Document Summarization with VHTM: Variational Hierarchical Topic-Aware Mechanism

Xiyan Fu,[1] **Jun Wang,**[2*] **Jinghan Zhang,**[1] **Jinmao Wei,**[1] **Zhenglu Yang**[1*]
[1]College of Computer Science, Nankai University, China
[2]Ludong University, China
{fuxiyan, junwang, jhzhang}@mail.nankai.edu.cn, {weijm, yangzl}@nankai.edu.cn

Abstract

Automatic text summarization focuses on distilling summary information from texts. This research field has been considerably explored over the past decades because of its significant role in many natural language processing tasks; however, two challenging issues block its further development: **(1)** how to yield a summarization model embedding topic inference rather than extending with a pre-trained one and **(2)** how to merge the latent topics into diverse granularity levels. In this study, we propose a variational hierarchical model to holistically address both issues, dubbed VHTM. Different from the previous work assisted by a pre-trained single-grained topic model, VHTM is the first attempt to jointly accomplish summarization with topic inference via variational encoder-decoder and merge topics into multi-grained levels through topic embedding and attention. Comprehensive experiments validate the superior performance of VHTM compared with the baselines, accompanying with semantically consistent topics.

Introduction

Automatic text summarization refines integrant information from long texts to a short summary for convenient understanding. It has provided great benefits for many natural language processing (NLP) tasks, such as text classification, information retrieval, and question answering (Gambhir and Gupta 2017).

Dozens of summarization approaches have been introduced in the literature, and they can be roughly categorized into three groups, i.e., *extractive* (Nallapati, Zhai, and Zhou 2017; Narayan, Cohen, and Lapata 2018b), *abstractive* (Tan, Wan, and Xiao 2017; Cao et al. 2018), and *unified* (See, Liu, and Manning 2017; Chen and Bansal 2018), which focus on picking up original crucial objects, generating new expressions, and combining the process of selection and generation. To tackle the limitations of inaccurate factual details and low relevancy of original document in these traditional approaches, topic-aware models are introduced, in which topic serves as guidance to help generate abundant topic-related words and maintain the original ideas of documents (Harabagiu and Lacatusu 2005; Wang et al. 2009). Recently, some deep learning-based approaches have achieved remarkable summarizing performance and drawn extensive attentions (Wang et al. 2018; Narayan, Cohen, and Lapata 2018a). In this work, we devote our efforts to approach the unified summarization tasks with deep learning by fusing topic information.

Most existing deep learning-based summarizaiton works are assisted by an external topic model, e.g., latent dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). This widely adopted strategy accompanies three major weaknesses:

(1) The summarizing process is driven by the externally generated topics beyond summarization models, which may neither characterize the specified flavor of each article nor satisfy the individual requirement of each task. For example, LDA is performed under the assumption that topics are drawn from multinomial distributions across an article; this apparently does not hold for all of the articles fetched from diverse domains. Furthermore, as to the summarizing task, it prefers the most informative topics, while LDA extracts all of the latent topics.

(2) The pre-trained topics are fetched in the document level, neglecting the non-trival difference between paragraphs (i.e., segments with an equal length). An established belief in summarization is that different paragraphs typically possess distinct levels of importance. Thus, excluding the sub-topics hidden in paragraphs may compromise the summarizing performance. In other words, we need the paragraph-gained topics and document-gained ones.

(3) The fusion effect of a deep learning approach coupled with an arbitrary topic model is doubted according to some empirical evidence (Dieng et al. 2017). We argue that a co-optimization framework that incorporates summarization and topic inference is an applicable solution to dispense with the intricate selection of appropriate topic models.

In this study, we present a general deep learning-based framework equipped with a variational hierarchical topic-aware mechanism, dubbed VHTM. This model aims to accomplish topic inference and summarization in an end-to-end manner via variational encoder decoder (VED) and embed the latent topics into the summarization under different

[*] Corresponding authors.

granularity levels. Specifically, multi-scale topics including paragraph-level and document-level ones are induced to capture local and global semantic and syntactic information of a document, promising a comprehensive insight into its latent key parts. We hierarchically incorporate these induced topic vectors into word embedding and paragraph attention, to expose the critical words and paragraphs for summarization.

We employ topic embedding model to produce document-level topic vector and merge it to traditional dense word embedding obtained by an extension of BERT, which can enhance discriminativeness of every word. Moreover, we separate an document into several paragraphs given that different paragraphs own diverse substances and concern them respectively. To the best of our knowledge, our model is the first to employ topic attention on topic representation of every paragraph and combine its context with traditional attention context (Bahdanau, Cho, and Bengio 2014).

