Improving Neural Abstractive Document Summarization with Structural Regularization*

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

Recent neural sequence-to-sequence models have shown significant progress on short text summarization. However, for document summarization, they fail to capture the long-term structure of both documents and multi-sentence summaries, resulting in information loss and repetitions. In this paper, we propose to leverage the structural information of both documents and multi-sentence summaries to improve the document summarization performance. Specifically, we import both *structural-compression* and *structural-coverage* regularization into the summarization process in order to capture the information compression and information coverage properties, which are the two most important structural properties of document summarization. Experimental results demonstrate that the structural regularization improves the document summarization performance significantly, which enables our model to generate more informative and concise summaries, and thus significantly outperforms 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. Recent success of neural sequence-to-sequence (seq2seq) architecture on text generation tasks like machine translation (Bahdanau et al., 2014) and image caption (Vinyals et al., 2015), has attracted growing attention to abstractive summarization research. Huge success has been witnessed in abstractive sentence summarization (Rush et al., 2015; Takase et al., 2016; Chopra et al., 2016; Cao et al., 2017; Zhou et al., 2017), which builds one-sentence summaries from one or two-sentence input. However, the extension of sentence abstractive methods to document summarization task is not straightforward.

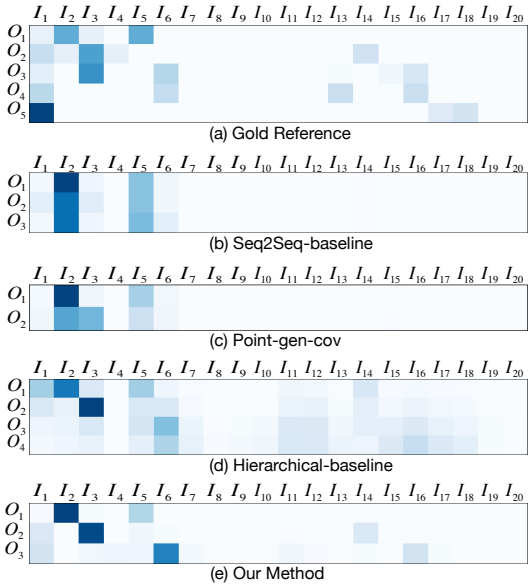

Figure 1: Comparison of sentence-level attention distributions for the summaries in Table 1 on a news article. (a) is the heatmap for the gold reference summary, (b) is for the *Seq2seq-baseline* system, (c) is for the *Point-gen-cov* (See et al., 2017) system, (d) is for the *Hierarchical-baseline* system and (e) is for our system. I_i and O_i indicate the i-th sentence of the input and output, respectively. Obviously, the seq2seq models, including the *Seq2seq-baseline* model and the *Point-gen-cov* model, lose much salient information of the input document and focus on the same set of sentences repeatedly. The *Hierarchical-baseline* model fails to detect several specific sentences that are salient and relevant for each summary sentence and focuses on the same set of sentences repeatedly. On the contrary, our method with structural regularizations focuses on different sets of source sentences when generating different summary sentences and discovers more salient information from the document.

As long-distance dependencies are difficult to be captured in the recurrent framework (Bengio et al., 1994), the seq2seq models are not yet able to achieve convincing performance in encoding and decoding for a long sequence of multiple sentences (Chen et al., 2017; Koehn and Knowles,

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 4078–4087
Brussels, Belgium, October 31 - November 4, 2018.

