Topic Attentional Neural Network for Abstractive Document Summarization

Hao Liu, Hai-Tao Zheng(✉), and Wei Wang

Tsinghua-Southampton Web Science Laboratory, Graduate School at Shenzhen,
Tsinghua University, Shenzhen, China
{liuhao17,w-w16}@mails.tsinghua.edu.cn, zheng.haitao@sz.tsinghua.edu.cn

Abstract. Abstractive summarization is a renewed and challenging task of document summarization. Recently, neural networks, especially attentional encoder-docoder architecture, have achieved impressive progress in abstractive document summarization. However, the saliency of summary, which is one of the key factors for document summarization, still needs improvement. In this paper, we propose Topic Attentional Neural Network (TANN) which incorporates topic information into neural networks to tackle this issue. Our model is based on attentional sequence-to-sequence structure but has paired encoders and paired attention mechanisms to deal with original document and topic information in parallel. Moreover, we propose a novel selection method called *topic selection*. This method uses topic information to improve the standard selection method of beam search and chooses a better candidate as the final summary. We conduct experiments on the CNN/Daily Mail dataset. The results show our model obtains higher ROUGE scores and achieves a competitive performance compared with the state-of-the-art abstractive and extractive models. Human evaluation also demonstrates our model is capable of generating summaries with more informativeness and readability.

Keywords: Abstractive summarization · Neural network · Topic information · Attention mechanism

1 Introduction

Document summarization is a task to produce a concise and condensed summary which covers the core information of the original document. Automatic summarization models can be divided into two categories: extractive and abstractive. Extractive approaches select important segments from the original document and rearrange them to construct a summary. Totally different from extractive, abstractive approaches potentially generate new phrases or sentences. Requiring deeper understanding of natural language, abstractive approaches are more difficult and face great challenges, such as saliency, coherence and readability.

Recently, the models based on attentional sequence-to-sequence (seq2seq) framework have demonstrated great advantages for abstractive document summarization [12,15,17]. However, the saliency of summary, which plays a vital role

Q. Yang et al. (Eds.): PAKDD 2019, LNAI 11440, pp. 70–81, 2019.
https://doi.org/10.1007/978-3-030-16145-3_6

in document summarization, is still not satisfactory and needs improvement. Meanwhile, topic information, which is one of the most important features of original document, can also help to identify the key information but has not been paid enough attention yet.

In this paper, to increase the saliency of summaries and make generated summaries cover more core information, we propose Topic Attentional Neural Network (TANN) for abstractive document summarization. Our model expands attentional seq2seq structure by using paired encoders and paired attention mechanisms to deal with original document and topic information in parallel. To obtain topic information, we employ two classic topic-extracted methods: Latent Dirichlet Allocation (LDA) [1] and TF-IDF. Then, we use these two types of topic information to train our model respectively.

Moreover, we also utilize topic information to improve the selection method of beam search. Beam search algorithm is widely used for generating outputs in neural networks [3,12,17]. It can generate several sequences as candidates and chooses one of them as the final output. In summarization task, the standard selection method chooses the candidate with the highest conditional probability as final summary. However, the conditional probability of language model is based on the whole training dataset, which take no specific features of source document into account. To solve the problem, we propose a selection method called *topic selection*. Our method utilizes topic information to calculate a score for each candidate and choose the one with the highest score. The experimental results show that *topic selection* help to produce more informative summaries. Our main contributions can be listed as follows:

- TANN introduces topic information into neural networks to produce summaries with more salient information. As far as we know, it is the first time that topic information is used to improve document-level abstractive summarization.
- We propose a novel selection method *topic selection* which further utilizes topic information to improve the standard selection method of beam search. With the help of topic information, our selection method helps to improve the informativeness of summary.
- Experiment on the CNN/Daily Mail dataset demonstrate that our model achieves competitive results compared with state-of-the-art abstractive and extractive models. Human evaluation also demonstrates our model produces informative summaries with high readability.

This paper is organized as follows. Section 2 introduces related work about abstractive summarization. Section 3 describes our model. Section 4 describes the experiment and gives discussion. In Sect. 5, we conclude this paper.

2 Related Work

While a large number of past works for document summarization are extractive approaches [2,10,13], abstractive approaches generate summaries by understanding the source document, which are closer to the way human writes summaries.