Our main contributions are as follows:

- We introduce a joint model that combines topic inference and summarization in an end-to-end manner. Our work is the first attempt to perform summarization without resorting to a pre-trained topic model and can be expected to advance the research of topic-based summarization.
- We employ a hierarchical topic-aware technique that incorporates topic information into words embedding and paragraph attention. The topic-related parts with different granularities in original documents are positioned and extracted to build appropriate summaries.
- The extensive experiments demonstrate that besides achieving superior summarizing performance, our proposed model can also yield similar topic relevance summaries compared with those written by humans as well as achieve a high training efficiency.

Related Work

Text summarization has been widely researched over the past decades for its significant role in NLP. The extant models for text summarization can be divided into three categories, i.e., extractive, abstractive, and unified.

Extractive-based models adopt a natural summarization strategy, that is, extracting key sentences and objects without any modification. Traditional models, such as machine learning techniques (Conroy and O'leary 2001) and graph-based methods (Wan and Yang 2006), have been predominantly applied in previous research. These models have been recently replaced by deep neural network models. The work in (Nallapati, Zhai, and Zhou 2017) treated summarization as a sequence-labeling task. Narayan, Cohen, and Lapata(2018b) conceptualized summarization as a sentence ranking task via reinforced learning. The work in (Jadhav and Rajan 2018) modeled the interaction between key words and salient sentences by using a new two-level pointer network based architecture.

Abstractive-based models paraphrase the content-related sentences verbatim after comprehending the original document. Various strategies for generating abstractive summaries have been presented, among which the sequence-to-sequence framework is considered as the most effective one (Nallapati et al. 2016). Nallapati et al.(2016) concatenated the embedding vectors and performed discretization manipulation to enrich the encoder. Li et al.(2017) equipped their model with a latent structure incorporating components to discover the common structure. Tan, Wan, and Xiao(2017) proposed a graph-based attention mechanism in the sequence-to-sequence framework. The work in (Cao et al. 2018) retrieved proper existing summaries as candidate soft templates and extended the basic framework to jointly perform template reranking and template-aware summary generation.

Unified models integrate the extractive and abstractive-based approaches by utilizing their respective strengths. That is, they can both copy words from original documents and generate novel words simultaneously. Gu et al.(2016) proposed CopyNet to ensure that certain segments in the input sequence can be selectively replicated in the output sequence. Gulcehre et al.(2016) extended the traditional soft-attention-based shortlist softmax by using pointers over the input sequence. See, Liu, and Manning(2017) improved the unified mechanism by providing an explicit switch probability and recycling the attention distribution as the copy distribution. In addition, some works also endeavor to optimize the summarization models via the generative adversarial network (GAN) or reinforced learning (Liu et al. 2018; Paulus, Xiong, and Socher 2018).

A significant drawback of the aforementioned studies is the ignorance of topic information, which is deemed as the most important signature of documents. To address this issue, some studies have introduced topic-aware multi-document summarization through traditional machine learning strategies (Harabagiu and Lacatusu 2005; Wang et al. 2009). By contrast, our work aims to develop a general deep learning-based framework via a hierarchical topic-aware mechanism. Our work seems to be similar to the works in (Narayan, Cohen, and Lapata 2018a; Wang et al. 2018) that also consider topic information, yet show four essential differences: (1) our model can jointly explore topic inference and summarization in an end-to-end manner, whereas the other works have to depend on the pre-trained topic models; (2) our model employs the topic embedding mechanism to merge dense word vectors and Dirichlet-distributed document-level topic vectors, while Wang et al.(2018) only embeds the words in the topic vocabulary (constructed in advance) and Narayan, Cohen, and Lapata(2018a) simply concatenates word vectors according to their topical relevance; (3) we divide a document into segments and apply topic attention on their topic representations, while Wang et al.(2018) utilizes attention on topic words; and (4) the long-document summarizaiton is approached in this work, rather than the short summarization issues in the above works.

Preliminary

Variation encoder decoder is an extension of variational autoencoder (VAE) (Kingma and Welling 2013) and sequence-to-sequence model, which promises the model variational and achieves transforming the source information X into the target information Y. Intuitively, this extension is available for most real applications, such as machine transla-