Original Text (truncated): the family of conjoined twin sisters who died 19 days after they were born have been left mortified$_{(2)}$ after they arrived at their gravesite to find cemetery staff had cleared the baby section of all mementos and tossed them in the rubbish$_{(3)}$. 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~~ $_{(1)}$. 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$_{(2)}$. scroll down for video . 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~~$_{(1)}$. 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$_{(3)}$. ' we took renee's aunts to see the girls for the first time and we found everything had been stripped away , ' mr howie told daily mail australia . ' we were devastated and mortified . we 've had a little shrine set up and we 've been adding to it since the funeral . ' it 's heartbreaking to know we 've set this up and it has been treated like rubbish . ' faith and hope were buried in a pink coffin and their family and friends released doves and pink and white balloons at their funeral . their family and friends had built up a small memorial with pink and white statues , flowers , pebbles and toys over the past 11 months . when they arrived on thursday , everything had been removed apart from a bunch of flowers . the twins were buried at pinegrove memorial park in western sydney after they died after just 19 days$_{(2)}$. their family and friends had built a small shrine at their gravesite , which they have added to since the funeral . family members have visited the grave every week to leave mementos and flowers for faith and hope , but when parents simon howie and renee young arrived on thursday they found the site completely bare$_{(3)}$.
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 .
Point-cov (See et al., 2017): 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.
Hierarchical-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 . they died in hospital less than a month after they died in hospital less than a month after they were born and 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 they were born on thursday they found the site completely bare . family members have visited the grave every week to leave mementos and flowers for faith and hope , but when they found the site completely bare .
Our Method: 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$_{(1)}$. they died in hospital less than a month after they were born and their parents laid them to rest at pinegrove memorial park in sydney 's west$_{(2)}$. 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$_{(3)}$.

Table 1: Comparison of the generated summaries of four abstractive summarization models and the gold reference summary on a news article. The summaries generated by the seq2seq models, both the Seq2seq-baseline model and the Point-cov model, lose some salient information. The Seq2seq-baseline model even contains serious information repetitions. The Hierarchical-baseline model not only contains serious repetitions, but also makes non-grammatical or non-coherent sentences. On the contrary, the summary generated by our model contains more salient information and is more concise. Our model also shows the ability to generate a summary sentence by compressing several source sentences, such as ~~shortening a long sentence~~.

2017). In document summarization, it is also difficult for the seq2seq models to discover important information from too much input content of a document (Tan et al., 2017a,b). The summary generated by the seq2seq models usually loses salient information of the original document or even contains repetitions (see Table 1).

In fact, both document and summary naturally have document-sentence hierarchical structure, instead of being a flat sequence of words. It is widely aware that the hierarchical structure is necessary and useful for neural document modeling. Hierarchical neural models have already been successfully used in document-level language modeling (Lin et al., 2015) and document classification (Yang et al., 2016). However, few work makes use of the hierarchical structure of document and multi-sentence summary in document summarization. The basic hierarchical encoder-decoder model (Li et al., 2015) is also not yet able to capture the structural properties of both document and summary (see Figure 1[1]), resulting in more serious repetitions and even nonsensical sentences (see Table 1).

In document summarization, information compression and information coverage are the two most important structural properties. Based on the hierarchical structure of document and summary, they can be realized at the sentence-level as: (1) *Structural-compression*: each summary sentence is generated by compressing several specific source sentences; (2) *Structural-coverage*: different summary sentences usually focus on different sets of source sentences to cover more salient information of the original document. Figure 1(a) intuitively shows the two properties in human-written gold reference summaries. We import both structural-compression and structural-coverage regularizations into the document summarization process based on a hierarchical encoder-decoder with hybrid sentence-word attention model. Typically, we design an effective learning and inference algorithm to explicitly model the structural-compression and structural-coverage properties of document summarization process, so as to generate more informative and concise summaries (see Table 1).

We conduct our experiments on benchmark datasets and the results demonstrate that properly modeling the *structural-compression* and *structural-coverage* properties based on the hier-

[1]To simulate the sentence-level attention mechanism on the gold reference summary, we compute the words-matching similarities (based on TF-IDF cosine similarity) between a reference-summary sentence and the corresponding source document sentences and normalize them into attention distributions. The sentence-level attention distributions of the Seq2seq-baseline model and the Point-gen-cov model are computed by summing the attention weights of all words in each sentence and then normalized across sentences.

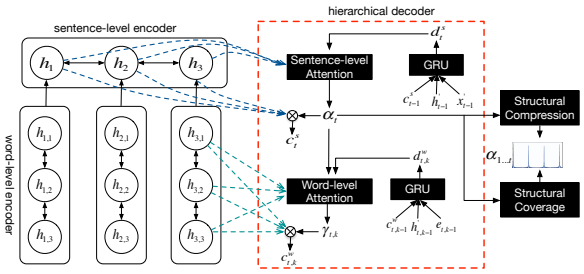

Figure 2: Our hierarchical encoder-decoder model with structural regularization for abstractive document summarization.

archical structure of document and summary, improves document summarization performance significantly. Our model is able to generate more informative and concise summaries by enhancing sentences compression and coverage, and significantly outperforms state-of-the-art seq2seq-based abstractive methods, especially on summarizing long documents with long summaries.

