Improving Neural Abstractive Document Summarization with Explicit Information Selection Modeling*

Wei Li[1,2,3] **Xinyan Xiao**[2] **Yajuan Lyu**[2] **Yuanzhuo Wang**[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Baidu Inc., Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
weili.ucas.ict@gmail.com, {xiaoxinyan,lvyajan}@baidu.com,
wangyuanzhuo@ict.ac.cn

Abstract

Information selection is the most important component in document summarization task. In this paper, we propose to extend the basic neural encoding-decoding framework with an information selection layer to explicitly model and optimize the information selection process in abstractive document summarization. Specifically, our information selection layer consists of two parts: gated global information filtering and local sentence selection. Unnecessary information in the original document is first globally filtered, then salient sentences are selected locally while generating each summary sentence sequentially. To optimize the information selection process directly, distantly-supervised training guided by the golden summary is also imported. Experimental results demonstrate that the explicit modeling and optimizing of the information selection process improves document summarization performance significantly, which enables our model to generate more informative and concise summaries, and thus significantly outperform state-of-the-art neural abstractive methods.

1 Introduction

Document summarization is the task of generating a fluent and condensed summary for a document while retaining the gist information. There are two prominent approaches: extractive methods and abstractive methods. Extractive methods generate summary for a document by directly selecting salient sentences from the original document. On the contrary, abstractive methods synthesize information from the input document to generate summary using arbitrary words and expressions - as human usually do. Recent neural models enable an end-to-end framework for natural language generation, which inspires the research on abstractive document summarization.

Most existing work directly apply the neural encoding-decoding framework, which first encodes the input into an abstract representation and then decodes the output based on the encoded information. Although the encoding-decoding framework has achieved huge success on some text generation tasks like machine translation (Bahdanau et al., 2014) and image caption (Vinyals et al., 2015), the performance on abstractive document summarization is much less convincing. Since document summarization is a special natural language generation task that requires information selection, the performance of current neural abstractive methods even has a considerable gap from extractive methods.

The most essential prerequisite for a practical document summarization system is that the generated summary must contain the salient information of the original document. Since a document is a long sequence of multiple sentences, both global document information and local inter-sentence relations need to be properly modeled in the information selection process. Although the encoding-decoding framework has implicitly modeled the information selection process via end-to-end training, we argue that abstractive document summarization shall benefit from explicitly modeling and optimizing it by capturing both the global document information and local inter-sentence relations.

In this paper, we propose to extend the encoding-decoding framework to model the information selection process explicitly. We treat the document summarization as a three-phase task: document encoding, information selection and summary decoding. Correspondingly, our model consists of three layers: a document encoder layer, an information selection layer and a

*This work was done while the first author was doing internship at Baidu Inc.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1787–1796
Brussels, Belgium, October 31 - November 4, 2018.

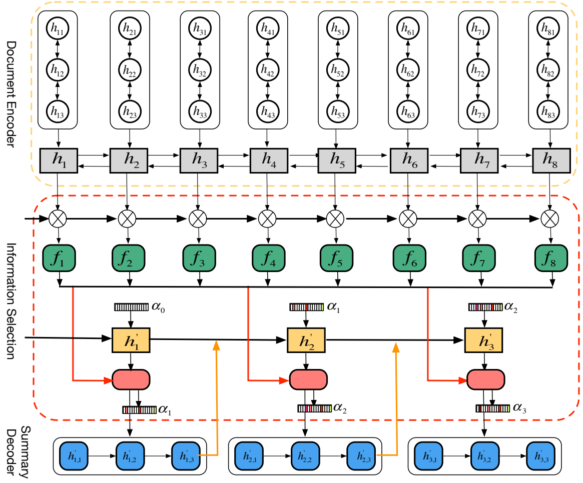

Document Encoder
Information Selection
Summary Decoder
h_{11} h_{12} h_{13}
h_{21} h_{22} h_{23}
h_{31} h_{32} h_{33}
h_{41} h_{42} h_{43}
h_{51} h_{52} h_{53}
h_{61} h_{62} h_{63}
h_{71} h_{72} h_{73}
h_{81} h_{82} h_{83}
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8
f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8
α_0 α_1 α_2
h'_1 h'_2 h'_3
α_1 α_2 α_3
$h'_{1,1}$ $h'_{1,2}$ $h'_{1,3}$
$h'_{2,1}$ $h'_{2,2}$ $h'_{2,3}$
$h'_{3,1}$ $h'_{3,2}$ $h'_{3,3}$

