A Neural Abstractive Summarization Model Guided with Topic Sentences

Kai Chen[1], Baotian Hu[2], and Qingcai Chen[1(✉)] Hong Yu[2]

[1] Harbin Institute of Technology (Shenzhen), HIT Campus, Shenzhen University Town, Shenzhen, GuangDong, China
ckai.hit@gmail.com, qingcai.chen@hit.edu.cn
[2] University of Massachusetts Lowell, 220 Pawtucket St, Lowell, MA 01854, USA
baotianchina@gmail.com, Hong_Yu@uml.edu

Abstract. The topic information plays a critical role in comprehending the main idea of a document. However, most of the recent neural sequence to sequence models fail to utilize the topic information. In this paper, we propose a novel neural abstractive summarization model guided with topic sentences, named NASTS. We first use the extractive summarization system to extract topic sentences from the document. And then, a well designed weighted Attention-over-Attention mechanism is exploited to integrate the topic information into NASTS. In this way, NASTS can effectively take advantages of extractive and abstractive summarization methods. Experimental results on the benchmark dataset CNN/Daily Mail demonstrate the superiority of our proposed model.

Keywords: Abstractive summarization · Topic sentences · Attention-over-Attention mechanism.

1 Introduction

Abstractive text summarization is the process of generating a piece of brief text by distilling the most important information from the given long documents [1, 2]. Recently, the neural Seq2Seq model with attention has become the dominant approach [3, 4], due to its powerful ability of representation and generation. Most competitive models [5–7] use the representation of the previously generated word to compute the current attention distribution on the input document. Intuitively, the generated word only contains partial topic information of the document. Especially, at the beginning of the generation, there is barely any information can be used to guide the generation.

When one writes a summary, the topic can give a hint about which part of the document is essential. Suppose if the topic of a document is about *"Donald Trump and Xi Jinping declare trade truce at G20,"* the summarization should pay more attention to the content of the talk between Xi and Trump instead of the background (e.g., issues related to G20). In contrast to abstractive summarization, extractive summarization approaches [8, 9] aim to extract topic sentences as the summary. Topic sentences represent the main idea of a document. Intuitively, topic sentences can be used to guide abstractive summarization.

The contributions of our work are as follows:

- In this paper, we present a novel approach to integrating the topic information into abstractive summarization. The topic information is represented as the sentences extracted via extractive summarization methods. Our model can combine both the advantages of extractive and abstractive summarization.
- We proposed a novel weighted Attention-over-Attention mechanism to force the decoder to generate text close to the main topic of the document. Automatic evaluation experiments show that our model outperforms the strong baseline pointer-network [10] with a large margin on CNN/Daily Mail dataset.

2 Related Work

The predominant abstractive summarization models are based on RNNs with attention [11, 12], which directly process the article with an encoder and produce the summary by an attentional decoder. However, these methods ignore topic information which has been proved to be helpful for summarization [13, 14]. Wang et al. [15] first incorporate the topic information for abstractive summarization, but a pre-trained topic model is needed for topic embedding initialization. Here, we construct a topic encoder which could be integrated into the pipeline summarization system to mining the topic information of the article.

Attention-over-Attention mechanism (AoA mechanism) was proposed by Cui et al. [16] to merge sentence-level and word-level attention in reading comprehension task. Inspired by it, we propose a weighted Attention-over-Attention mechanism (WAoA mechanism) for abstractive summarization. Whereas our weighted model not only

integrates topic attention and article attention, but also dynamically updates the encoder hidden states to ensure they are more relevant to the topic.

Extract-then-abstract methods which take advantage of extractive and abstractive summarization have been widely studied and achieve better performance. Hsu et al.[17] propose a novel inconsistency loss function to penalize the inconsistency between sentence-level and word-level of attentions, but need to optimize four loss functions(extractor loss, abstractor loss, coverage loss, and inconsistency loss). Chen et al. [18] first select salient sentences via an RNN-extractor and employ an abstractor which was guided with reinforcement learning to rewrite them, then rerank the generated summaries to avoid repetition. In contrast, NASTS utilizing topic encoder and weighted attention mechanism is more concise and easier to be reduplicated.

