic model is adopted, it can generate some accurately delivered topic words which are not in

| **Examples of summaries** |
|---|
| **D**: the sri lankan government on wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country.<br>**R**: sri lanka closes schools as war escalates<br>**OR**: sri lanka closes schools with immediate effect<br>**OT**: sri lanka closes schools in wake of military attacks |
| **D**: a us citizen who spied for communist east germany was given a suspended jail sentence of ## months here friday.<br>**R**: us citizen who spied for east germans given suspended sentence<br>**OR**: us man gets suspended jail term for communist spying<br>**OT**: us man jailed for espionage |
| **D**: malaysian prime minister mahathir mohamad indicated he would soon relinquish control of the ruling party to his deputy anwar ibrahim.<br>**R**: mahathir wants leadership change to be smooth<br>**OR**: malaysia's mahathir to relinquish control of ruling party<br>**OT**: malaysia's mahathir to submit control of ruling party |
| **D**: a french crocodile farm said it had stepped up efforts to breed one of the world's most endangered species, the indian UNK, with the hope of ultimately returning animals to their habitat in south asia.<br>**R**: french farm offers hope for endangered asian crocs UNK picture<br>**OR**: french crocodile farm steps up efforts to breed endangered species<br>**OT**: french crocodile farm says steps up efforts to save endangered species |

Table 4: Examples of generated summaries on the Gigaword corpus. **D**: source document, **R**: reference summary, **OR**: output of the **R**einforced-ConvS2S model, **OT**: output of the Reinforced-**T**opic-ConvS2S model. The words marked in blue are topic words not in the reference summaries. The words marked in red are topic words neither in the reference summaries nor in the source documents.

| | RG-1 (R) | RG-2 (R) | RG-L (R) |
|---|---|---|---|
| ABS [Rush *et al.*, 2015] | 26.55 | 7.06 | 22.05 |
| ABS+ [Rush *et al.*, 2015] | 28.18 | 8.49 | 23.81 |
| RAS-Elman [Chopra *et al.*, 2016] | 28.97 | 8.26 | 24.06 |
| words-lvt5k-1sent [Nallapati *et al.*, 2016b] | 28.61 | 9.42 | 25.24 |
| RNN+MLE [Shen *et al.*, 2016] | 24.92 | 8.60 | 22.25 |
| RNN+MRT [Shen *et al.*, 2016] | 30.41 | **10.87** | 26.79 |
| SEASS (beam) [Zhou *et al.*, 2017] | 29.21 | 9.56 | 25.51 |
| ConvS2S [Gehring *et al.*, 2017] | 30.44 | 10.84 | 26.90 |
| Topic-ConvS2S | 31.08 | 10.82 | 27.61 |
| Reinforced-ConvS2S | 30.74 | 10.68 | 27.09 |
| Reinforced-Topic-ConvS2S | **31.15** | 10.85 | **27.68** |

Table 5: Accuracy on the DUC-2004 dataset in terms of the recall-only RG-1, RG-2, and RG-L. Best performance on each score is displayed in **boldface**.

| | RG-1 (F) | RG-2 (F) | RG-L (F) |
|---|---|---|---|
| character-based preprocessing | | | |
| RNN context [Hu *et al.*, 2015] | 29.90 | 17.40 | 27.20 |
| COPYNET [Gu *et al.*, 2016] | 34.40 | 21.60 | 31.30 |
| RNN+MLE [Shen *et al.*, 2016] | 34.90 | 23.30 | 32.70 |
| RNN+MRT [Shen *et al.*, 2016] | 38.20 | 25.20 | 35.40 |
| word-based preprocessing | | | |
| RNN context [Hu *et al.*, 2015] | 26.80 | 16.10 | 24.10 |
| COPYNET [Gu *et al.*, 2016] | 35.00 | 22.30 | 32.00 |
| Topic-ConvS2S | 38.94/44.42 | 21.05/32.65 | 37.03/42.09 |
| Reinforced-ConvS2S | 36.68/42.61 | 18.69/29.79 | 34.85/40.03 |
| Reinforced-Topic-ConvS2S | 39.93/45.12 | 21.58/33.08 | 37.92/42.68 |

Table 6: Accuracy on the LCSTS dataset in terms of the full-length RG-1, RG-2, and RG-L. In last three rows, the word-level ROUGE scores are presented on the left and the character-level on the right.

the reference summaries or the original texts. It is believed that the joint learning with a pre-trained topic model can offer more insightful information and improve the diversity and readability for the summarization.

### 5.2 DUC-2004 Dataset

Since the DUC-2004 dataset is an evaluation-only dataset, we train the models on the Gigaword corpus first and then evaluate their performance on the DUC dataset. As the standard practice, we report the recall-based scores of the RG-1, RG-2, and RG-L metrics in this experiment, which are given in Table 5. From Table 5 we can observe that the proposed Reinforced-Topic-ConvS2S model achieves best scores of the RG-1 and RG-L metrics, and is comparable on the RG-2 score. Due to the similarity of the two datasets, we do not provide qualitative summarization examples in this experiment.