Figure 1: Our abstractive document summarization model, which mainly consists of three layers: document encoder layer (the top part), information selection layer (the middle part) and summary decoder layer (the bottom part).

summary decoder layer, as shown in Figure 1. In our model, both the document and summary are processed sentence by sentence, to better capture the inter-sentence relations. The information selection layer consists of two parts: gated global information filtering and local sentence selection. Unnecessary information in the original document are first globally filtered by a gated network, then important sentences are selected locally while generating each summary sentence sequentially. Moreover, we propose to optimize the information selection process with distantly-supervised training. Our proposed method combines the strengths of extractive methods and abstractive methods, which is able to tackle the factors of saliency, non-redundancy, coherence and fluency under a unified framework. We conduct extensive experiments on benchmark datasets and the results demonstrate that the explicit modeling and distantly-supervised optimizing of the information selection process improves document summarization performance significantly, which enables our model to significantly outperforms previous state-of-the-art neural abstractive methods.

2 Our Model

As shown in Figure 1, our model consists of a hierarchical document encoder, an information selection layer and an attention-equipped decoder. Firstly, the hierarchical encoder encodes the document sentence by sentence, and word by word in each sentence. Then the information selection layer selects and filters the sentence representations based on the global document representation. A sentence selection RNN is used to select salient and relevant sentences while generating each summary sentence sequentially based on the tailored sentence representations. At last, the summary decoder produces the output summary to paraphrase and generalize the selected sentences.

In the following, we denote h_i, $h_{i,j}$ as the hidden state of the i-th sentence and the j-th word of the i-th sentence in the document encoder part, respectively. In the information selection and summary decoder part, we denote h'_t, $h'_{t,k}$ as the hidden state of the t-th summary sentence and the k-th word in the t-th summary sentence, respectively.

2.1 Document Encoder

A document d is a sequence of sentences $d = \{s_i\}$, and each sentence is a sequence of words $s_i = \{w_{i,j}\}$. A hierarchical encoder, which consists of two levels: word level and sentence level similar to (Nallapati et al., 2016), is used to encode the document from both word and sentence level.

The word-level encoder is a bidirectional Gated Recurrent Unit (GRU) (Chung et al., 2014), which encodes the words of a sentence into sentence representation. The word encoder sequentially updates its hidden state after receiving a word, which is formulated as:

$$h_{i,j} = BiGRU(h_{i,j-1}, e_{i,j}) \tag{1}$$

where $h_{i,j}$ and $e_{i,j}$ denotes the hidden state and embedding of word $w_{i,j}$, respectively.

The concatenation of the forward and backward final hidden states in the word-level encoder is indicated as the vector representation x_i of the sentence s_i, which is used as input to the sentence-level encoder. The sentence encoder is also a bidirectional GRU, which updates its hidden state after receiving each sentence representation by:

$$h_i = BiGRU(h_{i-1}, x_i) \tag{2}$$

where h_i denotes the hidden state of sentence s_i.

The concatenation of the forward and backward final states in the sentence-level encoder is used as the vector representation of document $\hat{\mathbf{d}}$.

2.2 Information Selection

Document summarization is a special natural language generation task which requires information compression. It needs to remove the unnecessary information and select salient information from the input document to produce a condensed summary. However, it is difficult for the basic encoder-decoder framework to learn the process of salient information selection, which has also been noticed by several previous work (Tan et al., 2017a,b). To tackle the challenge, we extend the basic encoder-decoder framework by adding an information selection layer to model the information selection process explicitly. Our information selection layer consists of two parts: gated global information filtering that used to remove the unnecessary information of a document, and local sentence selection that used to select salient sentences from a document sequentially to produce summary sentences.

Gated Global Information Filtering

Inspired by studies on how human write text summaries by first skimming the document and deleting unnecessary material (Brown and Day, 1983), we design a gated global information filtering network to filter unnecessary information of a document based on the global document representation before the summary decoder generates summary. Concretely, the gated information filtering network makes use of the document representation $\hat{\mathbf{d}}$, which represents the global information of a document, to filter sentences based on the sentence representation h_i.

