A Gist Information Guided Neural Network for Abstractive Summarization

Yawei Kong[1,2⋆][0000−0001−8823−866X], Lu Zhang[1,2⋆][0000−0001−9693−1122], and Can Ma[1⋆⋆]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing
{kongyawei,zhanglu0101,macan}@iie.ac.cn

Abstract. Abstractive summarization aims to condense the given documents and generate fluent summaries with important information. It is challenging for selecting the salient information and maintaining the semantic consistency between documents and summaries. To tackle these problems, we propose a novel framework - Gist Information Guided Neural Network (GIGN), which is inspired by the process that people usually summarize a document around the gist information. First, we incorporate multi-head attention mechanism with the self-adjust query to extract the global gist of the input document, which is equivalent to a question vector questions the model "What is the document gist?". Through the interaction of the query and the input representations, the gist contains all salient semantics. Second, we propose the remaining gist guided module to dynamically guide the generation process, which can effectively reduce the redundancy by attending to different contents of gist. Finally, we introduce the gist consistency loss to improve the consistency between inputs and outputs. We conduct experiments on the benchmark dataset - CNN/Daily Mail to validate the effectiveness of our methods. The results indicate that our GIGN significantly outperforms all baseline models and achieves the state-of-the-art.

Keywords: Abstractive Summarization · Multi-Head Attention · Global Gist · Consistency Loss

1 Introduction

Recently, document summarization has attracted growing research interest for its promising commercial values. It aims to produce fluent and coherent summaries with the original documents. Existing approaches for building a document summarization system can be categorized into two groups: extractive and abstractive methods. The extractive methods focus on extracting sentences from the original document, which can produce more fluent sentences and preserve the meaning of the original documents but tend to information redundancy and

⋆ Equal contribution.
⋆⋆ Can Ma is the corresponding author.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

incoherence between sentences. In contrast, the abstractive methods effectively avoid these problems by utilizing arbitrary words and expressions that are more consistent with the way of humans. However, the abstractive methods are much more challenging due to the sophisticated semantic organization.

Encouraged by the success of recurrent neural network (RNN) in NLP, most typical approaches [4,14,18,20] employ the sequence-to-sequence (seq2seq) framework to model the document summarization system that consists of an encoder and a decoder. Given the input documents, the encoder first encodes them into semantic vectors and then the decoder utilizes these vectors to generate summaries in the decoding step. Although the popular methods are improved from various perspectives, such as introducing reinforcement learning [3, 16] or incorporating topic information [18, 23], they still fail to achieve convincing performance for ignoring the global gist information.

Intuitively, humans tend to generate a document summary around the gist information. Different from the topic that just focuses on the talking point of the original documents, the gist information represents the essence of text that contains more wealth of information. Thus, the gist is more suitable for the document summarization task and can guide the model how to generate relevant, diverse, and fluent summaries.

Towards filling this gap, we propose an effective Gist Information Guided Neural Network (GIGN) for abstractive summarization. To distill the gist information from the semantic representation of the original documents, we first introduce a Gist Summary Module (GSM) that consists of the multi-head attention mechanism with self-adjust query. The query effectively questions the model "What is the document gist?". Through the interaction between this query vector and the hidden state of each token in the document, we obtain a global representation of the gist, which contains several pieces of salient information. Obviously, the gist not only plays a global guidance role but also is required to have the capability of attending to different contents of gist during the process of decoding. Thus, we propose a Remaining Gist Information Guided Module (RGIGM). The remaining gist information is calculated by the global gist and the salient information of the generated summary, which dynamically guides the decoder to generate tokens at each step. And this mechanism can effectively reduce redundant information and makes the gist contents express completely. Furthermore, we also propose a gist consistency loss to guarantee that the main information of the generated summary is consistent with the input document. Finally, we introduce policy gradient reinforcement learning [16] to reduce the exposure bias problem.

We conduct experience on the benchmark CNN/Daily Mail dataset to validate the effectiveness of GIGN. The experimental results indicate that our GIGN significantly outperforms all baselines. In summary, this paper makes the following contributions:

– To the best of our knowledge, this is the first work to introduce the gist information for abstractive summarization.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

- We introduce a gist summary module, which contains several pieces of salient information. And then, a remaining gist information guided module is employed to attending to different contents of gist. They dynamically incorporate the gist information into the generation process, which effectively improves the model performance.
- We further propose a novel gist consistency loss to ensure the generated summary is coherent with the inputs. And the reinforced learning can further improve performance by reducing the exposure bias problem.
- The empirical results demonstrate that our approach outperforms all baselines in both automatic metrics and human judgments. Further analysis shows that our method can generate more salient and relevant summaries.