2 Hierarchical Encoder-Decoder Model

In this section, we introduce our baseline hierarchical encoder-decoder model which consists of two parts: a hierarchical encoder and a hierarchical decoder, as shown in Figure 2. Similar to (Li et al., 2015), both the encoder and decoder consists of two levels: a sentence level and a word level. The main distinction is that we design a hybrid sentence-word attention mechanism on the hierarchical decoder to help organize summary content and realize summary sentences.

2.1 Hierarchical Encoder

The goal of the encoder is to map the input document to a hidden vector representation. We consider a source document X as a sequence of sentences: $X = \{s_i\}$, and each sentence s_i as a sequence of words: $s_i = \{w_{ij}\}$. The word-level encoder encodes the words of a sentence into a sentence representation, and the sentence-level encoder encodes the sentences of a document into the document representation. In this work, both the word-level encoder and sentence-level encoder use the bidirectional Gated Recurrent Unit (BiGRU) (Chung et al., 2014). The word-level encoder sequentially updates its hidden state upon each received word, as $h_{i,j} = BiGRU(h_{i,j-1}, e_{i,j})$ where $h_{i,j}$ and $e_{i,j}$ denote the hidden state and the 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 r_i of sentence s_i, which is used as input to the sentence-level encoder. The sentence-level encoder updates its hidden state after receiving each sentence representation, as $h_i = BiGRU(h_{i-1}, r_i)$ 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 $\mathbf{d}$.

In the hierarchical encoder architecture, long dependency problem will be largely reduced at both the sentence level and the word level, so it can better capture the structural information of the input document.

2.2 Hierarchical Decoder with Hybrid Sentence-Word Attention

The goal of the decoder is to generate output summary according to the representation of the input document. Let $Y = \{s'_i\}$ indicates a candidate summary of document X, and each sentence s'_i consists of a sequence of words $s'_i = \{w'_{ij}\}$. The hierarchical decoder organizes summary Y sentence by sentence, and realizes each sentence word by word. In this work, both the sentence-level decoder and word-level decoder use a single layer of unidirectional GRU. The sentence-level decoder receives the document representation $\mathbf{d}$ as initial state h'_0 and predicts sentence representations sequentially by $h'_t = GRU(h'_{t-1}, r'_{t-1})$, where h'_t denotes the hidden state of the tth summary sentence s'_t and r'_{t-1} denotes the encoded representation of the previously generated sentence s'_{t-1}. The word-level decoder receives a sentence representation h'_t as initial state $h'_{t,0}$ and predicts word representations sequentially by $h'_{t,k} = GRU(h'_{t,k-1}, e_{t,k-1})$ where $h'_{t,k}$ denotes the hidden state of word $w'_{t,k}$ in sentence s'_t and $e_{t,k-1}$ denotes the embedding of previously generated word $w'_{t,k-1}$ in sentence s'_t.

In this work, we design a hybrid sentence-word attention mechanism based on the hierarchical encoder-decoder architecture, which contains both **sentence-level attention** and **word-level attention**, to better exploit both the sentence-level information and word-level information from the input document and the output summary.

2.2.1 Sentence-level Attention

The sentence-level attention mechanism is designed on the sentence-level encoder and decoder,

which is used to help our model to detect important and relevant source sentences in each sentence generation step. α_t^i indicates how much the t-th summary sentence attends to the i-th source sentence, which is computed by $\alpha_t^i = e^{f(h_i,h'_t)}/\sum_l e^{f(h_l,h'_t)}$ where f is the function modeling the relation between h_i and h'_t. We use the function $f(\mathbf{a},\mathbf{b}) = v^T tanh(W_a\mathbf{a} + W_b\mathbf{b})$, where v, W_a, W_b are all learnable parameters. Then the sentence level context vector c_t^s when generating the tth sentence s'_t can be computed as: $c_t^s = \sum_i \alpha_t^i h_i$, which is incorporated into the sentence-level decoding process.