For each source sentence s_i, the gate network takes the document representation $\hat{\mathbf{d}}$ and sentence representation h_i as inputs to compute the gate vector g_i:

$$g_i = \sigma(W_g h_i + U_g \hat{\mathbf{d}} + b_g) \tag{3}$$

where W_g and U_g denote weight matrices, b_g the bias vector, and σ the sigmoid activation function.

Then each sentence s_i can be filtered by the gate vector g_i as follows:

$$f_i = h_i \odot g_i \tag{4}$$

where f_i indicates the representation of sentence s_i after information filtering, and $\odot$ denotes element-wise multiplication.

Note that, we filter sentences in micro semantic dimensions rather than filtering whole sentences. The tailored sentence representations are used as input to the sentence selection network and summary decoder, which can help to detect salient sentences and improve informativeness of the generated summary.

Local Sentence Selection

We explicitly model the local sentence selection process which selects several target sentences to generate a summary sentence. Concretely, we apply a RNN layer to sequentially select target sentences for each summary sentence, shown as in Figure 1. The sentence-selection RNN uses the document representation $\hat{\mathbf{d}}$ as initial state $h_0^{'}$, and sequentially predicts the sentence selection vector α_t as follows:

$$\alpha_t^i = \frac{e^{\phi(f_i, h_t')}}{\sum_l e^{\phi(f_l, h_t')}} \tag{5}$$

$$\phi(f_i, h_t') = v^T tanh(W_f f_i + W_h h_t' + b). \tag{6}$$

where α_t^i indicates the weight of source sentence s_i when generating the t-th summary sentence, and $h_t^{'}$ denotes the hidden state of sentence selection layer when generating the t-th summary sentence. v, W_f and W_h are weight matrices, and b is the bias vector. Note that, the sentence selection vector α_t is computed based on the tailored sentence representation f_i.

The sentence-selection RNN uses a single unidirectional GRU, which updates its state by:

$$h_t^{'} = GRU(h_{t-1}^{'}, x_t^{'}) \tag{7}$$

where $x_t^{'}$ denotes the input of current sentence-selection step. $x_t^{'}$ combines both the previous sentence selection vector α_{t-1} and the encoded representation of previous generated sentence $r_{t-1}^{'}$ by $x_t^{'} = tanh(W_r r_{t-1}^{'} + W_\alpha \alpha_{t-1} + b_x)$, where W_r, W_α, and b_x denote learnable parameters.

The representation of the selected source sentences is computed by:

$$q_t = \sum_j \alpha_t^j f_j \tag{8}$$

which is used as initial state of the summary decoder to generate a summary sentence to paraphrase and generalize the selected sentences.

2.3 Summary Decoder

On top of the document encoder and the information selection layer, we use GRU with attention as the summary decoder to realize each summary sentence word by word.

At each word decoding step k in the t-th summary sentence, the GRU reads the previous word embedding $e_{t,k-1}$ and context vector $c_{t,k-1}$ as inputs to compute the new hidden state $h^{'}_{t,k}$ by:

$$h^{'}_{t,k} = GRU(h^{'}_{t,k-1}, c_{t,k-1}, e_{t,k-1}) \quad (9)$$

We import attention mechanism to help locate relevant words to be copied or paraphrased within the selected source sentences in each word generation step. The attention distribution $\beta^{i}_{t,k}$ of the kth word of the tth summary sentence over the sentences in the ith document can be computed as:

$$\beta^{i,j}_{t,k} = \alpha^{i}_{t} \frac{e^{\phi(h_{i,j}, h'_{t,k})}}{\sum_{l} e^{\phi(h_{i,l}, h'_{t,k})}} \quad (10)$$

where α^{i}_{t} denotes the weight of the ith source sentence, used to normalize the word attention distributions. Then the word-level context vector when generating the kth word at the tth sentence generation step can be computed as: $\mathbf{c}_{t,k} = \sum_i \sum_j \beta^{i,j}_{t,k} h_{i,j}$, which is also incorporated into the word decoder.