2 Related Work

In recent years, abstractive summarization [4,14,19,20,25] has received increasing attention for its promising commercial values. Different from extractive methods that directly selects salient sentences from the original documents, abstractive summarization aims to generate summaries word-by-word from the final vocabulary distribution, which is more consistent with the way of human beings.

The abstractive methods are more challenging for the following conspicuous problems: Out-of-vocabulary (OOV), repetition, and saliency. Therefore, some previous works [14, 16, 20] pay attention to tackle the OOV problem by introducing the pointer network. To eliminate repetitions, See et al. [20] propose a coverage mechanism that is a variant of the coverage vector from Machine Translation. However, the most difficult and concerning problem is how to improve the saliency.

To tackle this problem, some studies attempt to introduce the template discovered from the training dataset to guide the summary generation.For example, Cao et al. [2] employ the IR platform to retrieve proper summaries as candidate templates and then jointly conduct template reranking as well as template-aware summary generation. Wang et al. [22] propose a novel bi-directional selection mechanism with two gates to extract salient information from the source document and executes a multi-stage process to extract the high-quality template from the training corpus. Moreover, You et al. [24] extend the basic encoder-decoder framework with an information selection layer, which can explicitly model and optimize the information selection process. However, they are difficult to optimize for retrieving template first. And the wrong template will introduce noise to the model, which significantly hurt the generation performance.

Different from the works that incorporate templates, many researchers improve saliency by introducing topics. For example, Krishna et al. [11] take an article along with a topic of interest as input and generates a summary tuned to the target topic of interest. Li et al. [23] improve the coherence, diversity, and informativeness of generated summaries through jointly attending to topics and word-level alignment. To incorporate the important information, Li et al. [12] utilize keywords extracted by the extractive model to guide the generation process. Furthermore, Perez-Beltrachini et al. [18] train a Latent Dirichlet Allocation

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

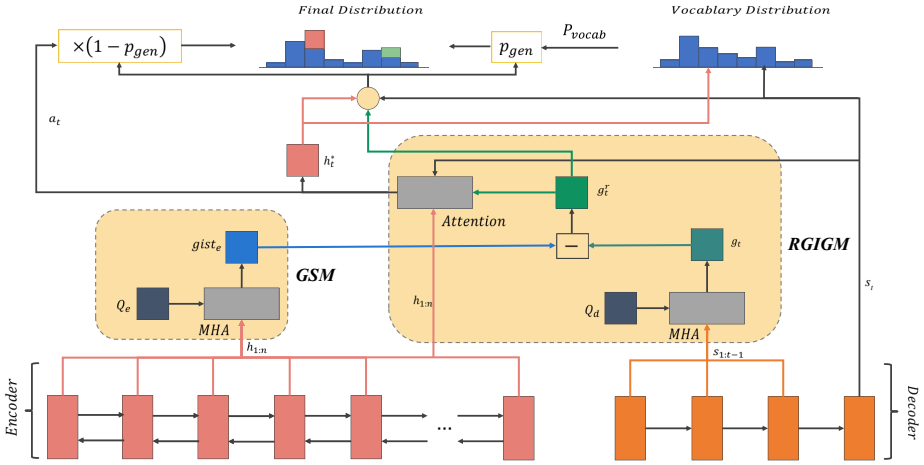

Fig. 1. The architecture of GIGN. It extends the pointer generator network with a gist summary module (GSM) and a remaining gist information guided Module (RGIGM).

model [1] to obtain sentence-level topic distributions. Although these methods have made some progress, they still fail to ignore the fact that the decoder should focus on different contents at different time steps and the information contains in topics is not enough to support the entire abstract.

In this paper, we propose a new framework, namely GIGN, to tackle these problems. We first introduce a gist summary module to obtain a global representation of the gist that contains several pieces of salient information. Then, the remaining gist information guided module is employed to attend to different contents during the decoding process.

