GRETEL: Graph Contrastive Topic Enhanced Language Model for Long Document Extractive Summarization

Qianqian Xie
National Centre for Text Mining,
Department of Computer Science,
The University of Manchester
qianqian.xie@manchester.ac.uk

Jimin Huang
Chancefocus AMC.
jimin@chancefocus.com

Tulika Saha and **Sophia Ananiadou**
National Centre for Text Mining, Department of Computer Science,
The University of Manchester
{tulika.saha,sophia.ananiadou}@manchester.ac.uk

Abstract

Recently, neural topic models (NTMs) have been incorporated into pre-trained language models (PLMs), to capture the global semantic information for text summarization. However, in these methods, there remain limitations in how they capture and integrate the global semantic information. In this paper, we propose a novel model, **G**raph cont**R**astiv**E** **T**opic **E**nhanced **L**anguage model (GRETEL), that incorporates the graph contrastive topic model with the pre-trained language model, to fully leverage both the global and local contextual semantics for long document extractive summarization. To better capture and incorporate the global semantic information into PLMs, the graph contrastive topic model integrates the hierarchical transformer encoder and the graph contrastive learning to fuse the semantic information from the global document context and the gold summary. To this end, GRETEL encourages the model to efficiently extract salient sentences that are topically related to the gold summary, rather than redundant sentences that cover sub-optimal topics. Experimental results[1] on both general domain and biomedical datasets demonstrate that our proposed method outperforms SOTA methods.

1 Introduction

Due to the well-known limitation of pre-trained language models (PLMs) (Devlin et al., 2019; Wang et al., 2021) that they fail to capture long-range dependencies (Beltagy et al., 2020), attempts have been proposed to integrate neural topic models (NTMs) (Cao et al., 2015; Peng et al., 2018; Xie et al., 2021) into PLMs, which have shown significant improvement in the performance of the text summarization task (Wang et al., 2020b; Cui and Hu, 2021b; Nguyen et al., 2021; Fu et al., 2020). In addition to the local contextual information captured in PLMs, NTMs can provide an approximation of the global semantics captured from document contents, i.e., latent topics, as well as their posterior topic representations. The global semantics are further used to guide the model, to generate coherent summaries which cover the most relevant topics discussed within the document, via the attention mechanism (Wang et al., 2020b; Aralikatte et al., 2021; Nguyen et al., 2021; Fu et al., 2020) or graph neural networks (GNNs) (Cui and Hu, 2021b; Cui et al., 2020).

However, there exists the **semantic gap** between latent topics as the approximation global semantics, and the true global semantics due to two major limitations for existing methods. The first limitation concerns the nature of **unsupervised topic inference** in these methods, where topics and posterior topic distributions are learned from documents in an unsupervised manner, without considering the key semantic information conveyed in the gold summary (mostly the abstract). Existing methods using document-word features, without accessing the semantic information of the gold summary, can extract sub-optimal topics with high-frequency words. However, sub-optimal topics with high-frequency words, do not necessarily cover the true global semantics that is condensed in the gold summary. This results in the wrong assignment of the document with sub-optimal topics, and con-

[1] https://github.com/xashely/GRETEL_extractive

Proceedings of the 29th International Conference on Computational Linguistics, pages 6259–6269
October 12–17, 2022.

sequently the model extract redundant sentences containing high-frequency topic words.

Another limitation is that existing methods rely on **document word features** such as Bag-of-Words (BOWs) to extract latent topics, and disregard the sequential and syntactic dependency between words. This may lead to sequentially and syntactically correlated words being allocated to different topics. One solution is to provide NTMs with contextual representations from PLMs, which is nevertheless challenging to directly apply in existing methods for text summarization. PLMs in existing methods are forced to truncate the input to a limited length owing to the complexity of language models. Thus the representations from partial content of the document cannot necessarily help NTMs to mine informative topics that cover the whole content of the document, especially for long documents. Overall, although existing methods encourage the summary with sentences that are topically similar to the document topic distribution, the summary focuses more on sentences with high-frequency words and can have low semantic similarity with the gold summary. To help understand these limitations, we make a detailed analysis based on the benchmark datasets in section 3.