Recently, neural networks applied in abstractive approaches have been intensively studied. Rush et al. are the first to introduce neural network for abstractive text summarization [15]. Their model is based on convolutional encoder-decoder architecture and shows a promising path of applying seq2seq in abstractive summarization. Chopra et al. [3] and Nallapati et al. [12] extend this work by using RNN in place of CNN. However, because of the fixed vocabulary of these models, the generated summaries tend to cause the out-of-vocabulary (OOV) problem. To overcome this issue, Gu et al. propose CopyNet [6] and Gulcehre et al. propose pointer network [7], which both extend seq2seq structure by copying OOV words directly from the source text.

Due to the lack of large document-level dataset, most past works are sentence-level summarization models [3,15], which summarize a document to one sentence. Nallapati et al. address this issue by introducing the CNN/Daily Mail dataset which consists of news from CNN and Daily Mail website [12]. Then, Paulus et al. propose the intra-attention networks with reinforcement learning [14]. See et al. introduce the coverage mechanism into summarization system to address the repetition on the CNN/Daily Mail dataset [17].

So far, few works have considered about using external information to improve abstractive summarization. Nallapati et al. use feature-rich word embedding as the input of model [12]. This embedding expands original word embedding with some linguistic information such as POS tags and named-entities. Li et al. use keyword representation as the extra input for attention to guide the summarization [9]. In this paper, we use topic information as one of the external information and utilizes it to produce more salient summaries.

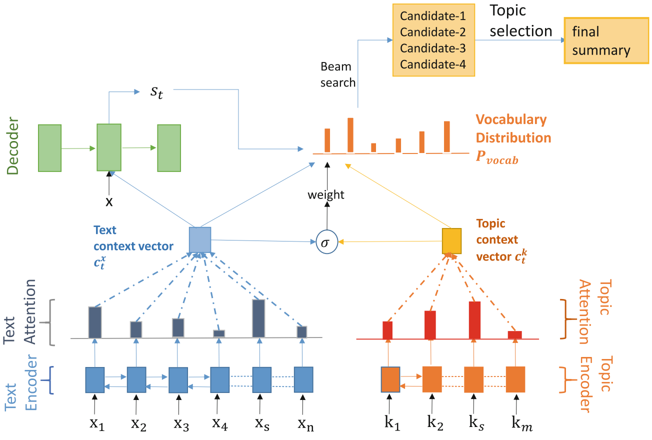


Fig. 1. TANN model with paired encoders and paired attention mechanisms.

3 Our Model

3.1 Overview

We build our model based on attentional encoder-decoder framework. In particular, our model has paired encoder and paired attention mechanisms. We use copying mechanism and coverage mechanism to avoid the OOV and repetition. We show the structure of our model in Fig. 1. In encoding step, the inputs are original document $x = \{x_1, x_2, ..., x_n\}$ and pre-extracted topic information $k = \{k_1, k_2, ..., k_m\}$. Then, the text context vector c_t^x and topic context vector c_t^k are calculated by the respective attention mechanisms. Next, these two context vectors are both used to compute vocabulary distribution in the decoder. Specially, we use a learnable parameter *weight* to make model learn the importance of topic information. Finally, we use beam search algorithm to generate candidate sequences and apply *topic selection* method to choose the final summary.

3.2 Paired Encoder

The input of our model consists of original document text x and topic words k. The text encoder and topic encoder both use a bidirectional LSTM [16]. For the text encoder, it uses the source word x as input and produces two LSTM states as forword state and backward state:

$$\overrightarrow{h_i^x}, \overleftarrow{h_i^x} = biLSTM\left(x\right) \tag{1}$$

Then, we combine these two states into the final hidden state h_i^x as:

$$h_i^x = relu\left(W\left[\overrightarrow{h_i^x}, \overleftarrow{h_i^x}\right] + b\right) \tag{2}$$

where W and b are learnable. Using all the words $x = \{x_1, x_2, ..., x_n\}$ as input, text encoder produces a sequence of hidden states $\{h_1^x, h_2^x, ..., h_n^x\}$. Similarly, the topic encoder reads m topic words $k = \{k_1, k_2, ..., k_m\}$ and builds $\left\{h_1^k, h_2^k, ..., h_m^k\right\}$ as the representation of topic information.