3 Method

In this section, we present the detailed architecture of the proposed NASTS, which consists of an attention-based topic encoder, a weighted Attention-over-Attention article encoder, and a pointer-network decoder [10]. Figure 1 illustrates the NASTS architecture.

3.1 Attention-based Topic Encoder

The topic encoder is a type of statistical mechanism for discovering salient information of the article. The input of the topic encoder is a token list of topic sentences $T = [t_1, t_2, ..., t_i, ...]$ where i is the word index. We apply a Bi-LSTM encoder to abstract topic sentences, then at decode time t, we can obtain the attention distribution a'^t as below:

$$e_i'^t = v^{'t} tanh(W_s' s_t + W_h h_i' + b_{attn}') \tag{1}$$

$$a'^t = softmax(e'^t) \tag{2}$$

where $h_i^{'}$ represents the encoder hidden state of the ith word in topic sentences, s_t is the decoder state, $v_T^{'}$, $W_s^{'}$, W_h and $b_{attn}^{'}$ are learnable parameters. And calculated the topic vector tv_t by:

$$tv_t = \sum\nolimits_i a_i'^t h_i' \tag{3}$$

3.2 Weighted Attentional Article Encoder

Here we introduce the article encoder with the weighted Attention-over-Attention mechanism. The tokens of the sequence $X = [w_1, w_2, ..., w_i, ...]$ are fed one-by-one into the Bi-LSTM encoder and a sequence of encoder hidden states h_i are produced. Then, the topic vector tv_t could be used to act on the attention process of the article in two steps.

Firstly, we employ a weighted mechanism to model the selection process for the encoder hidden states of the article which are fixed in conventional attention-based Seq2Seq models. For each time step t, the topic-weighted hidden state $\hat{h}_i$ can be computed with tv_t and h_i as inputs:

$$\hat{h}_i = tv_t \odot h_i \tag{4}$$

where $\odot$ is element-wise multiplication. After the weighted mechanism, we obtain another sequence of vectors $[\hat{h}_1, \hat{h}_2, ..., \hat{h}_n]$ which was seen as topic related word representations.

Then, we incorporate the topic vector tv_t into the model through the AoA mechanism. During decoding, the article attention distribution a_t is given by:

$$e_i^t = v^T tanh(W_s s_t + W_{\hat{h}} \hat{h}_i + W_{tv} tv_t + b_{attn}) \tag{5}$$

$$a^t = softmax(e^t) \tag{6}$$

where v^T, W's and b_{attn} are learnable parameter. Then the context vector c_t^* will be computed with the article attention distribution like equation (3).