3 Model

In this section, we introduce our Gist Information Guided Neural Network (GIGN) in detail. Given the source document $X = (x_1, x_2, \cdots, x_n)$, our model aims to generate the corresponding summary $Y = (y_1, y_2, \cdots, y_m)$. As shown in Figure 1, our GIGN mainly includes the Gist Summary Module (GSM) and the Remaining Gist Information Guided Module (RGIGM). We briefly describe the pointer generator network in Section 3.1 firstly. Then, Section 3.2 introduces our gist summary module. And the remaining gist information guided module is described in Section 3.3. Finally, we introduce our training objective in the Section 3.4, which includes the gist consistency loss and the reinforcement learning objective. Notably, all W and b are learnable parameters.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

3.1 Pointer Generator Network

The Pointer-Generator Network (PGN) aims to solve the out-of-vocabulary (OOV) problem, which extends ses2seq networks by adding a copy mechanism that allows tokens to be copied directly from the source. First, the encoder is a single layer BiLSTM that produces a sequence of hidden states $h_i, i \in [1, n]$ by feeding in the input tokens x_i. Then, the final hidden state h_n is served as the initial hidden state of the decoder, which is an un-directional LSTM. Finally, at each decoding time step t, the calculation of hidden state s_t is formulated as:

$$s_t = BiLSTM(s_{t-1}, y_{t-1}) \tag{1}$$

where y_{t-1} and s_{t-1} are the word embedding of the output and hidden state at the previous step, respectively.

To solve the repetition problem, we employ the coverage mechanism, which ensures the attention mechanism's current decision (choosing where to attend next) only informed by a reminder of its previous decisions. At the time step t, we maintain a coverage vector $c_t = \sum_{j=0}^{t-1} a^j$, which is the sum of attention distributions over all previous decoder time steps. And the attention distribution a_t are calculated as follows:

$$a_i^t = softmax(V^T tanh(W_h h_i + W_s s_t + W_c c_i^t + b_{attn})) \tag{2}$$

Based on the coverage vector and the previous hidden state, the vocabulary distribution of the next token is computed as follows:

$$P_{vocab} = softmax(W_{v'}(W_v[s_t; c_t] + b_v) + b_{v'}) \tag{3}$$

To solve the OOV problem, the pointer mechanism aims to copy rare or unknown words from the original document via pointing. Thus, the model first computes a generation probability:

$$p_{gen} = \sigma(W_{h^*}^T h_t^* + W_s^T s_t + W_x^T x_t + b_{ptr}) \tag{4}$$

where σ is the sigmoid function and $h_t^* = \sum_{j=1}^{n} a_j^t h_j$ is the context vector. Moreover, p_{gen} acts as a soft switch to decide whether to generate a word from the vocabulary by sampling from P_{vocab} or copy a word from the input sequence by sampling from the attention distribution a^t. Therefore, the final probability distribution is computed as follows:

$$P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t \tag{5}$$

3.2 Gist Summary Module

Inspired by the interaction between Question and Passage in Machine Reading Comprehension task [9], we introduce a Gist Summary Module (GSM) to obtain the gist of input document. As illustrated in Figure 1, the GSM consists of a

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

self-adjust query Q^e and a Multi-Head Attention (MHA) mechanism [21]. In our model, the self -adjust query Q^e is fed as the question, while the representation of source document $H = [h_1, \cdots, h_n]$ serves as the passage. The query vector seems to question the model "What is the document gist?". From the interaction between Q^e and H, we can get a question-aware global representation $gist_e$ that attends to several pieces of salient information for the document. The gist vector can be formulated as follows:

$$\begin{aligned} gist_e &= MHA(Q^e, H) \\ MHA &= [head_1; head_2 \ldots; head_k]W^O \\ head_j &= softmax(\frac{Q^e W_j^Q (HW_j^K)^T}{\sqrt{d_k}})HW_j^V \end{aligned} \tag{6}$$

where T represents transpose function and k is the number of heads. Notably, the Q^e is a learnable vector.

3.3 Remaining Gist Information Guided Module

The gist of a document contains several pieces of salient information, which needs to be distributed in different parts of the generated abstract. Therefore, we introduce a Remaining Gist Information Guide Module(RGIGM), which obtains the remaining gist information by calculating the difference between the generated semantics and the global gist. In this way, the decoder can attend to the most needed information of the gist for the current timestamp. As the continuation of the generation process, the information contained in the gist is constantly expressed, which can prevent semantic duplication and improve the conciseness of the summary. At time step t, we utilize the Gist Summary Module to obtain the generated information based on the previous hidden states $s_{1:t-1}$:

$$gist_t = MHA(Q^d, s_{1:t-1}) \tag{7}$$

where Q^d is the self-adjust question vector. Intuitively, $gist_t$ represents the expressed gist information contained in the generated sequence $y_{1:t-1}$. Then, we obtain the current remaining gist information $gist_t^r$ by subtracting $gist_t$ from the global $gist_e$:

$$gist_t^r = gist_e - gist_t \tag{8}$$

Furthermore, we utilize the remaining gist information $gist_t^r$ to guide the whole process of sequence generation. First, we propose the gist-aware attention distribution, which incorporates the context vector into the generation process. Then, Equation 2 is modified as follows:

$$a_i^t = softmax(V^T tanh(W_h h_i + W_{sg}[s_t; gist_t^r] + W_c c_i^t + b_{attn})) \tag{9}$$

In this way, this distribution is not only related to the current hidden state s_t but also affected by the remaining gist information. Then, we also apply the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

remaining gist to the calculation of vocabulary distribution. And Equation 3 can be modified as follows:

$$P_{vocab} = softmax(W_{v'}(W_v[s_t; gist_t^r; c_t] + b_v) + b_{v'}) \tag{10}$$

We further introduce the $gist_t^r$ into the pointer mechanism, which enables the pointer to identify the words that are relevant to the remaining salient information. And Equation 4 is modified as follows:

$$p_{gen} = \sigma(W_{h^*}^T h_t^* + W_{sg}^T[s_t; gist_t^r] + W_x^T x_t + b_{ptr}) \tag{11}$$

Finally, the whole RGIGM mechanism allows the decoder can generate unexpressed gist semantic information, which not only effectively prevents semantic repetition but also significantly enhances the saliency of the generated abstracts.

3.4 Model Training Objective

To train the model, we use a mixed training objective that jointly optimizes four loss functions, including the negative log-likelihood loss, the coverage loss, the gist consistency loss and the reinforcement learning loss.

Negative Log-Likelihood (NLL) Loss The model is first pre-trained to optimize NLL loss, which is widely used in sequence generation tasks. We define (X, Y) is a document-summary pair in training set. The function is formulated as follows:

$$\mathcal{L}_{NLL} = -\sum_{t=1}^{m} \log p(y_t|y_1 ... y_{t-1}, X; \theta) \tag{12}$$

Coverage Loss We utilize the coverage loss to alleviate the repetition problem, which aims to penalize the attention mechanism to focus on the same locations frequently. The formula can be described as follows:

$$\mathcal{L}_{Coverage} = \sum_i \min(a_i^t, c_i^t) \tag{13}$$

Gist Consistency Loss To further ensure the consistency of the original document and the generated abstracts, we propose a gist consistency loss, which maximizes the similarity between the gist of source document and the salient information of generated results. At the time step t, we get the current salient semantics $gist_t$ for the generated tokens $y_1, \cdots, y_{t-1}$. Thus, we obtain the salient information of the entire generated summary $gist_m$ when decoding to the last word:

$$\mathcal{L}_{GCL} = cos(gist_m, gist_e) \tag{14}$$

where $cos(\cdot)$ represents the cosine similarity. By maximizing the similarity between the global gist and the generated gist, the salient information of documents can be expressed completely.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

Reinforcement Learning (RL) Loss In order to improve the naturalness of the generated sequence and alleviate the exposure bias problem, we utilize reinforcement learning [16] to directly optimize the ROUGE evaluation metric [13] of the discrete target, which is non-differentiable. For each training example X, two output sequences are generated: $\hat{y}_t$ is sampled from the probability distribution $p(\hat{y}_t|\hat{y}_1 \cdots \hat{y}_{t-1}, X; \theta)$ at each time step and $\tilde{y}_t$ is the baseline output that is greedily generated by decoding from $p(\tilde{y}_t|\tilde{y}_1 \cdots \tilde{y}_{t-1}, X; \theta)$. The training objective can be formulated as follows:

$$\mathcal{L}_{RL} = (r(\tilde{y}) - r(\hat{y})) \sum_{t=1}^{M} \log p(\hat{y}_t|\hat{y}_1 ... \hat{y}_{t-1}, X; \theta) \tag{15}$$

where $r(\cdot)$ denotes the reward score calculated by ROUGE-L. Intuitively, minimizing $\mathcal{L}_{RL}$ is equivalent to maximize the conditional likelihood of the sampled sequence $\hat{y}_t$ if it obtains a higher reward than the baseline $\tilde{y}_t$, thus increasing the reward expectation of our model.