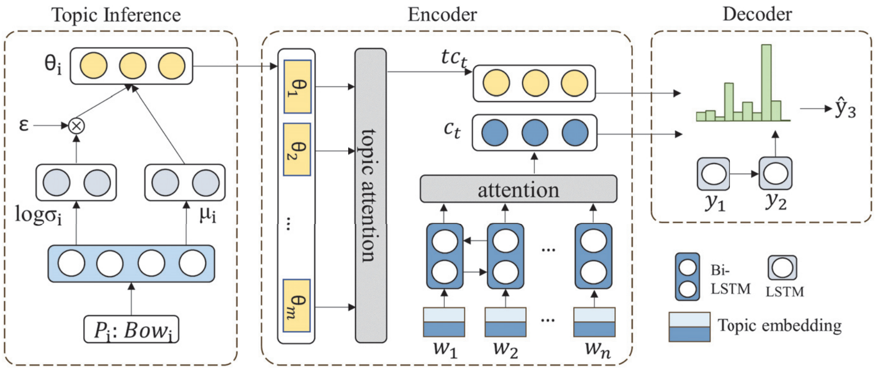

Figure 1: Overview of the variational hierarchical topic-aware model.

tion (Zhang et al. 2016), dialog generation (Cao and Clark 2017), and summarization, even though they own distinct formats between source and target sequences. Following the idea of autoencoder, VED assumes that there exists a continuous latent variable **z** in the underlying semantic space and the training process is $P(Y|z, X)$. Then, the conditional probability is described as

$$p(Y|X) = \int_z p(Y|z, X)p(z)\, dz. \tag{1}$$

This latent variable **z** can bring abundant semantic signal that is complementary for models with deterministic variables. In the training process, it is sampled from a proposal distribution $q(z|X, Y)$ and used for approximating the posterior $p(z|X, Y)$. However, different from dialog generation or machine translation whose information in X and Y is incoordinate, X in summarization tasks has already involved full-scale signals and Y can be ignored. Hence, we follow the assumption in (Bahuleyan 2018) to regard Y as a function of X, as $Y = Y(X)$. Consequently, the proposal distribution $q(z|X, Y) \triangleq q(z|X, Y(X)) \triangleq q(z|X)$. Thus, the evidence lower bound for the log-likelihood of data can be given as

$$\begin{aligned}\log p(Y|X) \geq & -KL(q(z|X)||p(z)) \\ & +\mathbb{E}_{z\sim q} \log p(Y|z, X).\end{aligned} \tag{2}$$

Variational Hierarchical Topic-Aware Model

We propose a variational hierarchical topic-aware framework for automatic text summarization. In the encoder part, the tokens of document $X = (x_1, x_2, ..., x_n)$ are sequentially fed into a single layer Long Short Term Memory (LSTM) network to build its representation $h = (h_1, h_2, ..., h_n)$. The topic embedding block combines the dense word embedding w_i with document embedding t_d to further pinpoint the meaning of words. We also assume that there exists K topic numbers. The topic inference model creates K dimensional topic representations θ_1,...,θ_m for m paragraphs that are applied to topic attention. In the decoder part, which is also constituted by LSTM, the summary is produced by merging the context vector c_t and the topic context vector tc_t additionally. Figure 1 presents a sketch of VHTM.

Encoder

We utilize an encoder to read the tokens of a document and to induce representations for them. We use LSTM as our basic model. As the single directional LSTM suffers from the weakness of extracting contextual information from future tokens, we employ a bidirectional LSTM (BiLSTM) as our recurrent unit. The BiLSTM is constituted by a forward LSTM and a backward LSTM, and thus, it processes the sequence in two directions. The original states of BiLSTM are initialized with uniform distribution, and the representation of each word in the document is obtained by concatenating two kinds of hidden states, i.e., $h_i = [h_i^{fwd}, h_i^{bwd}]$, in the subsequent encoding process.

Topic Inference

Our topic inference model is inspired by the VAE-based neural topic model (Dieng et al. 2017; Miao, Grefenstette, and Blunsom 2017), which uses a latent variable θ as a topic representation. Given the difference between source S and target T, we choose VED as our base framework. Similar to LDA-style topic models, we assume that there exists a K dimensional vector as a topic representation of inputs. The bag-of-words sent to the topic reference block should be removed stop words depicted by S^{NS} to eliminate their negative influence on the topic extraction. The traditional LDA uses Dirichlet distribution as a prior distribution to generate the parameter $\theta \sim Dirichlet(\alpha)$, and we choose the Gaussian distribution $\theta \sim N(\mu, diag(\sigma^2))$ given its high flexibility in the sequence-to-sequence model. We calculate the Gaussian parameters $\mu \in \mathbb{R}^K, \log\sigma \in \mathbb{R}^K$ by applying a linear transformation of $f(S^{NS})$ as

$$\begin{aligned}\mu &= W_1 f(S^{NS}) + b_1, \\ log\sigma &= W_2 f(S^{NS}) + b_2,\end{aligned} \tag{3}$$