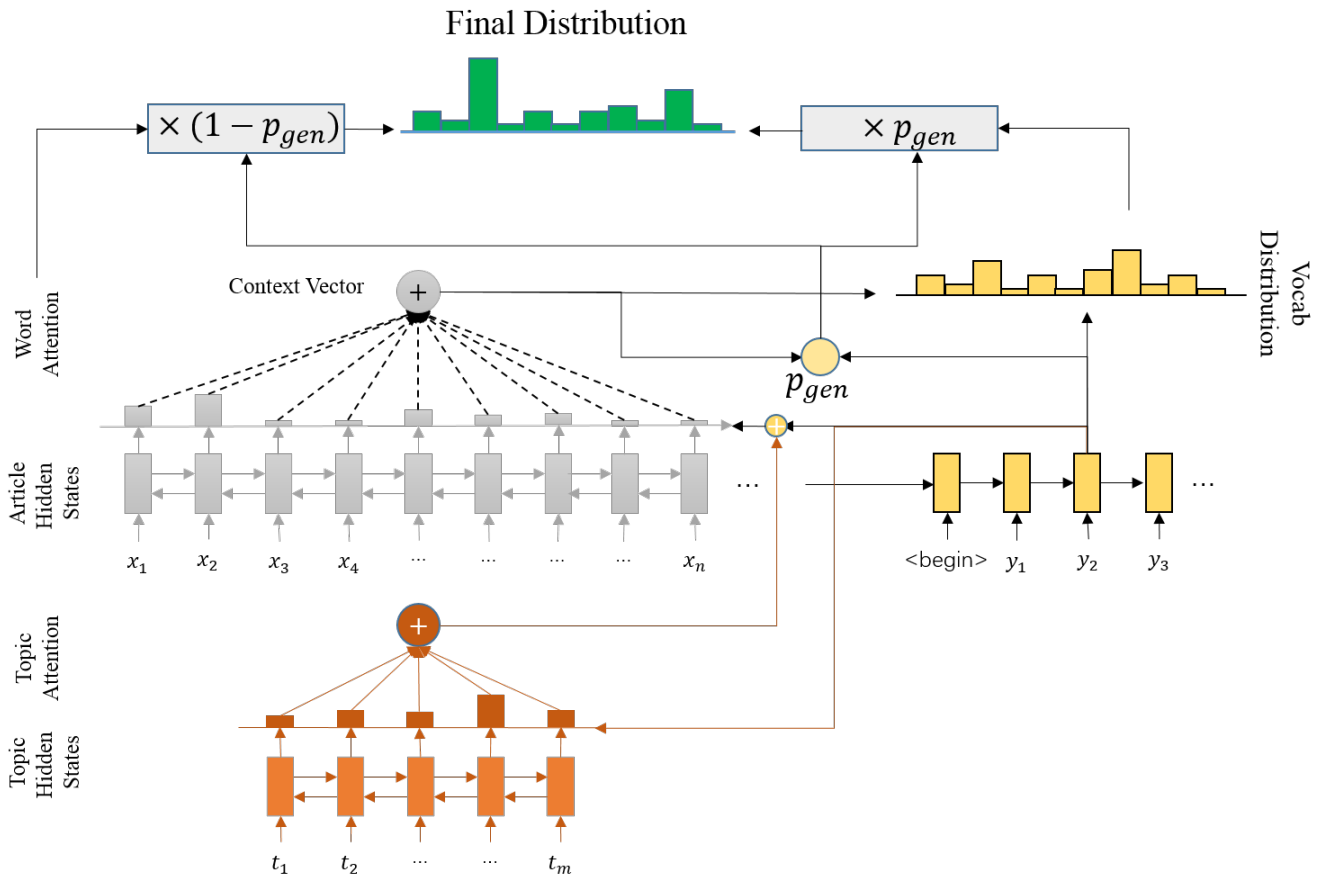


Fig. 1. Architecture of NASTS. Topic sentences and article are encoded by two associated encoders. For each decoder step, the topic vector and previous decoder state jointly participate in the attention process of article. Finally, the target sequence is produced according to the final distribution.

3.3 Summary Decoder

In decoder part, we use pointer-network decoder [10]. Following topic encoder and article encoder introduced before, we can obtain the vocabulary distribution P_{vocab} by learnable hyper-parameters $V^{'}$, V, b and $b^{'}$ as follow:

$$P_{vocab} = softmax(V'(V[s_t, c_t^*, tv_t] + b) + b') \tag{7}$$

And the generation probability p_g can be calculated as below:

$$p_g = \sigma(W_{c^*}^T c_t^* + W_{tv}^T tv_t + W_s^T s_t + W_x^T x_t + b_g) \tag{8}$$

where x_t represents the decode input, W's and b_g are parameters to be learned and σ represents the sigmoid function. Then p_g is used to further compute a weighted sum of the probability of vocabulary distribution and article attention distribution. Additionally, We adopt coverage mechanism [10, 19] to eliminate the repetition problem. At each decoding step t, we compute the coverage vector $cov_t = \sum_{t'=1}^{t-1} \hat{a}^{t'}$. Then the coverage vector will be used to calculate word attention, changing equation (5) to:

$$e_i^t = V^T tanh(W_s s_t + W_h h_t + W_{tv} tv_t + W_{cov} cov_t + b_{attn}) \tag{9}$$

Moreover, the coverage loss $covloss_t = \sum_i min(a_i^t, cov_i^t)$ will be calculated to penalize the repetition. Finnally, the objective function changes from cross-entropy loss to a weighted sum of coverage loss and cross-entropy.