2.2.2 Word-level Attention with Sentence-level Normalization

The word-level attention is designed on the word-level encoder and decoder, which is used to help our model to realize the summary sentence by locating relevant words in the selected source sentences in each word generation step. Let $\beta_{t,k}^{i,j}$ denotes how much the j-th word in source sentence s_i contributes to generating the k-th word in summary sentence s'_t, which is computed by $\beta_{t,k}^{i,j} = e^{f(h_{i,j},h'_{t,k})}/\sum_l e^{f(h_{i,l},h'_{t,k})}$.

Since the word-level attention above is within each source sentence, we normalize it by sentence-level attentions to get word attention over all source words, as $\gamma_{t,k}^i = \beta_{t,k}^i \alpha_t^i$. Then the word-level context vector when generating word $w'_{t,k}$ can be computed as: $\mathbf{c}_{t,k}^w = \sum_i \sum_j \gamma_{t,k}^{i,j} h_{i,j}$, which is also incorporated into the word-level decoding process.

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

$$P_{vocab}(w'_{t,k}) = softmax(W_v(W_c[h_{t,k}^{'}, c_{t,k}^w] + b_c) + b_v) \quad (1)$$

where W_v, W_c, b_c and b_v are learned parameters. We also incorporate the copy mechanism (See et al., 2017) based on the normalized word-level attention to help generate out-of-vocabulary (OOV) words during the sentence realization process.

3 Structural Regularization

Although the above hierarchical encoder-decoder model is designed based on the document-sentence hierarchical structure, it can't capture the basic structural properties of document summarization (see Figure 1(d) and Table 1). However, the hierarchical architecture makes it possible for importing structural regularization to capture the sentence-level characteristics of document summarization process. In this work, we propose to model the *structural-compression* and *structural-coverage* properties based on the hierarchical encoder-decoder model by adding structural regularization during both the model learning phase and inference phase.

3.1 Structural Compression

Compression is a basic property of document summarization, which has been widely explored in traditional document summarization research, such as sentence compression-based methods which shorten sentences by removing non-salient parts (Li et al., 2013; Durrett et al., 2016) and sentence fusion-based methods which merge information from several different source sentences (Barzilay and McKeown, 2005; Cheung and Penn, 2014). As shown in Figure 1, each summary sentence in the human-written reference summary is also created by compressing several specific source sentences.

In this paper, we propose to model the *structural-compression* property of document summarization based on sentence-level attention distributions by:

$$strCom(\alpha_t) = 1 - \frac{1}{logN}\sum_{i=1}^{N} \alpha_t^i log\alpha_t^i \quad (2)$$

where α_t denotes the sentence-level attention distribution when generating the tth summary sentence and N denotes the length of distribution α_t. The right part in the above formula is actually the entropy of the distribution α_t. As the attention distribution becomes sparser, the entropy of the distribution becomes lower, so the value of $strCom(\alpha_t)$ defined above will become larger. Sparse sentence-level attentions help the model compress and generalize several specific source sentences which are salient and relevant in the sentence generation process. Note that, $0 \leq strCom(\alpha_t) \leq 1$.

3.2 Structural Coverage

A good summary should have the ability to cover most of the important information of an input document. As shown in Figure 1, the human-written reference summary covers the information of many source sentences. Coverage has been

used as a measure in many traditional document summarization research, such as the submodular-based methods which optimize the information coverage of the summary with similarity-based coverage metrics (Lin and Bilmes, 2011; Chali et al., 2017).

In this work, we simply model the *structural-coverage* property of summary based on the hierarchical architecture by encouraging different summary sentences to focus on different sets of source sentences so that the summary can cover more salient sentences of the input document. We measure the *structural-coverage* of summary based on the sentence-level attention distributions:

$$strCov(\alpha_t) = 1 - \sum_i \min(\alpha_t^i, \sum_{t'=0}^{t-1} \alpha_{t'}^i) \quad (3)$$

which is used to encourage different summary sentences to focus on different sets of source sentences during the summary generation process. As the sentence-level attention distributions of different summary sentences become more diversified, the summary will cover more source sentences, which is effective to improve the informativeness and conciseness of summaries. Note that, $0 \leq strCov(\alpha_t) \leq 1$.