where $f(\cdot)$ denotes a three-layer feed-forward neural network. The weights W_1, W_2 and biases b_1, b_2 are learnable parameters shared among different inputs. The latent topic representation can be calculated by applying the reparameterization as

$$\theta = \mu + \sigma \otimes \varepsilon, \quad \varepsilon \sim N(0, I), \tag{4}$$

where $\varepsilon \in \mathbb{R}^K$ is an auxiliary noise variable. The output of the topic inference block θ can be regarded as a dense vector that is filled with an abundant amount of semantic information. Accordingly, the loss function of this block can be expressed as follows:

$$\mathcal{L}_{topic} = KL(q(\theta|S)||p(\theta)) - \mathbb{E}_{\theta \sim q} \log p(T|\theta, S). \tag{5}$$

Hierarchical Topic Mechanism

Topic information is critical in capturing the latent semantic meanings of original documents. Using the related topic information as a guide in generating a summarization can help one understand the meaning of a document and generate an abundant amount of topic-oriented words. Moreover, the topic objects from each level always contain diverse semantic contents that can-not be reflected on their own. Therefore, we embed topic information into a sequence-to-sequence model and then construct a hierarchical structure that advances topic-aware summary generation comprehensively. In sum, our hierarchical topic mechanism includes two components, namely, topic embedding and topic attention, which will be delicately described as below.

Topic Embedding Various models for producing dense word vectors have been proposed and demonstrated to be effective in capturing token-level semantic and syntactic regularities in language. Although these models have been trained on a large-scale text corpus and produce a dedicated long vector to represent words, they still suffer from polysemy and weak relationship with the original document. Inspired by (Moody 2016), we propose the topic embedding mechanism that creates lda2vec which uses the latent topic information in a document for the disambiguation. Figure 2 presents a flow diagram of the mechanism process. Unlike the previously proposed fixed model that is only applicable for training materials, our model demonstrates superior performance for dynamic inputs and can be combined with an advanced pre-trained word embedding model. We use BERT to obtain the pre-trained word embedding because of its excellent performance in our task as demonstrated in the experiments section. By assuming that D denotes a document comprising a sequence of words $X = (x_1, x_2, ..., x_n)$, the topic embedding te_i of each word x_i is formulated as

$$te_i = \varphi(r_d)M_t + w_i, \tag{6}$$

where the word embedding $w_i \in \mathbb{R}^d$ of each word can be acquired by using an extension of BERT and the representation of its belonging document $r_d \in \mathbb{R}^d$ is the average of all included words. A feed forward network $\varphi(\cdot)$ reduces the document representation into a K dimensional vector, which can be regarded as a topic distribution. The learnable parameter topic mapping matrix $M_t \in \mathbb{R}^{K \times d}$ further transforms this vector into a document topic representation vector t_d. We merge the word embedding w_i with the corresponding document representation t_d to produce a topic-specific word embedding instead of a concatenation for in-depth fusion.

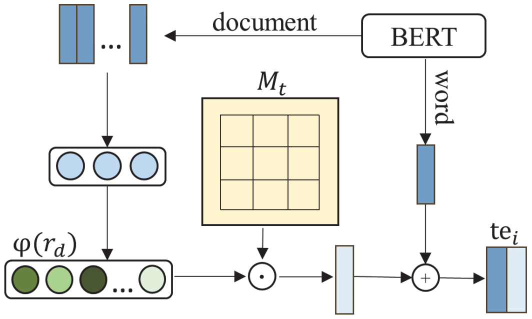

Figure 2: Overview of the topic embedding mechanism: $\varphi(r_d)$ is the topic distribution, M_t is the topic mapping matrix, and te_i is the topic embedding of word x_i.

Topic Attention Documents comprise a combination of several paragraphs, with each paragraph being a self-contained unit of a discourse. Each paragraph shares the same topic represented in the whole document, from the coarse-grained perspective and contains independent sub-points from the fine-grained perspective. Given that the whole document topic information has been included through topic embedding mechanism by document representation t_d, we propose topic attention, which induces the model to take notice of diverse subtopics in the decoding process. We assume that each document includes m paragraphs $p_1, p_2, ..., p_m$, and the $\lfloor \frac{n}{m} \rfloor$ words without stop words in each paragraph are sent to the topic inference block to infer its corresponding topic representation θ_i. We achieve topic attention by following the idea of (Bahdanau, Cho, and Bengio 2014) and obtain the topic context vector as

$$\begin{aligned} tc_t &= \sum_{i=1}^{m} a_i^t \theta_i, \quad a_i^t = softmax(e_i^t), \\ e_i^t &= v^T \tanh(W_h \theta_i + W_s s_t + b_{attn}), \end{aligned} \tag{7}$$

where v, W_h, W_s, and b_{attn} are learnable parameters. Aside from the abundant topic information, we pay attention to information extraction in long texts as well. Similar to topic attention mechanism, we analyze the feedback information between the decoder state s_t and encoder hidden state h_i to determine which parts in the original documents should be concerned. As a result, context vector c_t can be obtained by $c_t = \sum_{i=1}^{n} s(v'^T \tanh(W_h' h_i + W_s' s_t + b_{attn}'))h_i$, where $s(\cdot)$ is the softmax function.