At each word generation step, the vocabulary distribution is calculated from the context vector $\mathbf{c}_{t,k}$ and the decoder state $h^{'}_{t,k}$ by:

$$P_{vocab}(w'_{t,k}) = softmax(W_v(W_c[h^{'}_{t,k}, c_{t,k}] + b_c) + b_v) \quad (11)$$

where W_v and W_c are learned parameters. The copy mechanism based on the word attention is also imported into the decoder to alleviate the OOV problems as in (See et al., 2017).

2.4 Model Learning with Distant Supervision

Despite the end-to-end training for the performance of generated summary, we also directly optimize the sentence selection decisions by importing supervision for the sentence selection vector α_t in Equation 5. While there is no explicit supervision for sentence selection, we define a simple approach for labeling sentences based on the reference summaries. To simulate the sentence selection process on human-written abstracts, we compute the words-matching similarities (based on TF-IDF cosine similarity) between a reference-summary sentence and corresponding source document sentences and normalize them into distantly-labelled sentence selection vector p_t. Then the sentence selection loss is defined as:

$$loss_{sel} = \sum_{t} D_{KL}(\alpha_t, p_t) \quad (12)$$

where $D_{KL}(\alpha_t, p_t)$ indicates the KL-divergence between distribution α_t and p_t. The sentence selection loss is imported into the final loss function to be optimized with the summary generation component together.

The loss function $\mathcal{L}$ of the model is the mix of the negative log-likelihood of generating summaries over training set $\mathcal{T}$, and the sentence selection loss of distantly-supervised training:

$$\mathcal{L} = \sum_{(X,Y)\in\mathcal{T}} -logP(Y|X;\theta) + \lambda loss_{sel} \quad (13)$$

where λ is a hyper-parameter tuned on the validation set. (X, Y) denotes a document-summary pair in the training set.

3 Experiments

3.1 Dataset

We conduct our experiments on a large-scale corpus of *CNN/DailyMail*, which has been widely used for exploration on summarizing documents with multi-sentence summaries. The corpus are originally constructed in (Hermann et al., 2015) by collecting human generated highlights from news stories in the CNN and DailyMail Website, which contains input document of about 800 tokens on average and multi-sentence summaries of up to 200 tokens. We use the same version of data with (See et al., 2017), which totally has 280,125 training pairs, 13,367 validation pairs and 11,489 test pairs after discarding the examples with empty article text. Some of previous work (Nallapati et al., 2016, 2017; Paulus et al., 2017; Tan et al., 2017a) use the anonymized version of data, which has been pre-processed to replace each named entity with an unique identifier. By contrast, we use the non-anonymized data similar to (See et al., 2017), which is a more favorable and challenging problem because it requires no pre-processing.

3.2 Implementation Details

Model Parameters For all experiments, the word-level encoder and summary decoder both use 256-dimensional hidden states, and the sentence-level encoder and sentence selection network both use 512-dimensional hidden states. We use pre-trained Glove (Pennington et al., 2014) vector for initialization of word embeddings. The dimension of word embeddings is 100, which will be further trained in the model. We use a vocabulary of 50k words for both encoder and decoder.

Method	Rouge-1	Rouge-2	Rouge-L
Lead-3	40.34	17.70	36.57
SummaRuNNer-abs	37.5	14.5	33.4
SummaRuNNer	39.6	16.2	35.3
Seq2seq-baseline	36.64	15.66	33.42
ABS-temp-attn	35.46	13.30	32.65
Graph-attention	38.1	13.9	34.0
Deep-reinforced	39.87	15.82	**36.90**
Coverage	39.53	17.28	36.38
Our Model	**41.54**	**18.18**	36.47

Table 1: Rouge F_1 scores on the test set. All our ROUGE scores have a **95% confidence interval of at most $\pm$0.25** as reported by the official ROUGE script.

Method	Informat.	Concise	Coherent	Fluent
Lead-3	3.49*	3.19*	3.86	4.07*
Seq2seq-b.	3.11*	2.95*	3.08*	3.51*
Coverage	3.41*	3.25*	3.37	3.72
Our Model	3.76	3.49	3.65	3.80

Table 2: Human evaluation results. * indicates the difference between **Our Model** and other models are statistic significant ($p < 0.1$) by two-tailed t-test.