3.3 Model Learning

Experimental results reveal that the properties of *structural-compression* and *structural-coverage* are hard to be captured by both the seq2seq models and the hierarchical encoder-decoder baseline model, which largely restricts their performance (Section 4). In this work, we model them explicitly by regulating the sentence-level attention distributions based on the hierarchical encoder-decoder framework. The loss function $\mathcal{L}$ of the model is the mix of negative log-likelihood of generating summaries over training set $\mathcal{T}$, the structural-compression loss and the structural-coverage loss:

$$\mathcal{L} = \sum_{(X,Y)\in\mathcal{T}} \{-logP(Y|X;\theta) + \underbrace{\lambda_1 \sum_t strCom(\alpha_t)}_{structural-compression\ loss} + \underbrace{\lambda_2 \sum_t strCov(\alpha_t)}_{structural-coverage\ loss} \} \quad (4)$$

where λ_1 and λ_2 are hyper-parameters tuned on the validation set. We use Adagrad (Duchi et al., 2011) with learning rate 0.1 and an initial accumulator value 0.1 to optimize the model parameters θ.

3.4 Hierarchical Decoding Algorithm

The traditional beam search algorithm that widely used for text generation can only help generate fluent sentence, and is not easy to extend to the sentence level. The reason is that the K-best sentences generated by a word decoder will mostly be similar to each other (Li et al., 2016; Tan et al., 2017a). We propose a hierarchical beam search algorithm with *structural-compression* and *structural-coverage* regularization.

The hierarchical decoding algorithm has two levels: K-best word-level beam search and N-best sentence-level beam search. At the word-level, the vanilla beam search algorithm is used to maximize the accumulated score $\hat{P}(s'_t)$ of generating current summary sentence s'_t. At the sentence-level, N-best beam search is realized by maximizing the accumulated score $score_t$ of all the sentences generated, including the sentences generation score, structural-compression score and structural-coverage score, which are defined as:

$$score_t = \sum_{t'=0}^{t} \{\hat{P}(s'_{t'}) + \zeta_1 strCom(\alpha_{t'}) + \zeta_2 strCov(\alpha_{t'})\} \quad (5)$$

where ζ_1 and ζ_2 are factors introduced to control the influence of structural regularization during the decoding process.

4 Experiments

4.1 Experimental Settings

We conduct our experiments on the *CNN/Daily Mail* dataset (Hermann et al., 2015), which has been widely used for exploration on summarizing documents with multi-sentence summaries (Nallapati et al., 2016; See et al., 2017; Tan et al., 2017a; Paulus et al., 2017). The CNN/DailyMail dataset contains input sequences of about 800 tokens in average and multi-sentence summaries of up to 200 tokens. The average number of sentences in documents and summaries are 42.1 and 3.8, respectively. We use the same version of non-anonymized data (the original text without pre-processing) as See et al. (2017), which has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs.

For all experiments, the word-level encoder and decoder both use 256-dimensional hidden states, and the sentence-level encoder and decoder both

Method	Rouge-1	Rouge-2	Rouge-L
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
Point-cov	39.53	17.28	36.38
Hierachical-baseline	34.95	14.79	32.68
Our Model	**40.30**	**18.02**	**37.36**

Table 2: Rouge F_1 scores on the test set. All our ROUGE scores have a **95% confidence interval of at most ± 0.25** as reported by the official ROUGE script.

length	Method	Rouge-1	Rouge-2	R.-L
< 100	Our M.	**39.66**	**17.28**	**36.69**
(94.47%)	Point-cov	39.44	17.20	36.30
$[100, 125)$	Our M.	**43.07**	**19.96**	**39.47**
(4.00%)	Point-cov	41.78	19.00	38.41
$[125, 150)$	Our M.	**43.25**	**19.21**	**40.08**
(1.07%)	Point-cov	41.31	18.02	37.75
> 150	Our M.	**40.64**	**18.30**	**38.00**
(0.46%)	Point-cov	35.64	17.76	33.12

Table 3: Comparison results w.r.t different length of reference summary. < 100 indicates the reference summary has less than 100 words (occupy 94.47% of test set).

use 512-dimensional hidden states. The dimension of word embeddings is 128, which is learned from scratch during training. We use a vocabulary of 50k words for both the encoder and decoder.