4 Experiment

4.1 Dataset

We evaluate our models on the non-anonymized version of the CNN/Daily Mail dataset [1, 9, 10], which includes 287,226 training pairs, 13,368 validation pairs, and 11,490 test pairs. To compare with previous work, we adopt

the same evaluation metrics as in [17, 18], full-length F1 of ROUGE-1, ROUGE-2, and ROUGE-L to evaluate our model.

To train NASTS, we need topic sentences to feed into the topic encoder. To address this problem, a greedy algorithm similar to [20] is used to extract topic sentences. while testing, we adopt four version extractors to extract topic sentences: Coherent Extractor (**Coh-Ext**, 41.25 ROUGE-1 score in extractive summarization tasks on CNN/Daily Mail) and Sentence-Level Extractor (**SL-Ext**, 41.47 ROUGE-1 score) are two state-of-the-art reinforced extractive summarization models [9, 18]; and **Human-Ext**-2 and **Human-Ext**-3 means topic sentences are extracted manually with maximum count 2 and 3, respectively.

4.2 Setting

The dimension of randomly initialized word embedding and LSTM hidden state is 256. The vocabulary size is 50k, and the batch size is 16. Adagrad optimizer [21] is used with learning rate 0.15 and an initial accumulator value of 0.1. We use gradient clipping with a maximum gradient norm of 2. For all training and testing time, we truncate the article to 400 tokens, topic sentences to 120 tokens, and limit the length of output to 100 tokens. At testing time, we use beam search (size=16) to generate summaries. Further, we train NASTS without coverage about 227k iterations (12.7 epochs). Finally, we added coverage mechanism and trained 4k iterations more.

4.3 Result

We compare our model with two kinds of robust baseline models. **P**ipeline **A**bstracive **S**ummarization **M**odels (**PASM**s) which include: See et al. [10] employed pointer-generator and coverage mechanism; Fan et al. [22] added controlling parameters to adapt the summary and Liu et al. [23] achieve high readability score on human evaluation using generative adversarial networks. Extract-then-abstract methods, Hsu [17] constructed an end-to-end training architecture with four loss functions, and Chen et al. [18] proposed a novel sentence-level reinforcement learning model to rewrite extracted summaries.

From Table 1, we can see that NASTS outperforms previous PASMs and achieves higher ROUGE-1 score compared with previous extract-then-abstract methods. From Row 7 to Row 10, we can see that the more topic information extra inputs contain, the better performance NASTS achieves. Especially, NASTS even can obtain 50.36 ROUGE-1 F1 score with Human-Ext-3. And it indicates that this is a pretty promising direction to use topic information guide the generation of the summary.

Table 1. Performance comparison on CNN/Daily Mail test set. All our ROUGE scores have a 95% confidence interval with respect to previous best models.

Models	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Pointer-network+cov [10]	39.53	17.28	36.38
Controlled [22]	39.75	17.29	36.54
GAN [23]	39.92	17.65	36.71
End2End + inconsistency [17]	40.68	**17.97**	37.13
RNN-ext+abs+RL+rerank [18]	40.88	17.80	**38.54**
NASTS + Coh-Ext	40.60	17.79	37.35
NASTS + SL-Ext	**40.94**	17.92	37.62
NASTS + Human-Ext-2	48.16	27.30	45.21
NASTS + Human-Ext-3	50.36	29.03	47.40

To further quantify the effectiveness of proposed policies in this paper, we make a step-by-step comparison (details in Table 2). First, the baseline with the topic encoder is tested and achieves about 2.29, 0.80 and 1.82 percent improvement. The reason is that the topic encoder can extract crucial elements to enhance the decoder. Then the weighted mechanism is added into the model and exceeds the previous step about 1.0, 0.7 and 1.17 percent. We may consider that the weighted mechanism is capable of integrating both topic information and document representation. Finally, we apply coverage mechanism to alleviate the repetition problem further. The improved ROUGE scores indicate that topic encoder and weighted mechanism are consistently producing better summaries.