Decoder

After obtaining the topic attention context tc_t and the basic attention context c_t, we feed them to the recurrent decoder part as auxiliary inputs to guide the generation. The final vocabulary distribution is calculated as

$$p(y_t) = softmax(W_p[y_{t-1}; s_{t-1}; c_t; tc_t] + b_p), \tag{8}$$

where y_{t-1} is the embedding of the target word, s_{t-1} denotes the last decoder state, and W_p, b_p represent the weights and bias for linear transformation, respectively.

Similar to most state-of-the-art works, we combine the pointer generator framework with the copy mechanism to directly copy some out of-vocabulary words and to reduce repetition (See, Liu, and Manning 2017). Refer to the related document for additional details.

Joint Training

As shown in the above sections, both the topic inference block and the recurrent deterministic encoder-decoder are designed based on neural networks. Therefore, all parameters in our model can be optimized in an end-to-end manner via back-propagation. Generally, the objective of our framework consists of two terms. One of these terms is the negative loglikelihood of the generated summaries. We use a widely adopted technology called "teaching forcing" for training, which minimizes the maximum-likelihood loss at each decoding step and then computes the overall loss function as

$$\mathcal{L}_{tf} = -\sum_{t=1}^{n'} \log p(y_t^*|y_1^*, ..., y_{t-1}^*, X, \theta_1, ..., \theta_m), \qquad (9)$$

where X is an input sequence, and $\{y_1^*, y_2^*, ..., y_{n'}^*\}$ is the corresponding ground-truth output sequence. The other term is the variational lower bound shown in Eq.(6). Given that we separate the original document into several paragraphs, Eq.(6) must be equal to the sum of each $\mathcal{L}_{topic}$. Since the variational lower bound also contains a likelihood term, we can merge it with the likelihood term of summaries. The final objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{tf} + \sum_{i=1}^{m} KL(q(\theta_i|p_i)||p(\theta_i)). \qquad (10)$$

Experiments

Dataset

Given that long documents contain rich words for semantic information extraction. We choose the **CNN/Daily Mail** corpus as the benchmark dataset.The CNN/Daily Mail dataset comprises online news documents (761 tokens on average) paired with multi-sentence summaries (46 tokens on average). For training and testing efficiency, we use the script of (See, Liu, and Manning 2017), which contains 287,226 training pairs, 13,368 validation pairs, and 11,490 test pairs. This script is a non-anoymized version of the CNN/Daily Mail dataset, which is highly practical and does not require any pre-processing.

Experimental Settings

For the hierarchical topic-aware mechanism of VHTM, we set the topic number K to 50, the dimension of topic representation to 200, and the scale of topic vocabulary to 20,000. Both topic embedding and topic attention share the same topic-related parameters. The dimension of $f(\cdot)$, which represents a three-layer feed-forward neural network, is set equal to the topic dimension. Meanwhile, we set the paragraphs number to 3 for the topic attention mechanism of VHTM. Because of the dimension of the word dense vector obtained from BERT is 768, we set the same dimension for the topic representation of each document for better fusion. In terms of the basic framework, we follow the settings in (See, Liu, and Manning 2017).

Evaluation

We employ **ROUGE** (Lin 2004) to comprehensively evaluate the proposed model. ROUGE is a popular automatic evaluation method that calculates the overlaps between peer and model summaries. We apply the ROUGE-1(R-1), ROUGE-2(R-2), and ROUGE-L(R-L) F-scores in the follow-up experiments to evaluate the overlapping of one word, bi-gram, and the longest common subsequence between decode summary and reference.