We use dropout (Srivastava et al., 2014) with probability $p = 0.5$. After tuning on the validation set, parameter λ is set as 0.2.

Model Training We use Adagrad (Duchi et al., 2011) algorithm with learning rate 0.1 and an initial accumulator value of 0.1 to optimize the model parameters θ. During training, we use gradient clipping with a maximum gradient norm of 2. Our model is trained on a single Tesla K40m GPU with a batch size of 16 and an epoch is set containing 10,000 randomly sampled documents. Convergence is reached within 300 epochs.

Hierarchical Beam Search To improve information correctness and avoid redundancy during the summary decoding process, we use the hierarchical beam search algorithm with reference mechanism (Tan et al., 2017a) to generate multi-sentence summaries. Similar to (Tan et al., 2017a), the beam sizes for word decoder and sentence decoder are 15 and 2, respectively.

3.3 Baselines

We compare our system with the results of state-of-the-art neural summarization approaches reported in recent papers, which contain both abstractive models and extractive models. The extractive models include **SummaRuNNer** (Nallapati et al., 2017), while **SummaRuNNer-abs** is similar to SummaRuNNer but is trained directly on the abstractive summaries. **Lead-3** is a strong extractive baseline which uses the first 3 sentences of the document as summary. The abstractive models include:

1) **Seq2seq-baseline**, which uses the basic seq2seq encoder-decoder structure with attention mechanism and incorporates with the copy mechanism as in (See et al., 2017).
2) **ABS-temp-attn** (Nallapati et al., 2016), which uses *Temporal Attention* on the seq2seq architecture to overcome the repetition problem.
3) **Graph-attention** (Tan et al., 2017a), which uses a graph-ranking based attention mechanism based on a hierarchical architecture to identify important source sentences.
4) **Deep-reinforced** (Paulus et al., 2017), which trains the seq2seq encoder-decoder model with reinforcement learning techniques.
5) **Coverage** (See et al., 2017), which is an extension of the Seq2seq-baseline model by importing coverage mechanism to control repetitions in summary.

3.4 Evaluation

ROUGE Evaluation

We evaluate our models with the standard ROUGE metric (Lin, 2004) and obtain ROUGE scores using the `pyrouge` package. Results in Table 1 show that our method has significant improvement over state-of-the-art neural abstractive baselines as well as extractive baselines. Note that, the **Deep-reinforced** model achieves the best ROUGE-L performance because it directly optimizes the ROUGE-L metric. Comparing with the current state-of-the-art model **Coverage**, our model achieves significant better performance on ROUGE-1 and ROUGE-2 metrics, and comparable performance on ROUGE-L metric, which demonstrates that our model is more effective in selecting salient information from a document to produce an informative summary while keeping the ability to generate fluent and correct sentences.

Human Evaluation with Case Analysis

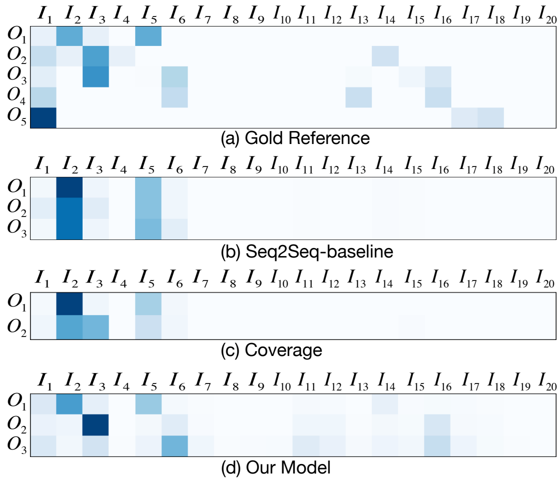
In addition to the ROUGE evaluation, we also conducted human evaluation on 50 random samples from CNN/DailyMail test set and compared the summaries generated by our method with the outputs of **Lead-3**, **Seq2seq-baseline** and **Coverage**. Three data annotators were asked to compare the generated summaries with the human summaries, and assess each summary from four independent