In response to the above, we propose a novel **G**raph cont**R**astiv**E** **T**opic **E**nhanced **L**anguage model (GRETEL), that incorporates the graph contrastive topic model (GCTM) empowered by the semantic information of the gold summary and the global document context, with PLMs for long document extractive summarization. The first distinguishing feature of GRETEL is the employment of the hierarchical transformer encoder (HTE) to fully embed the global context of long documents, to inform topic representations of documents and sentences. The global contextual information captured in HTE but missing in the BOW feature, allows the model to learn more discriminative document and sentence topic representations, and coherent topics.

Secondly, it utilizes graph contrastive learning with the supervised information from the gold summary. It pushes close topic representations of documents and sentences that have high semantic similarity with the gold summary and pulls away otherwise. This encourages the model to capture better global semantic information: latent topics that describe the most key information in the original document content. Therefore, it allows the model to select key sentences that are topically similar to the gold summary. Experimental results demonstrate that our method can effectively distinguish salient sentences from documents with both global and local semantics, leading to superior performance compared to previous methods.

2 Related Work

Topic enhanced PLMs for Text Summarization. Many studies have investigated the application of PLMs for extractive summarization, including BERTSum (Liu and Lapata, 2019), DiscoBert(Xu et al., 2020), MatchSum (Zhong et al., 2020) et al. However, these methods fail to capture the global context of long documents due to the limitation of PLMs. To address it, several studies combined topic modeling with PLMs to introduce global semantic information. Wang et al. (2020b) proposed to extract firstly latent topics independently and then use them to improve the summarization model. Other studies (Aralikatte et al., 2021; Nguyen et al., 2021; Cui et al., 2020) proposed to use the BOW as input features for neural topic modeling and improved the transformer encoder and decoder with the extracted latent topics with an attention mechanism for abstractive summarization (Aralikatte et al., 2021; Fu et al., 2020; Nguyen et al., 2021). Cui et al. (2020); Cui and Hu (2021b) proposed to use the graph neural network to infuse topics into contextual representations from PLMs, for multi-document abstractive summarization and extractive summarization. Fu et al. (2020) considered extract both document and paragraph-level topic distribution, and use them to guide the abstractive summarization.

PLMs for Long Document Summarization. To address the limitation of encoding the full context of long documents using PLMs, another direction is to design the efficient sparse self-attention or using a sliding window. Narayan et al. (2020) proposed a step-wise model with a structured transformer. Huang et al. (2021) proposed a computationally efficient method based on the head-wise positional strides, to identify salient content for long documents. Liu et al. (2021) employed a transformer with multi-granularity sparse attentions. Cui and Hu (2021a) used a sliding selector network with dynamic memory, in which the sliding window is used to encode input documents, segment by segment. Grail et al. (2021) divided long documents into multiple blocks and encoded them by independent transformer networks.

3 Dataset-dependent Analysis for Limitations

To better illustrate the limitations of existing methods, we present a dataset-dependent analysis, with the aim to answer two key questions: 1) Do the extracted topics from topic models tend to focus on high-frequency words? and 2) Due to it, would there be a semantic gap between topics as the approximation of global semantics and true global semantics in the gold summary?

We first present the top-10 words of topics learned by the traditional topic model LDA on the PubMed (Cohan et al., 2018) dataset, as shown in Table 1. It shows that there is a high overlap between words in learned topics and high-frequency words. This is also reported in previous studies (Griffiths and Steyvers, 2002; Steyvers and Griffiths, 2007; Chi et al., 2019), that words are mentioned more frequently, have a higher probability conditioned on topics on average. Since, they infer the posterior distribution of documents over topics, according to the co-occurrences of words in the whole document collections.