Baseline

In this paper, we compare our proposed model VHTM with following baselines:
words-lvt2k-temp-att (Nallapati et al. 2016) models abstractive text summarization using attentional encoder-decoder recurrent neural network, which is the basic framework for most other works.
ConvS2S + fixed control (Fan, Grangier, and Auli 2017) introduces a controlled summarization model (constant value for length and source-style) based on a convolutional sequence-to-sequence network.
pg + cov (See, Liu, and Manning 2017) proposes a hybrid pointer-generator network to deal with Out-Of-Vocabulary words problem and designs coverage mechanism to avoid words repetition.
pg + EG + QG (Guo, Pasunuru, and Bansal 2018) applies multi-task learning with the auxiliary tasks of question generation and entailment generation, leading to salient questioning-worthy details and ability of rewriting.
pg + GAN (Liu et al. 2018) learns an adversarial process, in which train a generative model and a discriminative model simultaneously.
pg + cbdec (Jiang and Bansal 2018) adds an additional 'closed-book' decoder forcing the encoder to be more selective in the information encoded in its memory state.

Results

Quantitative Analysis

We evaluate VHTM with other baseline models that mainly focus on unified summarization. Table 1 shows the experimental results based on the non-anoymized CNN/Daily Mail corpus. To evaluate the impact of VHTM on basic sequence-to-sequence framework, we compare VHTM with a classical RNN-based model *words-lvt2k-temp-att*, and a novel CNN-based model *ConvS2S* and its variant. Furthermore, we compare with *pointer-generator + coverage (pg + cov)*, which is a significant baseline and also the basic framework of VHTM. Table 1 presents that VHTM leads to +1.04, +0.77, +0.80 improvements over *pg + cov* in terms of R-1, R-2 and R-L. This indicates that by using topic embedding and topic attention, VHTM can produce high-quality summaries, since its joint model can provide relevant topic information

Model	R-1	R-2	R-L
words-lvt2k-temp-att	36.64	15.66	33.42
ConvS2S	38.23	16.68	34.77
ConvS2S + fixed control	39.75	17.29	36.54
pg	36.44	15.66	33.42
pg + cov	39.53	17.28	36.38
pg + EG + QG (w/ cov)	39.81	17.64	36.54
pg + GAN (w/ cov)	39.92	17.65	36.71
pg + cbdec (w/ cov)	40.05	17.66	36.73
VHTM	**40.57**	**18.05**	**37.18**

Table 1: Comparison of our VHTM with other baselines models. All the experiments are evaluated by the official ROUGE on the CNN/Daily Mail dataset.

by considering multi-grained features. Besides, we compare with some other models that focus on improving *pg + cov*. Although they are well constructed and simultaneously consider multi-task, adversarial process, and information selection, they are inferior to our model as reported in Table 1.

Ablation Study

To evaluate the effects of each component in VHTM, we perform the ablation study. Given that most current works are based on *pg + cov*, which is weak in abstractive ability and tends to copy words from the source document with high probability (low p_{gen}), we remove the pointer mechanism in ablation study and shorten training steps for simplicity.

Table 2 presents the results. We first build a base model (Base), which is a sequence-to-sequence model accompanied with traditional attention mechanism. In the first block, we test the models with pre-trained word dense embedding from Google[1] (BG) and BERT extension (BB) as mentioned above. We choose the latter as the elementary vector, considering its great power in semantic extraction. When the models are equipped with the hierarchical topic-aware mechanism, namely topic embedding (BB+topic_emb) and topic attention (BB+topic_attn) in the second block in Table 2, these models and their mixture (BB+topic_emb&attn) achieve higher ROUGE scores compared with the baselines.

Intuitively, topic embedding is beneficial for uni-gram generation, while achieves slight improvements. We speculate that current language models have already captured significant semantic information for uni-gram, while bi-gram still needs more assistance from topic materials. However, topic attention on paragraph level is more comprehensive, which can significantly improve the performance under all ROUGE metrics.

Evaluation of Training methods

We conduct experiments to assess how the performance of VHTM is affected by the training technique, and Table 3 shows the results. In the first two lines, we evaluate the effect of topic distribution vector under the loss function Eq.(9). Considering that topic representation of document is related to both the topic inference and topic embedding, we attempt to reuse the topic probability in these two modules named *joint_topic_prob*. Unexpectedly, we observe that its performance is slightly inferior to *indie_topic_inference*. We assume that, different granularities are better not to be mixed in case of the internecine results. After all, only a small performance decline is observed, joint topic distribution can be considered when the efficiency is prior to high accuracy.

Model	R-1	R-2	R-L
Base+Google_emb(BG)	24.14	7.98	22.02
Base+BERT_emb(BB)	26.76	8.24	24.11
BB+topic_emb	26.88	9.40	24.58
BB+topic_attn	28.42	10.08	25.86
BB+topic_emb&attn	29.13	10.68	26.47

Table 2: Ablation study for the hierarchical topic-aware mechanism of VHTM on the CNN/Daily Mail dataset.

Model	R-1	R-2	R-L
joint_topic_prob	27.16	9.69	24.83
indie_topic_inference	27.53	9.73	25.06
joint_training	28.42	10.08	25.86

Table 3: Ablation study for the joint training technique of VHTM on the CNN/Daily Mail dataset.