Gold Reference: faith and hope howie were born with one body and two faces on may 8 . they tragically died in hospital just 19 days after they were born . parents simon howie and renee young visit their grave at pinegrove in western sydney fortnightly . they arrived on thursday to find the grave bare of all the girls ' mementos . staff had cleared entire baby section and thrown belongings in rubbish .
Seq2Seq-baseline: faith and hope howie were dubbed the miracle twins when they were born on may 8 last year with one body and two faces due to an extremely rare condition known as disrosopus . faith and hope howie were dubbed the miracle twins when they were born on may 8 last year with one body and two faces due to an extremely rare condition known as disrosopus . faith and hope howie were dubbed the miracle twins when they were born on may 8 last year with one body and two faces due to an extremely rare condition known as disrosopus .
Coverage: faith and hope howie were dubbed the miracle twins when they were born on may 8 last year with one body and two faces due to an extremely rare condition known as disrosopus . they died in hospital less than a month after they were born and their parents , simon howie and renee young , laid them to rest at pinegrove memorial park in sydney 's west.
Our Model: faith and hope howie were dubbed the miracle twins when they were born on may 8 last year with one body and two faces due to an extremely rare condition known as disrosopus. they died in hospital less than a month after they were born and their parents , simon howie and renee young , laid them to rest at pinegrove memorial park in sydney 's west. family members have visited the grave every week to leave mementos and flowers for faith and hope , but when mr howie and ms young arrived on thursday they found the site completely bare .

Table 3: Examples of generated summaries. The Seq2Seq-baseline model generates repeated sentences and loses salient information. The Coverage model reduces repetitions, but also loses salient information. Our model can select more salient information from the original document and generate more informative summary.

perspectives: (1) **Informative**: How informative the summary is? (2) **Concise**: How concise the summary is? (3) **Coherent**: How coherent (between sentences) the summary is? (4) **Fluent**: How fluent, grammatical the sentences of a summary are? Each property is assessed with a score from 1(worst) to 5(best) by three annotators. The average results are presented in Table 2.

The results show that our model consistently outperforms the **Seq2seq-baseline** model and the previous state-of-the-art method **Coverage**. An example of comparison of the generated summaries by our model with the two abstractive models (w.r.t the reference summary) is shown in Table 3[1]. The summary generated by **Seq2Seq-Baseline** usually contains repetition of sentences, which seriously affects its informativeness, conciseness as well as coherence. For example, the sentence "*faith and hope howie were dubbed the miracle twins when they were born ...*" is repeated three times in Table 3. The **Coverage** model effectively alleviates the information repetition problem, however, it loses some salient information that should be included in the summary. For example, the information about "**mementos**" and "**family members visit the grave**" is lost in the example shown in Table 3. The summary generated by our method obviously contains more salient information, which shows the effectiveness of the information selection component in our model. According to the results in Table 2, the sentence-level modeling of document and summary in our model also makes the generated summaries achieve better inter-sentence coherence. Compared with the strong extractive baseline **Lead-3**, our model is able to generate more informative and concise summaries, which shows the advantage of abstractive methods. The fluency scores also show the good ability of our model to generate fluent and grammatical sentences. The human evaluation results demonstrate that our model is able to generate more informative, concise and coherent summaries than the baselines.



(a) Gold Reference

(b) Seq2Seq-baseline

(c) Coverage

(d) Our Model

Figure 2: Visualization of sentence selection vectors. I_i and O_i indicate the i-th sentence of the input and output, respectively. Obviously, our model can detect more salient sentences that are included in the reference summary.

The visualization of the sentence selection vectors of the gold reference summary and the three abstractive models when generating the presented examples in Table 3 are shown in Figure 2[2]. The figure shows that **Seq2Seq-baseline** fails to detect all important source sentences and attend to the same sentences repeatedly, which result in generating repeated summary sentences. **Coverage** learns to reduce repetitions, but fails to detect all the salient information. Obviously, our method is more effective in selecting salient and relevant source sentences from the document to generate more informative summary. Furthermore, our

[1] More examples are shown in the supplementary material

[2] The sentence selection vectors of the Seq2seq-baseline mode and the Coverage model are computed by summing the attention weights of all words in each sentence and then normalized across sentences.