3.3 Paired-Attentional Decoder

We adopt the decoder with paired attention mechanisms to deal with text information and topic information respectively. For text attention, the text context vector c_t^x at step t is calculated based on the encoder hidden state h_i^x and the decoder hidden state s_t:

$$e_{t,i}^x = v^T \tanh\left(linear\left(h_i^x, s_t\right)\right) \tag{3}$$

$$a_{t,i}^x = softmax\left(e_t^x\right) \tag{4}$$

$$c_t^x = \sum_i a_{t,i}^x h_i^x \tag{5}$$

where v is learned vector. Likewise, topic attention mechanism also computes the topic context vector c_t^k in the similar way.

Specially, a learnable parameter *weight* is computed as:

$$weight = \sigma\left(W_t c_t^x + W_k c_t^k + b\right) \tag{6}$$

where W_t, W_k, b are learnable. The *weight* is regarded as a parameter that indicates how much topic context vector is involved in the generation of words. Then, the vocabulary distribution is calculated as:

$$P_{vocab} = softmax\left(V'\left(V\left[s_t, c_t^x, weight \cdot c_t^k\right] + b\right) + b'\right) \tag{7}$$

where V', V, b, b' are learnable, s_t is decoder hidden state. Finally, the training loss is defined as:

$$loss_m = -\frac{1}{T}\sum_{t=0}^{T} logP_{vocab}\left(y_t\right) \tag{8}$$

3.4 Copying and Coverage

Copying Mechanism. A number of works for sequence generation tasks have solved the problem of OOV words by copying corresponding words directly from original document [6,17,18]. Following these works, we employ copying mechanism in our model. We use a variable *gate* as a switch to choose whether generating a word from the fixed vocabulary ($gate = 1$), or copying from the source ($gate = 0$). The *gate* at step t is computed as:

$$gate = \sigma\left(linear\left(c_t^x, s_t, x_t\right)\right) \tag{9}$$

Then, the next word y_t is predicted by:

$$P\left(y_t\right) = \begin{cases} P_{vocab}\left(y_t\right), & gate = 1 \\ \sum_{i:x_i=y_t} a_{i,t}^x, & gate = 0 \end{cases} \tag{10}$$

where $a_{i,t}^x$ is the text attention distribution of x_i.

Coverage Mechanism. As for the task of producing long summary, repetition is a common problem and impairs the quality of summaries. Following See et al. [17], we use the coverage mechanism to address this issue. Specially, at each step t, our model keep a coverage vector which sums the text attention distribution $a_{*,i}^x$ before t:

$$cov^t = \sum_{j=0}^{t-1} a_{j,i}^x \tag{11}$$

The vector cov^t indicates how much attention has the model paid for the input word x_i before step t. Then, We use coverage vector as an additional input for text attention mechanism and change the calculation of $e_{t,i}^x$ in Eq. (3) as:

$$e_{t,i}^x = v^T \tanh\left(linear\left(h_i^x, s_t, cov^t\right)\right) \tag{12}$$

We define the loss of coverage mechanism as:

$$loss_{cov} = \sum_{i=1}^{t} min\left(a_i^x, cov^t\right) \tag{13}$$

Then, the final loss of our model is defined as:

$$loss_{final} = loss_m + loss_{cov} \tag{14}$$

3.5 Topic Selection

Beam search algorithm, which is used for generating sequence, produces words step by step in decoder. For each decoding step, the beam search keeps top K sequences with high conditional probability, where K is a hyper-parameter called beam size. Therefore, at the end of beam search, the algorithm produces K sequences which are viewed as candidates for summary. The standard selection method selects the one with the highest conditional probability as the summary. Taking the concrete topic information into account, we propose a brand selection method *topic selection*. We firstly obtain the candidate sequences by using beam search algorithm and sort them by conditional probability. Then, for each of the sequence, we calculate a feature value for every word x_i as:

$$feature\left(x_i\right) = \begin{cases} \sum_t a_{t,i}^k, & x_i \in \; topic\ word \\ 0, & x_i \notin \; topic\ word \end{cases} \tag{15}$$

where $a_{t,i}^k$ is the topic attention distribution at timestep t. This distribution vector can stand for the degree of participation of each topic word, and it can be used for measuring the importance of the word. Next, we calculate the score of each candidate sequence s by suming the feature value of every word:

$$Score\left(s\right) = \sum_{x_i \in s} feature\left(x_i\right) \tag{16}$$

Finally, we resort the candidates and choose the one with the highest $Score$ as final summary. If the $Score$ of candidates are equal, we choose the one with higher conditional probability, which makes our method consider not only conditional probability but topic information.