Moreover, to demonstrate the improvement of topicality and readability by NASTS, we present some qualitative outputs by randomly extracting several summaries from the test set. We compare the reference summaries to the

Table 2. Results of NASTS+SL-Ext in different stages.

Models	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	36.44	15.66	33.42
+topic encoder	38.73	16.46	35.24
+ weighted mechanism	39.73	17.16	36.41
+article coverage	40.94	17.92	37.62

Table 3. Examples of generated summaries by the baseline and NASTS.

Reference: preacher trevor brooks has been refused a passport by the home office. brooks, 39, is an associate of islamist firebrand preacher anjem choudary. a home office letter said he is linked to banned group al muhajiroun. it added he was assessed as being likely to travel to syria to fight for isis.
Baseline: trevor brooks is a known associate of islamist firebrand preacher anjem choudary. the letter adds:' it is also assessed that you intend to travel overseas to syria where you would use this opportunity to engage in terrorism-related activity'. brooks, 39, was released from prison last month after serving part of his 11-month term for breaching bail. **(ROUGE-1: 36.96, ROUGE-2: 15.09, ROUGE-L: 36.65)**
NASTS + SL-Ext: trevor brooks, an associate of islamist firebrand anjem choudary, was told it was not in the public interest for him to have a passport because he is seen as a terror risk. a home office letter says that brooks, also known as abu izzadeen, is an extremist linked to members of the banned group al muhajiroun.' brooks is a known associate of islamist firebrand anjem choudary. choudary has been refused a passport after the home office accused him of planning to join islamic state. **(ROUGE-1: 62.00, ROUGE-2: 33.33, ROUGE-L: 59.70)**
NASTS + Human-Ext-3: trevor brooks is a known associate of islamist firebrand preacher anjem choudary. a home office letter says that brooks, also known as abu izzadeen, is an extremist linked to members of the banned group al muhajiroun.' a radical muslim preacher has been refused a passport after the home office accused him of planning to join islamic state. **(ROUGE-1: 62.92, ROUGE-2: 41.82, ROUGE-L: 62.78)**

summaries produced by our proposed model with different versions of topic sentences. As shown in Table 3, we can observe that the results produced by the baseline is readable but missed salient information like *"been refused a passport by the home office"*. However, summaries generated by NASTS (with SL-Ext or Human-Ext) are more readable and cover more topic information. Furthermore, NASTS with Human-Ext can even generate complex and logical sentences. It is considered that NASTS can grasp more insightful information and improve the topicality and readability for the summarization.

5 Conclusion

We propose a neural abstractive summarization model guided with topic sentences. Most importantly, a topic encoder is introduced to draw topic information and weighted Attention-over-Attention mechanism enables decoder focus on the prominent part of the article. By end-to-end training, our model significantly outperforms conventional Seq2Seq model on the CNN/Daily Mail dataset. In future work, we will focus on studying the topic-driven summarization model and integrate common sense into NASTS.

6 Acknowledgment

This work is supported by Natural Science Foundation of China (Grant No. 61872113, 61573118, U1813215, 61876052), Special Foundation for Technology Research Program of Guangdong Province (Grant No. 2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (Grant No. JCYJ20170307150528934, JCYJ2017 0811153836555, JCYJ20180306172232154), Innovation Fund of Harbin Institute of Technology (Grant No. HIT. NSRIF.2017052)

References

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

2. Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.
3. Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073*.
4. Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. *arXiv preprint arXiv:1708.00625*.
5. Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.
6. Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *Computer Science*, pages 2667–2671.
7. Preksha Nema, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. *arXiv preprint arXiv:1704.08300*.
8. Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. 2015. Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1954.
9. Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. *arXiv preprint arXiv:1804.07036*.
10. Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
11. Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. 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.
12. Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. *arXiv preprint arXiv:1805.03989*.
13. Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2017. Improving multi-document summarization via text classification. In *Thirty-First AAAI Conference on Artificial Intelligence*.
14. Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2101–2110.
15. Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *arXiv preprint arXiv:1805.03616*.
16. Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*.
17. Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.
18. Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.
19. Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.
20. Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081.
21. John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.
22. Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*.
23. Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative adversarial network for abstractive text summarization. *arXiv preprint arXiv:1711.09357*.