Method	Rouge-1	Rouge-2	Rouge-L
Our Model	**41.54**	**18.18**	**36.47**
– *distS*	40.02	17.54	34.87
– *distS&gateF*	39.26	16.96	33.92
– *infoSelection*	36.64	15.66	33.42

Table 4: Comparison results of removing different components of our method.

method tends to focus on different sets of source sentences when generating different summary sentences. The results verify that the information selection component in our model significantly improves the information selection process in document summarization.

4 Discussion

In this section, we first validate the effectiveness of each component of our model, then compare the performance of information selection of our method with several extractive methods, and finally analyze the effects of golden summary length on the performance of our model.

4.1 Model Validation

To further verify the effectiveness of each component in our model, we conduct several ablation experiments. "– *distS*" denotes removing the distant supervision for sentence selection (set λ as 0). "– *distS&gateF*" denotes removing both the distant supervision for sentence selection training and the global gated information filtering component. "– *infoSelection*" denotes removing the whole information selection layer and do not explicitly modeling the information selection process, which is actually the Seq2seq-baseline model.

Results on the test set are shown in Table 4. Our method much outperforms all the comparison systems and removing each component of our model one by one will leads to sustained significant performance declining, which verifies the effectiveness of each component in our model. The global gated information filtering network removes unnecessary information from the original document and helps generate more informative summary. The distantly-supervised training for sentence selection decisions helps the model learn to detect important and relevant source sentences for each summary sentence. The results verify that explicitly modeling the information selection process significantly improves the document summarization performance.

Method	Rouge-1	Rouge-2	Rouge-L
SummaRuNNer-abs	37.5	14.5	33.4
SummaRuNNer	39.6	16.2	35.3
OurExtractive	**40.41**	**18.30**	36.30
– *distS*	37.06	16.55	33.23
– *distS&gateF*	36.25	16.22	32.59

Table 5: Comparsion results of sentence selection.

length	Method	Rouge-1	Rouge-2	Rouge-L
< 75	Our Mod.	**39.90**	**16.91**	35.19
(81.82%)	Coverage	38.90	16.81	**35.82**
$[75, 100)$	Our Mod.	**47.13**	**22.44**	**40.81**
(12.64%)	Coverage	42.89	19.72	39.41
$[100, 125)$	Our Mod.	**50.49**	**24.23**	**43.68**
(4.00%)	Coverage	41.78	19.00	38.41
> 125	Our Mod.	**50.25**	**23.98**	**41.19**
(1.54%)	Coverage	39.57	17.93	36.33

Table 6: Comparison results w.r.t different length of reference summary. < 75 indicates the reference summary has less than 75 words (occupy 81.82% of test set), $[75, 100)$ denotes the number of words in reference summary is between 75 and 100 (occupy 12.64% of test set).

4.2 Effectiveness of Information Selection

To verify the performance of sentence selection in our model, we add a comparison system **OurExtractive** which is almost the same as our model, but replaces the summary decoder by a sentence extractor. The sentence extractor extracts the source sentence with the largest weight in each sentence generation step. "– *distS*" denotes removing the distant supervision for sentence selection training in our model. "– *distS&gateF*" denotes removing both the distant supervision for sentence selection training and the gated global information filtering component.

Results in Table 5 show that our simple extractive method **OurExtractive** significantly outperforms state-of-the-art neural extractive baselines, which demonstrates the effectiveness of the information selection component in our model. Moreover, **OurExtractive** significantly outperforms the two comparison systems which remove different components of our model one by one. The results show that both the gated global information filtering and distant supervision training are effective for improving information selection in document summarization. Our proposed method effectively combines the strengths of extractive methods and abstractive methods into a unified framework.

4.3 Effects of Summary Length

We further compare our method with the **Coverage** model by evaluating them on the test set with different length of golden reference summaries. The results are shown in Table 6, which demonstrate that our method is better at generating long summary for long document. As the golden summary becoming longer, our system will obtain larger advantages over the baseline (from +1.0 Rouge-1, +0.1 Rouge-2 and -0.63 Rouge-L for summary less than 75 words, rising to +10.68 Rouge-1, +6.05 Rouge-2 and +4.86 Rouge-L for summaries more than 125 words). The results also verify that our method is more effective in selecting salient information from documents, especially for long documents.

5 Related Work

Existing exploration on document summarization mainly can be categorized to **extractive methods** and **abstractive methods**.