Mixed Loss In the training process, we combine all loss functions described above. The composite training objective is as follows:

$$\mathcal{L}_{MIXED} = (1 - \gamma)(\mathcal{L}_{NLL} + \lambda_1 \mathcal{L}_{Coverage} + \lambda_2 \mathcal{L}_{GCL}) + \gamma \mathcal{L}_{RL} \tag{16}$$

where λ_1, λ_2 and γ are tunable hyper parameters.

4 Experiments

4.1 Dataset

We perform experiments on the large-scale dataset CNN/Daily Mail [9], which is widely used in abstractive document summarization with multi-sentence summaries. For the data prepossessing, we utilize the scripts provided by [20] to obtain the non-anonymized dataset version[3], which contains 287,226 training pairs, 12,368 validation pairs and 11,490 test pairs. In addition, the average number of sentences in documents and summaries are 42.1 and 3.8, respectively.

4.2 Baselines

To validate the correctness and effectiveness of our model, we choose the following representative and competitive frameworks for comparison. **PGN+Cov** [20] proposes a novel architecture that extends the standard seq2seq attention model with pointer mechanism and coverage loss. **ML+RL** [6] allows users to define attributes of generated summaries and applies the copy mechanism for source entities. **Fast-Abs** [4] selects salient sentences and then rewrites them abstractively (i.e., compresses and paraphrases) to generate a concise summary.

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

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

Table 1. Automatic evaluation of our proposed model against recently released summarization systems on CNN/DailyMail dataset. The best performance is highlighted in bold and the results of all baselines are taken from the corresponding papers.

Models	ROUGE-1	ROUGE-2	ROUGE-L
PGN + Cov	39.53	17.28	36.38
ML + RL	39.87	15.82	36.90
Fast-Abs	40.88	17.80	38.54
GPG	40.95	18.05	37.19
ROUGESal + Ent	40.43	18.00	37.10
Bottom-Up	41.22	18.68	38.34
DCA	41.11	18.21	36.03
Our Model (GIGN)	**42.04**	**19.08**	**39.15**

DCA [3] divides the hard task of encoding a long text across multiple collaborating encoder agents. **GPG** [8] promotes the standard attention model from both local and global aspects to reproduce most salient information and avoid repetitions. **Bottom-Up** [7] equips the seq2seq model with a data-efficient bottom-up content selector. **ROUGESal+Ent** [15] utilizes the reinforcement learning approach with two novel reward functions: ROUGESal and Entail. Notably, due to the limited computational resource, we don't apply the pre-trained contextualized encoder (i.e. BERT [5]) to our model. Thus, we only compare with the models without BERT for the sake of fairness.

4.3 Hyper-parameters Settings

For a fair comparison, we limit the vocabulary size to 50k and initialize the tokens with 128-dimensional Glove embeddings [17]. The dimensions of hidden units are all set to 256 same as [20]. And the number of heads in attention mechanism is 8. During training, we set the batch size to 16 and optimize the model with Adam [10] method that the initial learning rate is 0.1. At test time, we utilize the beam search algorithm to generate summaries and the beam size is set to 5. Moreover, trigram avoidance [16] is used to avoid trigram-level repetition as previous methods. We implement our model on a Tesla V100 GPU.

4.4 Evaluation Metrics

To evaluate our model comprehensively, we adopt both automatic metrics and human judgments in our experiments. For automatic metrics, We evaluate our models with the standard ROUGE metric, measuring the unigram, bigram and longest common subsequence overlap between the generated and reference summaries as ROUGE-1, ROUGE-2 and ROUGE-L, respectively.

Moreover, human judgments can further evaluate the quality of the generated summaries accurately, which has been widely applied in previous works. We invite six volunteers (all CS majored students) as human annotators. For

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

Table 2. Ablation study. The token "+" indicates that we add the corresponding module to the model.

Models	R-1	R-2	R-L
PGN + Cov	39.60	17.47	36.31
+ GSM	40.17	17.86	36.92
+ RGIGM	41.29	18.37	37.91
+ Consistency loss	41.72	**19.14**	38.78
+ RL	**42.04**	19.08	**39.15**

Table 3. The human evaluation results on Relevance(C1), Non-Redundancy (C2) and Readability(C3).