We trained our model 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. After tuning on the validation set, parameters λ_1, λ_2, ζ_1 and ζ_2, are set as -0.5, -1.0, 1.2 and 1.4, respectively. At the test time, we use the hierarchical decoding algorithm with sentence-level beam size 4 and word-level beam size 8.

4.2 Evaluation

ROUGE Evaluation. We evaluate our models with the widely used ROUGE (Lin, 2004) toolkit. We compare our system's results with the results of state-of-the-art neural summarization approaches reported in recent papers, including both abstractive models and extractive models. The extractive models include **SummaRuNNer** (Nallapati et al., 2017) and **SummaRuNNer-abs** which is similar to SummaRuNNer but is trained directly on the abstractive summaries. The abstractive models include:

1) **Seq2seq-baseline**, which uses the basic seq2seq encoder-decoder architecture with attention mechanism, and incorporates with copy mechanism (See et al., 2017) to alleviate the OOV problem.
2) **ABS-temp-attn** (Nallapati et al., 2016), which uses *Temporal Attention* on the seq2seq architecture to overcome the repetition problem.
3) **Point-cov** (See et al., 2017), which is an extension of the Seq2seq-baseline model by importing word-coverage mechanism to reduce repetitions in summary.
4) **Graph-attention** (Tan et al., 2017a), which uses a graph-ranking based attention mechanism based on a hierarchical architecture to identify important sentences.
5) **Hierachical-baseline**, which just uses the basic hierarchical encoder-decoder with hybrid attention model proposed in this paper.

Results in Table 2 show that our model significantly outperforms all the neural abstractive baselines and extractive baselines. An interesting observation is that the performance of the **Hierarchical-baseline** model are lower than the **Seq2seq-baseline** model, which demonstrates the difficulty for a traditional model to identify the structural properties of document summarization process. Our model outperforms the **Hierarchical-baseline** model by more than 4 ROUGE points, which demonstrates that the structural regularization improves the document summarization performance significantly.

To verify the superiority of our model on generating long summaries, we also compare our method with the best seq2seq model **Point-cov** (See et al., 2017) by evaluating them on a test set w.r.t. different length of reference summaries. The results are shown in Table 3, which demonstrate that our model is better at generating long summary than the seq2seq model. As the summary becomes longer, our system will obtain larger advantages over the baseline (from +0.22 Rouge-1, +0.08 Rouge-2 and +0.39 Rouge-L for summary less than 100 words, rising to +5.00 Rouge-1, +0.54 Rouge-2 and +4.88 Rouge-L for summaries more than 150 words).

Human Evaluation. 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 **Seq2seq-baseline** and **Point-cov** (See et al., 2017). Three data annotators were asked to compare the generated summaries

Method	Informat.	Concise	Coherent	Fluent
Seq2seq-b.	2.79*	2.52*	2.68*	3.57
Point-cov	3.17*	2.92*	3.00*	3.54
Our Model	3.67	3.39	3.51	3.70

Table 4: Human evaluation results. * indicates the difference between **Our Model** and other models are statistic significant ($p < 0.05$) by two-tailed t-test.

Method	R-1	R-2	R-L	strCom	strCov
Hierarchical-b.	34.95	14.79	32.68	0.22	0.31
+strCom	37.03	16.21	34.44	0.64	0.71
+strCov	39.52	17.12	36.44	0.65	0.87
+hierD	**40.30**	**18.02**	**37.36**	**0.68**	**0.93**

Table 5: Results of adding different components of our method in terms of ROUGE-1, ROUGE-2, ROUGE-L, strCom (Equation 1) and strCov (Equation 2) scores.

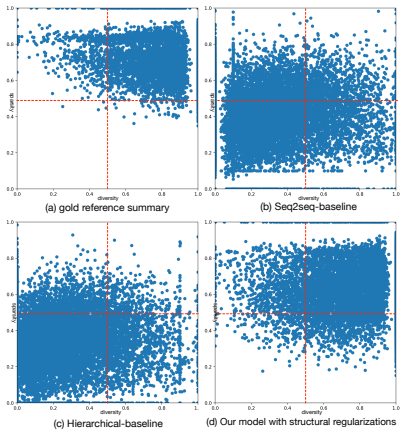

Figure 3: Comparisons of structural-compression and structural-coverage analysis results on random samples from CNN/Daily Mail datasets, which demonstrate that both the Seq2seq-baseline model and the Hierarchical-baseline model are not yet able to capture them properly, but our model with structural regularizations achieves similar behavior with the gold reference summary.

with the human summaries, and assess each summary from four independent 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). The average results are presented in Table 4.