5.1 Extractive Summarization Methods

Neural networks have been widely investigated on extractive document summarization task. Earlier work attempts to use deep learning techniques to improve sentence ranking or scoring (Cao et al., 2015a,b; Yin and Pei, 2015). Some recent work solves the sentence extraction and document modeling in an end-to-end framework. Cheng and Lapata (2016) propose an encoder-decoder approach where the encoder hierarchically learns the representation of sentences and documents while an attention-based sentence extractor extracts salient sentences sequentially from the original document. Nallapati et al. (2017) propose a recurrent neural network-based sequence-to-sequence model for sequential labelling of each sentence in the document. Neural models are able to leverage large-scale corpora and achieve better performance than traditional methods.

5.2 Abstractive Summarization Methods

As the seq2seq learning with neural networks achieve huge success in sequence generation tasks like machine translation, it also shows great potential in text summarization area, especially for abstractive methods. Some earlier researches studied the use of seq2seq learning for abstractive sentence summarization (Takase et al., 2016; Rush et al., 2015; Chopra et al., 2016). These models are trained on a large corpus of news documents which are usually shortened to be the first one or two sentences, and their headlines.

Later, some work explored the seq2seq models on document summarization, which produce a multi-sentence summary for a document. The seq2seq models usually exhibit some undesirable behaviors, such as inaccurately reproducing factual details, unable to deal with out-of-vocabulary (OOV) words and repetitions. To alleviate these issues, copying mechanism (Gu et al., 2016; Gulcehre et al., 2016; Nallapati et al., 2016) has been incorporated into the encoder-decoder architecture. Distraction-based attention model (Chen et al., 2016) and coverage mechanism (See et al., 2017) have also been investigated to alleviate the repetition problem. To better train the seq2seq model on tasks with long documents and multi-sentence summaries, a deep reinforced model was proposed to combine the standard words predication with teacher forcing learning and the global sequence prediction training with reinforcement learning (Paulus et al., 2017). Recently, Tan et al. (2017a) propose to leverage the hierarchical encoder-decoder architecture on generating multi-sentence summaries, and incorporate sentence-ranking into the summary generation process based on the graph-based attention mechanism. Different from these neural-based work, our model explicitly models the information selection process in document summarization by extending the encoder-decoder framework with an information selection layer. Our model captures both the global document information and local inter-sentence relations, and optimize the information selection process directly via distantly-supervised training, which effectively combines the strengths of extractive methods and abstractive methods.

6 Conclusion

In this paper, we have analyzed the necessity of explicitly modeling and optimizing of the information selection process in document summarization, and verified its effectiveness by extending the basic neural encoding-decoding framework with an information selection layer and optimizing it with distantly-supervised training. Our information selection layer consists of a gated global information filtering network and a local RNN sentence selection network. Experimental results demonstrate that both of them are effective for helping select salient information during the summary generation process, which significantly improves the document summarization performance. Our model combines the strengths of extractive methods and abstractive methods, which can generate more informative and concise summaries, and thus achieves state-of-the-art abstractive document summarization performance and is also competitive with state-of-the-art extractive models.

Acknowledgments

This work was supported by National Key Research and Development Program of China under grants 2016YFB1000902 and 2017YFC0820404, and National Natural Science Foundation of China under grants 61572469, 91646120, 61772501 and 61572473. We thank the anonymous reviewers for their helpful comments about this work.

References

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

Ann L Brown and Jeanne D Day. 1983. Macrorules for summarizing texts: The development of expertise. *Journal of verbal learning and verbal behavior*, 22(1):1–14.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015a. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159.

Ziqiang Cao, Furu Wei, Sujian Li, Wenjie Li, Ming Zhou, and WANG Houfeng. 2015b. Learning summary prior representation for extractive summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 829–833.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for document summarization. *arXiv preprint arXiv:1610.08462*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

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.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

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.

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

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. *arXiv preprint arXiv:1603.08148*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

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

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *AAAI*, 1:1.

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*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *EMNLP*, pages 1054–1059.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017a. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1171–1181.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017b. From neural sentence summarization to headline generation: A coarse-to-fine approach. *IJCAI*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *IJCAI*.