Models	C1	C2	C3
Reference	5.00	5.00	5.00
PGN + Cov	3.78	3.85	4.02
Fast-Abs	3.74	3.48	3.72
Our Model (GIGN)	**4.18**	**4.32**	**4.34**

the fair comparison, given 100 randomly sampled source-target pairs from the CNN/Daily Mail test dataset, volunteers are required to score the results of all models from 1 to 5 based on the following indicators: *Relevance* (**C1**) represents the correlation between the generated summaries and the ground truth. *Non-Redundancy* (**C2**) measures the diversity and informativeness of outputs. And *Readability* (**C3**) mainly evaluates whether the output is grammatically fluent.

5 Results

5.1 Automatic Evaluation

Table 1 presents the results of automatic evaluation on the CNN/DailyMail dataset. Obviously, our model significantly outperforms all the baselines on all metrics, which indicates the gist guided method can effectively generate more fluent summary. For the benchmark model PGN+Cov, our results are improved by 2.51, 1.80 and 2.77 in terms of ROUGE-1, ROUGE-2 and ROUGE-L, respectively. Particularly, the ROUGE-2 brings a 10.4% boost compared with PGN+Cov. Moreover, our model just utilizes end-to-end training instead of the two-step method like [7] to achieve the best results. In summary, we only add a few parameters to the baseline model, but we get a great improvement.

5.2 Ablation study

We conduct the ablation study to evaluate the correctness and effectiveness of different modules. On the basic model PGN + Cov, we gradually add the GSM module, RGIGM mechanism, gist consistency loss and reinforcement learning. As shown in Table 2, we first add the GSM module to distill the salient gist information to guide the whole generation process, which achieves better results than the baseline. Taking the ROUGE-L for an example, the score exceeds the basic model by 0.61 points. Hence, the model can generate more coherent and fluent summaries by introducing the GSM module. Then, we introduce the RGIGM for the decoder, which brings a great performance improvement. It proves the remaining gist guided method makes the decoder concern about the information to be expressed next. Finally, we equip the model with a gist consistency loss to

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

Table 4. The bold words in Article are salient parts contained in Reference Summary. The blue words in generated summaries are salient information and the red words are uncorrelated or error.

Article: **A video that was played during a preliminary hearing in a california courtroom on friday** showed a san diego police officer being hit with his own cruiser. **Officer Jeffrey Swett was allegedly run over by William Bogard in january after the suspect stole his car while it was running, according to prosecutors**. **Swett suffered two broken arms, a broken leg and severe head and neck trauma**, while Bogard has pleaded not guilty. Scroll down for video. A video from a hearing in a court on friday showed a san diego police officer being hit with his own cruiser. **William Bogard has pleaded not guilty after being charged with attempted murder, assault and vehicle theft**. Officer jeffrey ...
Reference Summary: Officer Jeffrey Swett was allegedly run over by William Bogard in january. Suspect stole officer 's car while it was running, according to prosecutors. Swett suffered broken arms, broken leg and severe head and neck trauma. Video of incident was played during preliminary hearing in court on friday. Bogard pleaded not guilty to charges including murder, assault and theft.
PGN+Cov: Officer jeffrey swett was charged with attempted murder, assault and vehicle theft. Swett suffered two broken arms, a broken leg and severe head and neck trauma. Bogard has pleaded not guilty after being charged with attempted murder.
Fast-Abs: Officer jeffrey swett was allegedly run over by William Bogard in january. Swett suffered two broken arms, a broken leg and severe head and neck trauma. William bogard has pleaded not guilty after being charged with attempted murder. Video shows san diego police officer being hit with his own cruiser. Bogard was smiling behind the wheel while running him down.
Our model: Officer Jeffrey Swett was allegedly run over by Billiam Bogard in january after the suspect stole his car while it was running, according to prosecutors. Swett suffered two broken arms, a broken leg and severe head and neck trauma, while Bogard has pleaded not guilty. The suspect was charged with attempted murder, assault and vehicle theft. The video was played during preliminary hearing in courtroom on friday.

further improve the consistency between original documents and generated summaries that ensures the salient information is expressed completely. It is worth noting that the model has achieved state-of-the-art results at this time. We also verify reinforcement learning can further promote the performance of our model.