4 Experiments

4.1 Dataset

We conducted experiments on the CNN/Daily Mail dataset[1] which consists of news stories in CNN and Daily Mail website [8,12]. The corpora has 312,085 articles paired with human-written multi-sentence summaries. This dataset has two version: non-anonymized and anonymized. Following See et al. [17], we obtain the non-anonymized version by same processing steps[2], which divide the dataset into 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs.

[1] https://cs.nyu.edu/~kcho/DMQA/.

[2] https://github.com/abisee/cnn-dailymail.

4.2 Topic Information Acquisition

We extract topic words of each document by using two classic topic models: TF-IDF and LDA. These two models are both trained on the full CNN/Daily Mail dataset. For TF-IDF, we firstly build individual vocabulary for each document and calculate TF and IDF for each word. Next, we select top 50 percent of words in the document vocabulary as the topic words. Finally, we rearrange topic words following the order of their appearance in the original document. As for LDA, we obtain topic information by GibbsLDA++[3] which using Gibbs sampling for parameter estimation and inference. We set the hyperparameters of LDA as $\alpha = 0.04$ and $\beta = 0.01$. For each document, we pick top 25 topics and each topic has 16 topic words as the final topic information.

4.3 Implementation

We train our model with 128-dimensional word embeddings. In particular, we train word embeddings directly from scratch. We use bi-LSTM for both text encoder and topic encoder. For each encoder, we use the hidden state dimension as 256. We select 50k most frequently used words from both source documents and human-written summaries, then put them together as the vocabulary. Our model is trained using Adagrad optimizer [4] with learning rate 0.15 and the initial accumulator value is set to 0.1. We use mini-batches of size 16 and the encoder size is set to 400. For decoding time, we set the decoder size as 100 for training and 120 for testing. Beam size is fixed as 4 during beam search.

4.4 Results and Discussion

In this section, we firstly report and analysis the results of ROUGE scores. Then, we give discussion on the performance of two types of topic information (TF-IDF and LDA) and *topic selection*. Finally, we report the result of human evaluation.

Quantitative Analysis. We evaluate our model with ROUGE [5] scores which are widely used in summarization task. We compare our models with some state-of-the-art extractive models (lead-3 [12] and SummaRuNNer [11]) and abstractive models. The overall performance evaluation is demonstrated in Table 1. Due to the different topic-words acquisition methods, we use several notations (**m1** to **m8**) to represent our models. From the Table 1, we can see that our basic model **m1** and **m5** both outperform baseline PG model (the sixth row in Table 1) on all ROUGE scores, which shows the effectiveness of topic attention. After using *topic selection* to choose the final summaries, **m2** and **m6** obtain ROUGE scores improvement, indicating *topic selection* method helps to differentiate the candidate sequences. With coverage mechanism employed, **m7** model has shown a competitive performance in abstractive models. Moreover, it is observed that **m8** achieves best performance on non-anonyized dataset and even exceeds strong

[3] http://gibbslda.sourceforge.net/.