The results show that our model consistently outperforms the **Seq2seq-baseline** model and the previous state-of-the-art method **Point-cov**. As shown in Table 1, the summary generated by **Seq2Seq-Baseline** usually contains repetition of sentences or phrases, which seriously affects its informativeness, conciseness as well as coherence. The **Point-cov** model effectively alleviates the information repetition problem, however, it usually loses some salient information and mainly copies original sentences directly from the input document. The summaries generated by our method obviously contains more salient information and are more concise through sentences compression, which shows the effectiveness of the structural regularization in our model. The results also show that the sentence-level modeling of document and summary in our model makes the generated summaries achieve better inter-sentence coherence.

5 Discussion

5.1 Model Validation

To verify the effectiveness of each component in our model, we conduct several ablation experiments. Based on the *Hierarchical-baseline* model, several different structural regularizations are added one by one: *+strCom* indicates adding structural-compression regularization during model learning, *+strCov* indicates adding structural-coverage regularization during model learning, *+hierD* indicates using the hierarchical decoding algorithm with both structural-compression and structural-coverage regularizations during inference.

Results on the test set are shown in Table 5. Our method much outperforms all the compared systems, which verifies the effectiveness of each component of our model. Note that, both the structural-compression and structural-coverage regularization significantly affect the summarization performance. The higher structural-compression and structural-coverage scores will lead to higher ROUGE scores. Therefore, we can conclude that the structural-compression and structural-coverage regularization based on our hierarchical model have significant contributions to the increase of ROUGE scores.

5.2 Structural Properties Analysis

We further compare the ability of different models in capturing the structural-compression and structural-coverage properties of document summarization. Figure 3 shows the comparison results of 4000 document-summary pairs with 14771 reference-summary sentences sampled from CNN/Daily Mail dataset. Figure 3(a) shows that most samples (over 95%) fall into the right-top area in human-made summaries, which indicates high structural-compression and structural-

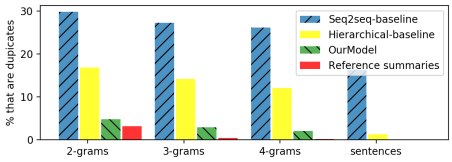

Figure 4: The structural regularization reduces undesirable repetitions while summaries from the Seq2seq-baseline and the Hierarchical-baseline contains many n-gram repetitions.

coverage scores. However, Figure 3 (b) and (c) show that in both the Seq2seq-baseline model and the Hierarchical-baseline model, most samples fall into the left-bottom area (low structural-compression and structural-coverage), and only about 13% and 7% samples fall into the right-top area, respectively. Figure 3 (d) shows that our system with structural regularization achieves similar behaviors to human-made summaries (over 80% samples fall into the right-top area). The results demonstrate that the structural-compression and structural-coverage properties are common in document summarization, but both the seq2seq models and the basic hierarchical encoder-decoder models are not yet able to capture them properly.

5.3 Effects of Structural Regularization

The structural regularization based on our hierarchical encoder-decoder with hybrid attention model improves the quality of summaries from two aspects: (1) The summary covers more salient information and contains very few repetitions, which can be seen both qualitatively (Table 1 and Figure 1) and quantitatively (Table 5 and Figure 4). (2) The model has the ability to shorten a long sentence to generate a more concise one or compress several different sentences to generate a more informative one by merging the information from them. Table 6 shows several examples of abstractive summaries produced by sentence compression in our model.