5.3 Human Evaluation

The human evaluation results are calculated by averaging all scores from six annotators and the scores of reference summaries are set to 5. As shown in Table 3, our model significantly outperforms all baseline models we have implemented, especially in terms of the C1 (Relevance) and C2 (Non-Redundancy). Moreover, C1 measures the correlation between the generated summaries and the ground truth, while C2 evaluates the diversity and informativeness of outputs. Therefore, the high scores of C1 and C2 suggest that the gist information can make the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

model pay more attention to salient information of the input documents and the RGIGM mechanism improves the diversity of results by reducing semantic duplication. Furthermore, all scores of our model are very close to the ground truth, which indicates that our model can generate relevant, diverse and fluent summaries as human beings.

5.4 Case Study

To verify whether the performance improvements are owing to the gist information, we show a sample of summaries generated by our model and baseline models. As shown in the Table 4, without the guidance of remaining gist information, PGN+Cov fails to obtain some pieces of salient information and even generates false facts (officer jeffrey swett is a victim, not a suspect). Moreover, Fast-Abs not only losts the salient information, but also generates a number of trivial facts. By contrast, our model, with the guidance of gist, can avoid redundancy and generate summaries containing most pieces of salient information.

6 Conclusion

In this paper, we propose a novel framework that first introduces the gist concept in abstractive summarization. We propose the self-adjust query in multi-head attention mechanism to distill the salient semantics as global gist and calculate the remaining gist to guide the generation process dynamically, which can effectively reduce the redundancy and improve the readability. And the gist consistency loss further improves the consistency between documents and summaries. We conduct experiments on the CNN/Daily Mail dataset and the results indicate that our method significantly outperforms all baselines.

In the future, we can extend the gist guided method in many directions. An appealing direction is to investigate the abstractive method on the multi-document summarization, which is more challenging and lacks training data.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993, http://portal.acm.org/citation.cfm?id=944937
2. Cao, Z., Li, W., Li, S., Wei, F.: Retrieve, rerank and rewrite: Soft template based neural summarization. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 152–161 (2018)
3. Celikyilmaz, A., Bosselut, A., He, X., Choi, Y.: Deep communicating agents for abstractive summarization. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1662–1675. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1150, https://www.aclweb.org/anthology/N18-1150

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

4. Chen, Y.C., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 675–686. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1063, https://www.aclweb.org/anthology/P18-1063
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
6. Fan, A., Grangier, D., Auli, M.: Controllable abstractive summarization. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. pp. 45–54. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/W18-2706, https://www.aclweb.org/anthology/W18-2706
7. Gehrmann, S., Deng, Y., Rush, A.: Bottom-up abstractive summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4098–4109. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1443, https://www.aclweb.org/anthology/D18-1443
8. Gui, M., Tian, J., Wang, R., Yang, Z.: Attention optimization for abstractive document summarization. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 1222–1228 (2019). https://doi.org/10.18653/v1/D19-1117, https://doi.org/10.18653/v1/D19-1117
9. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in neural information processing systems. pp. 1693–1701 (2015)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
11. Krishna, K., Srinivasan, B.V.: Generating topic-oriented summaries using neural attention. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1697–1705. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1153, https://www.aclweb.org/anthology/N18-1153
12. 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, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2009, https://www.aclweb.org/anthology/N18-2009

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5

13. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain (July 2004), https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/
14. Nallapati, R., Xiang, B., Zhou, B.: Sequence-to-sequence rnns for text summarization. ArXiv **abs/1602.06023** (2016)
15. Pasunuru, R., Bansal, M.: Multi-reward reinforced summarization with saliency and entailment. 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. 646–653. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2102, https://www.aclweb.org/anthology/N18-2102
16. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=HkAClQgA-
17. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
18. Perez-Beltrachini, L., Liu, Y., Lapata, M.: Generating summaries with topic templates and structured convolutional decoders. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5107–5116. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1504, https://www.aclweb.org/anthology/P19-1504
19. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 379–389. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1044, https://www.aclweb.org/anthology/D15-1044
20. 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, Vancouver, Canada (Jul 2017)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
22. Wang, K., Quan, X., Wang, R.: Biset: Bi-directional selective encoding with template for abstractive summarization. arXiv preprint arXiv:1906.05012 (2019)
23. Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., Du, Q.: A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 4453–4460. International Joint Conferences on Artificial Intelligence Organization (7 2018). https://doi.org/10.24963/ijcai.2018/619, https://doi.org/10.24963/ijcai.2018/619
24. You, Y., Jia, W., Liu, T., Yang, W.: Improving abstractive document summarization with salient information modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2132–2141 (2019)
25. Zheng, C., Zhang, K., Wang, H.J., Fan, L.: Topic-aware abstractive text summarization (2020)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_5