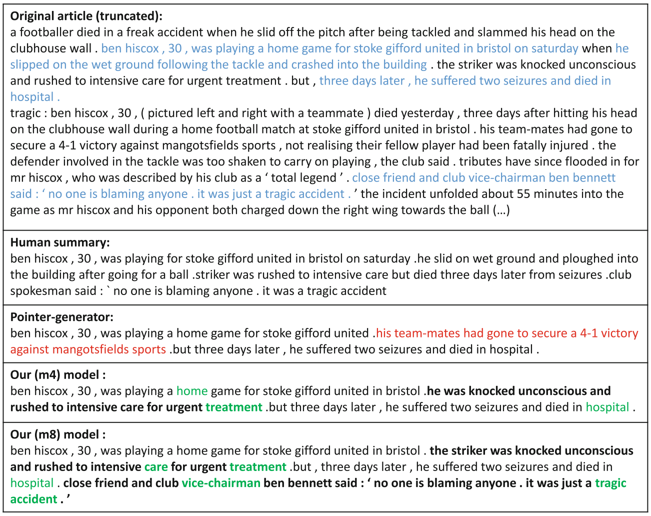
Table 1. The results is full-length F1 scores for ROUGE-1, ROUGE-2 and ROUGE-L on the CNN/daily mail test set. All ROUGE scores have a 95% confidence interval of at most ±0.25. Models with subscript * were trained and tested on the anonymized version dataset. Best results on non-anonymized dataset are bolded.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3 [12]	40.34	17.70	36.57
SummaRuNNer [11]*	39.60	16.20	35.30
Abstractive model [12]*	35.46	13.30	32.65
DeepRL, ML [14]*	39.87	15.82	36.90
DeepRL, Intra-attn [14]*	41.16	15.75	39.08
PG [17]	36.44	15.66	33.42
PG, coverage [17]	39.53	17.28	36.38
TANN, LDA (**m1**)	37.48	16.46	35.13
TANN, LDA, *topic selection* (**m2**)	38.32	16.96	35.45
TANN, LDA, coverage (**m3**)	39.75	17.45	36.62
TANN, LDA, coverage, *topic selection* (**m4**)	40.09	17.78	36.92
TANN, TF-IDF, (**m5**)	37.78	16.56	35.34
TANN, TF-IDF, *topic selection* (**m6**)	38.82	17.16	35.48
TANN, TF-IDF, coverage (**m7**)	40.29	17.89	36.92
TANN, TF-IDF, coverage, *topic selection* (**m8**)	**40.56**	**18.01**	**37.15**

extractive baseline lead-3. Besides, though different version of dataset may cause some deviations in the comparison of ROUGE, **m8** still achieves best ROUGE-2 score.

To illustrate the effectiveness of our models, we show an example in Fig. 2. Compared to human summary, it can be seen that with the help of topic information (green font), **m4** and **m8** both capture the important information (bold font) which Pointer-generator misses. Moreover, our model may identify the core information accurately and generate less unnecessary information (red font).

LDA or TF-IDF. It can be observed that our models with TF-IDF yield higher ROUGE scores than the models with LDA (such as the result of **m8** and **m4**). Meanwhile, the final value of parameter *weight* also supports this result. We show the learning curve of *weight* in Fig. 3. As we can see, the initial value of *weight* is about 0.5 for both **m4** and **m8**. After the training, the *weight* of **m8** stabilizes around to 0.65 and that of **m4** finally drops to about 0.45. It indicates that the topic information acquired by TF-IDF may be more involved in the generation process. This result in part because TF-IDF method extracts topic words directly from original document which may represent more important details. Moreover, we observe that these two learning curves both quickly increase



Original article (truncated):
a footballer died in a freak accident when he slid off the pitch after being tackled and slammed his head on the clubhouse wall . ben hiscox , 30 , was playing a home game for stoke gifford united in bristol on saturday when he slipped on the wet ground following the tackle and crashed into the building . the striker was knocked unconscious and rushed to intensive care for urgent treatment . but , three days later , he suffered two seizures and died in hospital .
tragic : ben hiscox , 30 , (pictured left and right with a teammate) died yesterday , three days after hitting his head on the clubhouse wall during a home football match at stoke gifford united in bristol . his team-mates had gone to secure a 4-1 victory against mangotsfields sports , not realising their fellow player had been fatally injured . the defender involved in the tackle was too shaken to carry on playing , the club said . tributes have since flooded in for mr hiscox , who was described by his club as a ' total legend ' . close friend and club vice-chairman ben bennett said : ' no one is blaming anyone . it was just a tragic accident . ' the incident unfolded about 55 minutes into the game as mr hiscox and his opponent both charged down the right wing towards the ball (...)

Human summary:
ben hiscox , 30 , was playing for stoke gifford united in bristol on saturday .he slid on wet ground and ploughed into the building after going for a ball .striker was rushed to intensive care but died three days later from seizures .club spokesman said : ` no one is blaming anyone . it was a tragic accident

Pointer-generator:
ben hiscox , 30 , was playing a home game for stoke gifford united .his team-mates had gone to secure a 4-1 victory against mangotsfields sports .but three days later , he suffered two seizures and died in hospital .

Our (m4) model :
ben hiscox , 30 , was playing a home game for stoke gifford united in bristol **.he was knocked unconscious and rushed to intensive care for urgent treatment** .but three days later , he suffered two seizures and died in hospital .