### 5.3 LCSTS Dataset

We now consider the abstractive summarization task on the LCSTS dataset. Since this is a large-scale Chinese dataset, suitable data preprocessing approaches should be proposed first. Basically, there are two approaches to preprocessing the Chinese dataset: character-based and word-based. The former takes each Chinese character as the input, while the latter splits an input sentence into Chinese words. [Hu *et al.*, 2015] provides a baseline result on both preprocessing approaches. [Shen *et al.*, 2016] also conducts experiments on the LCSTS corpus based on character inputs. [Gu *et al.*, 2016] proposes a neural model, the COPYNET, with both character-based and word-based preprocessing by incorporating the copying mechanism into the sequence-to-sequence framework. In this work, we adopt the word-based approach as we believe that in the case of Chinese, words are more relevant to latent knowledge of documents than characters are.

Since the standard ROUGE package[2] is usually used to evaluate the English summaries, directly employing the package to evaluate Chinese summaries would yield underrated results. In order to evaluate the summarization on the LCSTS dataset, we follow the suggestion of [Hu *et al.*, 2015] by mapping Chinese words/characters to numerical IDs, on which we then perform the ROUGE evaluation. Since not all previous work explicitly mentioned whether word-based or character-based ROUGE metrics were reported, we evaluate our proposed model with both metrics in order to obtain a comprehensive comparison. The results of both scores are presented in Table 6, which are displayed as word-based score/character-based score.

From the results shown in Table 6, we see that one can always achieve higher ROUGE scores in the character level than that based on Chinese words by our proposed model. We can also observe that the character-based results of our Reinforced-Topic-ConvS2S model outperforms every other method. Regarding to word-based ROUGE scores, our model obtains the best performance in terms of RG-1 and RG-L metrics. However, our best model does not achieve a good RG-2 score as its RG-1 and RG-L scores. We suspect that it may be partly caused by the biased probability generation mechanism that influences word order, which requires further studies.

In addition to ROUGE scores, we also present some randomly picked examples of generated summaries in Table 7. The original examples (in Chinese) are shown and all the texts are carefully translated to English for the convenience of reading. The examples demonstrate that the topic-aware mechanism can also improve the diversity in Chinese summarization tasks.

## 6 Conclusion and Future Work

In this work, we propose a topic-aware ConvS2S model with reinforcement learning for abstractive text summarization. It is demonstrated that the new topic-aware attention mechanism introduces some high-level contextual information for summarization. The performance of the proposed model advances state-of-the-art methods on various benchmark datasets. In addition, our model can produce summaries with better informativeness, coherence, and diversity.

Note that the experiments in this work are mainly based on the sentence summarization. In the future, we aim to evaluate our model on the datasets where the source texts can be long paragraphs or multi-documents. Moreover, we also note that how to evaluate the performance on Chinese summaries remains an open problem. It is also of great interest to study on this subject in the future.

## Acknowledgements

Qiang Du is supported in part by the US NSF TRIPODs project through CCF-170483.