Given that VHTM consists of two modules as Figure 1 illustrates, i.e., topic inference block and recurrent deterministic encoder-decoder, we apply the joint training loss function Eq.(10) to optimize VHTM. The second part in Table 3 demonstrates that joint training strategy improves the performance with almost 1 point under ROUGE-1, revealing that the topic inference block can provide useful document representation and supplement the missing semantic information.

Topic Mechanism Parameters

In this paper, we propose the hierarchical topic mechanism to merge topic information into multi-grained levels. To evaluate the impact of topic information in VHTM, we evaluate the sensibility of following key hyper parameters:
topic number In the topic embedding or topic attention mechanism, the topic number K plays a significant role because of its potential effects on the convergence rate and summarizing performance. We change the value of K from 20 to 150 independently, and conduct a sequence of experiments to evaluate their effects. The performance of VHTM evaluated by ROUGE when varing K value is reported in Table 4. It shows a unimodal trend, peaking at 50. Besides, a slight decline trend is observed when we increase K continuously, which may suggest that a proper topic number is sufficient for capturing the most significant topic distribution, while a larger one may distract and bring some noises.
paragraph number To further explore the effect of topic attention on paragraph level, we implement experiments specific to paragraph number with a scale from 2 to 5. Table 5 shows the results. Considering that DCA (Celikyilmaz et al. 2018) also takes advantage of paragraph separation, we compare our paragraph number experiments to

[1] https://code.google.com/archive/p/word2vec/

Topic Num	R-1	R-2	R-L
20	36.78	15.02	33.98
50	38.69	16.56	35.62
70	37.75	15.89	34.76
100	37.52	15.92	34.58
150	37.25	15.49	34.35

Table 4: Effects of different topic numbers of VHTM on the CNN/Daily Mail dataset.

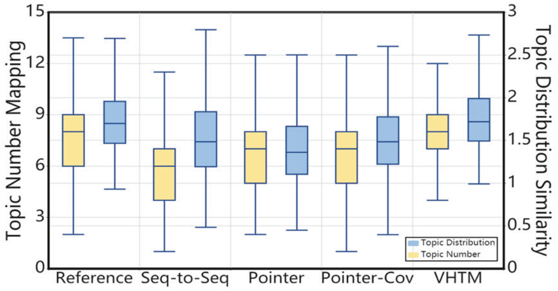

Figure 3: The similarity of topic distributions and the topic number mapping between documents and summaries generated by human or the learning models.

explore further information. VHTM achieves the best performance when we separate the whole document into three paragraphs, which share the same results with DCA. However, different from the similar performance of two extreme paragraph numbers (2 and 5) in DCA, we observe that larger paragraphs will weaken the effects. We speculate that LDA is attributed to this phenomenon, that is, a few words in one paragraph bring difficulties of capturing enough semantic information for topic representation. Specifically, three paragraphs are suitable for CNN/Daily Mail dataset, while it can be tuned according to materials with different length.

Topic Similarity

Except for text summarization, we also apply VHTM to extract and combine topics, which are excellent carriers of semantic information contained in documents. Intuitively, an excellent summarization model can capture abundant topics from documents and preserve the most essential information. Therefore, we evaluate topic similarity based on the topic number and the similarity of the topic distributions.

We construct a pre-trained LDA model on the CNN/Daily Mail corpus for calculating the topic distribution of articles and summaries. Further, topic similarity can be measured based on the KL divergence between two topic distributions. Blue boxes in Figure 3 illustrates that VHTM achieves a high topic similarity with reference. Besides, LDA model can predict the topic probability for each text (articles and summaries), and we count the number of the related topics whose probabilities are greater than the threshold of 0.001. Same topic index between articles and summaries are counted and reported as mapping topic number. The first yellow box in Figure 3 shows that high-quality summary preserve the topic information consisted in original document even they are more abstractive, while others demonstrates that VHTM keeps the topic information effectively compared with other significant baselines.

Para Num	R-1	R-2	R-L
2	37.87	15.78	34.67
3	38.01	16.22	34.99
4	37.57	15.69	34.39
5	37.19	15.34	33.91

Table 5: Effects of different paragraph numbers of VHTM on the CNN/Daily Mail dataset.