Our (m8) model :
ben hiscox , 30 , was playing a home game for stoke gifford united in bristol . **the striker was knocked unconscious and rushed to intensive care for urgent treatment** .but , three days later , he suffered two seizures and died in hospital . **close friend and club vice-chairman ben bennett said : ' no one is blaming anyone . it was just a tragic accident . '**

Fig. 2. Typical comparison. **m4** and **m8** both capture the most important information (bold font) which Pointer-generator misses. With the guidance of topics (green font), our models tend to generate less unnecessary information (red font). (Color figure online)

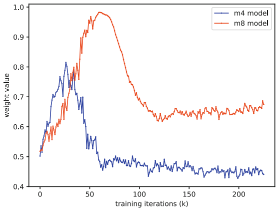


Fig. 3. Learning curve of *weight*.

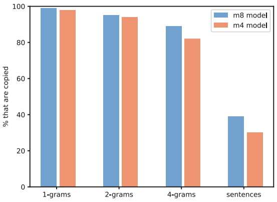
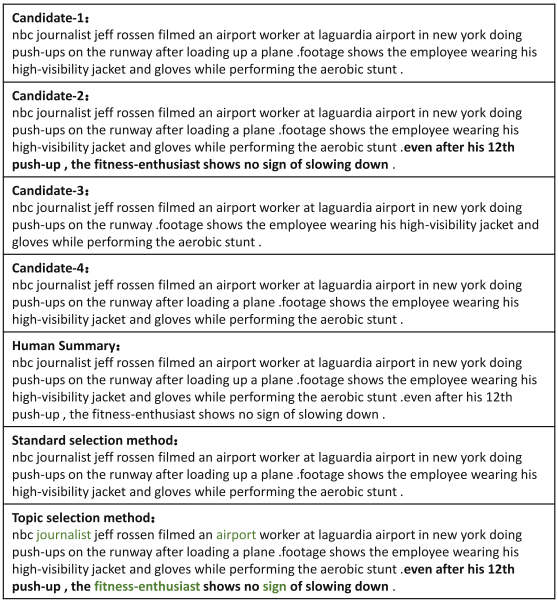


Fig. 4. Copying rate of n-grams.

and peak at early iterations of training. It may because without pre-trained word-embedding, topic context vector is more informative than text context vector at the beginning of training.

We also study whether it has a link between topic-extracted methods and how abstractive the models are. For our two final models **m4** and **m8**, we compare the copying rate of n-grams which computes n-grams that both appear in the source document and generated summary. The result is shown in Fig. 4. For 1-gram and 2-grams, our two models have similar performance. However, for sentence-level output, **m4** has lower copying rate (about 33%) than that of **m8** (about 41%), showing **m4** is more abstractive and tends to generate diverse summaries. The reason for this result may be that LDA method can generate some novel words, which enrich the topic information.

Topic Selection. The results show that *topic selection* method helps the model to achieve better performance on all ROUGE scores, indicating that topic words directly help to identify the key information. As an example shown in Fig. 5, the candidate sequences are similar but still contain different information (bold font). With the guidance of topic words *"fitness-enthusiast"* and *"sign"*, our *topic selection* select the best candidate sequence Candidate-2 as the final summary, even though Candidate-1 has the highest conditional probability and is selected by standard selection method.



Candidate-1:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading up a plane .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .

Candidate-2:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading a plane .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .**even after his 12th push-up , the fitness-enthusiast shows no sign of slowing down** .

Candidate-3:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .

Candidate-4:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading a plane .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .

Human Summary:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading up a plane .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .even after his 12th push-up , the fitness-enthusiast shows no sign of slowing down .

Standard selection method:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading up a plane .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .

Topic selection method:
nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading a plane .footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .**even after his 12th push-up , the fitness-enthusiast shows no sign of slowing down** .

Fig. 5. An example to show effectiveness of *topic selection*. All generated summaries is produced by **m8**. The topic words is green font. The important information missed by standard selection method is bold font. (Color figure online)

Human Evaluation. We also perform human evaluation to further evaluate our model. We use the following as evaluation criteria: (1) *Informativeness*, the main ideas and important details of article are shared; (2) *Coherence*, ideas are expressed clearly without repetition; (3) *Readability*, the generated summaries are fluent and grammatical.