6 Related Work

Recently some work explored the seq2seq models on document summarization, which exhibit some undesirable behaviors, such as inaccurately reproducing factual details, OOVs 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 to help generate information correctly. Distraction-based attention model

Original Text: luke lazarus ~~, a 23-year-old former private school boy ,~~ was jailed for at least three years on march 27 for raping an 18-year-old virgin in an alleyway outside his father 's soho nightclub ~~in kings cross , inner sydney~~ in may 2013 .(...) **Summary:** luke lazarus was jailed for at least three years on march 27 for raping an 18-year-old virgin in an alleyway outside his father 's soho nightclub in may 2013 .
Original Text: (...) amy wilkinson , 28 , claimed housing benefit and council tax benefit even though she was living in a home owned by her mother and her partner ~~,~~ ~~who was also working~~ .wilkinson ~~, who was a british airways cabin crew attendant ,~~ was ordered to pay back a total of 17,604 that she claimed over two years when she appeared at south and east cheshire magistrates court last week . (...) **Summary:** amy wilkinson , 28 , claimed housing benefit and council tax benefit even though she was living in a home owned by her mother and her partner . *she* was ordered to pay back a total of 17,604 that she claimed over two years when she appeared at south and east cheshire magistrates court last week .
Original Text: (...) a grand jury charged durst with possession of a firearm by a felon , and possession of both a firearm and an illegal drug : 5 ounces of marijuana , said assistant district attorney chris bowman , spokesman for the district attorney . millionaire real estate heir robert durst was indicted wednesday on the two weapons charges that have kept him in new orleans even though his lawyers say he wants to go to los angeles as soon as possible to face a murder charge there . his arrest related to those charges has kept durst from being extradited to los angeles , where he 's charged in the december 2000 death of longtime friend susan berman .(...) **Summary:** *durst entered his plea during an arraignment in a new orleans court on weapons charges that accused him of possessing both a firearm and an illegal drug , marijuana . the weapons arrest has kept durst in new orleans even though he is charged in the december 2000 death of a longtime friend .*

Table 6: Examples of sentences compression or fusion by our model. The ~~link-through~~ denotes deleting the non-salient part of the original text. The *italic* denotes novel words or sentences generated by sentences fusion or compression.

(Chen et al., 2016) and word-level coverage mechanism (See et al., 2017) have also been investigated to alleviate the repetition problem. Reinforcement learning has also been studied to improve the document summarization performance from global sequence level (Paulus et al., 2017).

Hierarchical Encoder-Decoder architecture is first proposed by Li et al. (2015) to train an auto-encoder to reconstruct multi-sentence paragraphs. In summarization field, hierarchical encoder has first been used to alleviate the long dependency problem for long inputs (Cheng and Lapata, 2016; Nallapati et al., 2016). Tan et al. (2017b) also propose to use a hierarchical encoder to encode multiple summaries produced by several extractive summarization methods, and then decode them into a headline. However, these models don't model the decoding process hierarchically.

Tan et al. (2017a) first use the hierarchical encoder-decoder architecture on generating multi-sentences summaries. They mainly focus on incorporating sentence ranking into abstractive document summarization to help detect important sentences. Different from that, our work mainly tends to verify the necessity of leveraging document structure in document summarization and studies how to properly capture the structural properties of document summarization based on the hierarchical architecture to improve the performance of document summarization.

7 Conclusions

In this paper we analyze and verify the necessity of leveraging document structure in document summarization, and explore the effectiveness of capturing structural properties of document summarization by importing both *structural-compression* and *structural-coverage* regularization based on the proposed hierarchical encoder-decoder with hybrid attention model. Experimental results demonstrate that the structural regularization enables our model to generate more informative and concise summaries by enhancing sentences compression and coverage. Our model achieves considerable improvement over state-of-the-art seq2seq-based abstractive methods, especially on long document with long summary.

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*.

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

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization. *arXiv preprint arXiv:1711.04434*.

Yllias Chali, Moin Tanvee, and Mir Tafseer Nayeem. 2017. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 418–424.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. *arXiv preprint arXiv:1707.05436*.

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*.

Jackie Chi Kit Cheung and Gerald Penn. 2014. Unsupervised sentence enhancement for automatic summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 775–786.

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.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. *arXiv preprint arXiv:1603.08887*.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Chen Li, Fei Liu, Fuliang Weng, and Yang Liu. 2013. Document summarization via guided sentence compression. In *Proceedings of the 2013 Conference on*

Empirical Methods in Natural Language Processing, pages 490–500.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

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.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling.

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*.

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*.

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489.

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