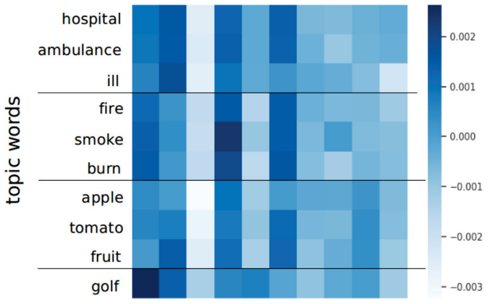

Figure 4: Topic distribution visualization of some extracted words which are consisted of three different topic groups and a random one.

Case Study of Topic Words

To further explore the topic information captured by the VHTM, and illustrate the effect of different topic words, we pick up some of them and depict their topic distribution in the Figure 4. They are categorized into three different groups, which words in one group share related information intuitively, and a random one. For limitation of space, we choose 10 out of 50 topics to depict. We observe that grouped topic words presents the similar topic distribution, comparing the each line in heatmap of Figure 4. Besides, different topic exhibits preference to some special group words discovered in each column. In a nutshell, VHTM is capable of capturing topic information efficiently.

Conclusion

In this work, we propose a variational hierarchical topic-aware mechanism for summarizing long texts, dubbed VHTM. This model simultaneously extracts and summarizes topic information by using a variational encoder-decoder framework as well as combines the hierarchical topic contents via embedding and paragraph attention. The comprehensive experiments demonstrate that VHTM holistically examines the topic substances in documents and shows promising summarizing performance.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No.U1636116, 11431006, the Research Fund for International Young Scientists under Grant No. 61750110530, and the Ministry of education of Humanities and Social Science project under grant 16YJC790123.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Bahuleyan, H. 2018. Natural language generation with neural variational models. *arXiv preprint arXiv:1808.09012*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Cao, K., and Clark, S. 2017. Latent variable dialogue models and their diversity. In *EACL*, 182–187.

Cao, Z.; Li, W.; Li, S.; and Wei, F. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *ACL*, 152–161.

Celikyilmaz, A.; Bosselut, A.; He, X.; and Choi, Y. 2018. Deep communicating agents for abstractive summarization. In *NAACL*, 1662–1675.

Chen, Y.-C., and Bansal, M. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *ACL*, 675–686.

Conroy, J. M., and O'leary, D. P. 2001. Text summarization via hidden markov models. In *SIGIR*, 406–407.

Dieng, A. B.; Wang, C.; Gao, J.; and Paisley, J. 2017. Topicrnn: A recurrent neural network with long-range semantic dependency. In *ICLR*.

Fan, A.; Grangier, D.; and Auli, M. 2017. Controllable abstractive summarization. In *ACL Workshop*.

Gambhir, M., and Gupta, V. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47(1):1–66.

Gu, J.; Lu, Z.; Li, H.; and Li, V. O. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, 1631–1640.

Gulcehre, C.; Ahn, S.; Nallapati, R.; Zhou, B.; and Bengio, Y. 2016. Pointing the unknown words. In *ACL*, 140–149.

Guo, H.; Pasunuru, R.; and Bansal, M. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *EMNLP*, 687–697.

Harabagiu, S., and Lacatusu, F. 2005. Topic themes for multi-document summarization. In *ACM SIGIR*.

Jadhav, A., and Rajan, V. 2018. Extractive summarization with swap-net: Sentences and words from alternating pointer networks. In *ACL*, 142–151.

Jiang, Y., and Bansal, M. 2018. Closed-book training to improve summarization encoder memory. In *EMNLP*, 4067–4077.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, P.; Lam, W.; Bing, L.; and Wang, Z. 2017. Deep recurrent generative decoder for abstractive text summarization. In *EMNLP*, 2091–2100.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* 8.

Liu, L.; Lu, Y.; Yang, M.; Qu, Q.; Zhu, J.; and Li, H. 2018. Generative adversarial network for abstractive text summarization. In *AAAI abstract*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *ICML*, 2410–2419.

Moody, C. E. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *SIGNLL*, 280–290.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, 3075–3081.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, 1797–1807.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL*, 1747–1759.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*, 1073–1083.

Tan, J.; Wan, X.; and Xiao, J. 2017. Abstractive document summarization with a graph-based attentional neural model. In *ACL*, 1171–1181.

Wan, X., and Yang, J. 2006. Improved affinity graph based multi-document summarization. In *NAACL*, 181–184.

Wang, D.; Zhu, S.; Li, T.; and Gong, Y. 2009. Multi-document summarization using sentence-based topic models. In *ACL-IJCNLP*, 297–300.

Wang, L.; Yao, J.; Tao, Y.; Zhong, L.; Liu, W.; and Du, Q. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *IJCAI*, 4453–4460.

Zhang, B.; Xiong, D.; Su, J.; Duan, H.; and Zhang, M. 2016. Variational neural machine translation. In *EMNLP*, 521–530.