[2]`http://www.berouge.com/Pages/default.aspx`

| **Examples of summaries** |
|---|
| $D$: 根据#### 年# 月# 日国家发改委等部门联合发布的《关于进一步做好新能源汽车推广应用工作的通知》，#### 年的补贴金额相比#### 年将降低##% 。（分享自@ 电动邦） |
| $D$: According to the notice *On the further promotion and application of new energy vehicles*, jointly released by the National Development and Reform Commission and other departments on ##/##/#### (date), the compensation of #### (year) will be reduced by ##% compared to #### (year). (reposted from @electric_nation) |
| $R$: 补贴金额再缩水#### 年新能源车政策解读 |
| $R$: The compensation has been reduced again: #### (year) policy analysis of new energy automobiles |
| $OR$: #### 年新能源汽车推广应用工作的通知 |
| $OR$: #### (year) notice on the promotion and application of new energy vehicles |
| $OT$: 国家发改委 发文 进一步做好 新能源汽车 推广应用工作 |
| $OT$: The National Development and Reform Commission issued a policy on further promotion and application of new energy vehicles |
| $D$: 成都市软件和信息技术服务业近年来一直保持快速增长势头，稳居中西部城市之首，已成为我国西部"硅谷"。《#### 年度成都市软件和信息技术服务产业发展报告》日前发布……详情请见: @ 成都日报@ 成都发布 |
| $D$: In recent years, the service industry of software and information technology in Chengdu has been growing rapidly, ranking first among the cities in Midwest China. Chengdu has become China's western "Silicon Valley". The *#### (year) Annual Chengdu Software and Information Technology Service Industry Development Report* has been released recently ... ... see details: @ Chengdu_Daily @ Chengdu_release |
| $R$: 成都倾力打造西部"硅谷" |
| $R$: Chengdu makes every effort to build the western "Silicon Valley" |
| $OR$: 成都软件 和信息技术服务业发展报告发布 |
| $OR$: The report of Chengdu software and information technology service industry development has been released |
| $OT$: 成都软件 和信息技术服务业 跃居 西部"硅谷" |
| $OT$: The service industry of software and information technology in Chengdu rockets to make it the western "Silicon Valley" |
| $D$: 新疆独特的区位优势，使其成为"一带一路"战略重要一环。记者从新疆发改委获悉，库尔勒至格尔木铁路先期开工段已进入招投标阶段，计划#### 年## 月中旬正式开工建设。#### 年计划完成投资## 亿元。 |
| $D$: Xinjiang's unique geographical advantages make it an important part of *The Belt and Road* strategy. The reporter learned from the Xinjiang Development and Reform Commission that the initial railway construction project from Korla to Golmud had been on tendering procedure. The project was scheduled to officially launch in mid ## (month) of #### (year) and attract the investment of ## billion yuan by #### (year). |
| $R$: "一带一路"战略惠及新疆<unk>, 铁路年底开建 |
| $R$: *The Belt and Road* strategy benefits Xinjiang <unk> and the railway construction starts by the end of #### (year) |
| $OR$: 新疆<unk> 至格尔木铁路计划#### 年开建 |
| $OR$: The railway from <unk> to Golmud is scheduled to start construction in #### (year) |
| $OT$: 库尔勒至格尔木铁路拟 ## 月开工建设 |
| $OT$: The railway construction project from Korla to Golmud is planned to launch in ## (month) |
| $D$: 昨日，商报记者从代表国内婚尚产业"风向标"的上海国际婚纱摄影器材展览会上了解到，部分商家开始将婚庆布置、婚礼流程、形式交给新人决定以迎合## 后新人的需求。此次展览会的规模超过# 万平方米，吸引参展企业超过### 家。 |
| $D$: The day before, the reporters of *Commercial News* learned from the Shanghai International Wedding Photographic Equipment Exhibition, which has been leading and defining the domestic wedding industry, that some companies began to cater for the requirements of ##s-generation newly married couples by self-decided wedding decoration, wedding process and forms. The venue of the exhibition is more than # tens of thousands square meters, attracting more than ### exhibitors. |
| $R$: 婚庆"私人定制"受## 后新人追捧 |
| $R$: The personalized wedding is admired by ##s-generation newly married couples |
| $OR$: 上海 国际婚纱摄影 器材展览会举行 |
| $OR$: Shanghai International Wedding Photographic Equipment Exhibition was held |
| $OT$: 上海 国际婚纱摄影 器材展览会昨 举行 |
| $OT$: Shanghai International Wedding Photographic Equipment Exhibition was held yesterday |

Table 7: Examples of generated summaries on the LCSTS dataset. **D**: source document, **R**: reference summary, **OR**: output of the **R**einforced-ConvS2S model, **OT**: output of the Reinforced-**T**opic-ConvS2S model. The words marked in blue are topic words not in the reference summaries. The words marked in red are topic words neither in the reference summaries nor in the source documents. All the texts are carefully *translated from Chinese*.

## References


[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Barzilay and McKeown, 2005] Regina Barzilay and Kathleen R McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328, 2005.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

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

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

[Dauphin *et al.*, 2016] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083*, 2016.

[Gehring *et al.*, 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

[Graff and Cieri, 2003] David Graff and C Cieri. English gigaword corpus. *Linguistic Data Consortium*, 2003.

[Gu *et al.*, 2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

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

[Hu *et al.*, 2015] Baotian Hu, Qingcai Chen, and Fangze Zhu. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*, 2015.

[Kraaij *et al.*, 2002] Wessel Kraaij, Martijn Spitters, and Anette Hulth. Headline extraction based on a combination of uni-and multidocument summarization techniques. In *Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002). ACL*, 2002.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

[Nallapati *et al.*, 2016a] Ramesh Nallapati, Bing Xiang, and Bowen Zhou. Sequence-to-sequence rnns for text summarization. 2016.

[Nallapati *et al.*, 2016b] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[Neto *et al.*, 2002] Joel Neto, Alex Freitas, and Celso Kaestner. Automatic text summarization using a machine learning approach. *Advances in Artificial Intelligence*, pages 205–215, 2002.

[Over *et al.*, 2007] Paul Over, Hoa Dang, and Donna Harman. Duc in context. *Information Processing & Management*, 43(6):1506–1520, 2007.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, and Soumith Chintala. Pytorch, 2017.

[Paulus *et al.*, 2017] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304, 2017.

[Ranzato *et al.*, 2015] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[Rennie *et al.*, 2016] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[Rush *et al.*, 2015] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[Shen *et al.*, 2016] Shiqi Shen, Yu Zhao, Zhiyuan Liu, Maosong Sun, et al. Neural headline generation with sentence-wise optimization. *arXiv preprint arXiv:1604.01904*, 2016.

[Sutskever *et al.*, 2013] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

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

[Williams and Zipser, 1989] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, June 1989.

[Xing *et al.*, 2017] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *AAAI*, pages 3351–3357, 2017.

[Zhou *et al.*, 2017] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073*, 2017.