We compare two final models **m4** and **m8** with lead-3 baseline [17] and PG with coverage [17]. For the process of human evaluation, we randomly pick 100 different samples from the test set. We show the original articles and four generated summaries to the human judges. The judges evaluate each summary by scoring 1–5 point according to each criterion described above. The 5-point means "best", while 1-point means"worst". Each sample is evaluated by 3 judges. The score of each criterion is averaged across all human judges.

Table 2. Human evaluation result. Best results are bolded

Model	Informativeness	Coherence	Readability
Lead-3	3.45	3.40	3.46
PGC	3.35	3.48	3.38
m4	3.52	**3.56**	3.52
m8	**3.56**	3.50	**3.60**

We invite 10 graduate students as our judges. The results are shown in Table 2. Both **m4** and **m8** outperform state-of-the-art abstractive pointer-generator and lead-3 extractive baseline. Compared to pointer-generator, it is observed that our models show competitive performance on informativeness and readability, indicating our models may provide more key information. Comparing the results of **m4** and **m8**, while being inferior to the other two criteria, **m4** shows advantage on Coherence. It indicates the richer topic information produced by LDA can help to improve the diversity of summaries in some extent.

5 Conclusion

In this paper, we propose topic attentional neural network (TANN) to utilize topic information for abstractive document summarization. We also propose a novel selection method named *topic selection* to improve the selection method of beam search. Experiments on the CNN/Daily Mail dataset demonstrate that, with the help of topic information, our model achieves a competitive performance with state-of-the-art abstractive and extractive methods and is able to produce summaries with more salient information. Human evaluation also demonstrates our model generates summaries with high informativeness and readability. In the future, we plan to extend our model with Generative Adversarial Network to generate more diverse summaries.

Acknowledgements. This research is supported by National Natural Science Foundation of China (Grant No. 61773229), Basic Scientific Research Program of Shenzhen City (Grant No. JCYJ20160331184440545), and Overseas Cooperation Research Fund of Graduate School at Shenzhen, Tsinghua University (Grant No. HW2018002).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. Arch. **3**, 993–1022 (2003)
2. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. In: Meeting of the Association for Computational Linguistics, pp. 484–494 (2016)
3. Chopra, S., Auli, M., Rush, A.M.: Abstractive sentence summarization with attentive recurrent neural networks. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–98 (2016)

4. Duchi, J.C., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
5. Flick, C.: Rouge: a package for automatic evaluation of summaries. In: The Workshop on Text Summarization Branches Out, p. 10 (2004)
6. Gu, J., Lu, Z., Li, H., Li, V.O.K.: Incorporating copying mechanism in sequence-to-sequence learning. Meeting of the Association for Computational Linguistics, pp. 1631–1640 (2016)
7. Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., Bengio, Y.: Pointing the unknown words. Meeting of the Association for Computational Linguistics, pp. 140–149 (2016)
8. Hermann, K.M., et al.: Teaching machines to read and comprehend. In: Neural Information Processing Systems, pp. 1693–1701 (2015)
9. Li, C., Xu, W., Li, S., Gao, S.: Guiding generation for abstractive text summarization based on key information guide network. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 55–60. Association for Computational Linguistics (2018). http://aclweb.org/anthology/N18-2009
10. McDonald, R.: A study of global inference algorithms in multi-document summarization. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 557–564. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71496-5_51
11. Nallapati, R., Zhai, F., Zhou, B.: SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In: National Conference on Artificial Intelligence, pp. 3075–3081 (2017)
12. Nallapati, R., Zhou, B., Santos, C.N.D., Gulcehre, C., Xiang, B.: Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Conference on Computational Natural Language Learning, pp. 280–290 (2016)
13. Nishikawa, H., Arita, K., Tanaka, K., Hirao, T., Makino, T., Matsuo, Y.: Learning to generate coherent summary with discriminative hidden semi-Markov model. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1648–1659. Dublin City University and Association for Computational Linguistics (2014). http://www.aclweb.org/anthology/C14-1156
14. Romain Paulus, C.X., Socher, R.: A deep reinforced model for abstractive summarization. In: The 2018 International Conference on Learning Representations (Submitted for Publication)
15. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. Empirical Methods in Natural Language Processing, pp. 379–389 (2015)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
17. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1073–1083. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1099, http://www.aclweb.org/anthology/P17-1099
18. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. Neural Information Processing Systems, pp. 2692–2700 (2015)