T1: type treatment consistent needed lower disorders sensitive patient acid way T2: male group treatment followed cells per side plasma american health T3: al dna clinical risk observed tube lower inflammatory et features T4: type al clinical mice bacteria high vs posterior conditions side T5: differences performed results side number higher size tube et patients T6: dna revealed smoking control mental number change sd light versus
Top high-frequency words: patients, study, using, cells, group, treatment, et, one, al, data, studies, two, patient, results, cell, time, however, figure, significant, reported, high, disease, analysis, clinical, found, age, years, associated, showed, different, compared, risk, levels

Table 1: top-10 words of topics learned by LDA on Pubmed dataset.

In Table 2, we further compare the mean score of ROUGE-1 (Lin and Hovy, 2003) F1 and ROUGE-2 F1 of the oracle summary, and summary based on the generated topics among all datasets used in our experiments. It shows a much lower rouge score on all datasets for summaries using generated topics, which indicates a semantic gap between the latent topics and the gold summary. The latent topics would guide the method to select sentences that are topically similar to the posterior distribution of documents, rather than informative sentences, that cover the semantics in the gold summary.

Dataset	Oracle Summary	Summary with Topics
CNN/DM	0.811	0.174
ArXiv	0.826	0.169
PubMed-Long	0.845	0.184
PubMed-Short	0.847	0.187
CORD-19	0.861	0.188
S2ORC	0.841	0.193

Table 2: The mean score of ROUGE-1 F1 and ROUGE-2 F1 between different summaries with the gold summary, averaged on all documents.

Oracle summary: this case report illustrates three learning points about cervical fractures in ankylosing spondylitis, and it highlights the need to manage these patients with the neck initially stabilised in flexion. We describe a case of cervical pseudoarthrosis that is a rare occurrence after fracture of the cervical spine with ankylosing spondylitis. This went undetected until the development of myelopathic symptoms many months later. The neck was initially stabilised in flexion using tongs, and then slowly extended before anterior and posterior fixation was performed. (**Mean score on ROUGE-1 F1 and ROUGE-2 F1: 0.994**)
Summary based on topic words: A patient's neurological condition may be made worse by extension of the neck, as the spinal cord may be compromised by the angle that is formed between the upper and lower rigid bony segments of the cervical spine. Over the previous 5 weeks, he had been experiencing increasing, although intermittent, symptoms including: sharp pains in the posterior aspect of his neck with head movement, abdominal pain and paraesthesia with numbness of his fingers and toes. Certainly significant trauma to a rigid and osteoporotic spine will cause fracture, and then the effect of instability at the fracture site (the fused spinal segments can be thought of as a long bone) will produce a pseudoarthrosis. (**Mean score on ROUGE-1 F1 and ROUGE-2 F1 : 0.172**)
Top-topic words: cervical, spine, neck, ankylosing, fractures spondylitis, fixation, spinal, stabilised, anterior, posterior, flexion, fracture, c7, trauma, traction, paraesthesia, weakness, immobilisation, immobilised, bony, limb, head, cord, post-operatively.....

Table 3: An example document. High-frequency topic words that appeared in sentences are marked with a red color.

4 Method

To address the aforementioned limitations of existing methods, we propose our method GRETEL, to better capture and incorporate the global semantics to improve PLMs, for long document extractive summarization. Given u sentences $\{s_1, \cdots, s_u\}$ of a document i from the corpus D, extractive summarization aims to select v informative sentences from u sentences ($v \ll u$) as the summary S for the document i. This task can be formulated as a binary sentence classification problem. We assign label $y_{i,j} = 1$ to sentence $s_{i,j} (j \in \{1, \cdots, u\})$ for the summary, or $y_{i,j} = 0$ otherwise.

As shown in Figure 1, different from previous methods, we leverage the contextual representations from PLMs, and gold summary to guide the topic inference. To this end, we first employ the hierarchical transformer encoder (HTE) to fully encode the global context of long documents, and then design the supervised graph contrastive loss, to push close the document topic distribution and topic distributions of salient sentences. This helps our method to capture better global semantics, that effectively distinguish between salient and non-salient sentences, according to their contextual and semantic connections to the gold summary.

4.1 Hierarchical Transformer Encoder

To fully encode the document contents, especially for long documents, we propose to use a the Hierarchical Transformer Encoder (HTE) based on

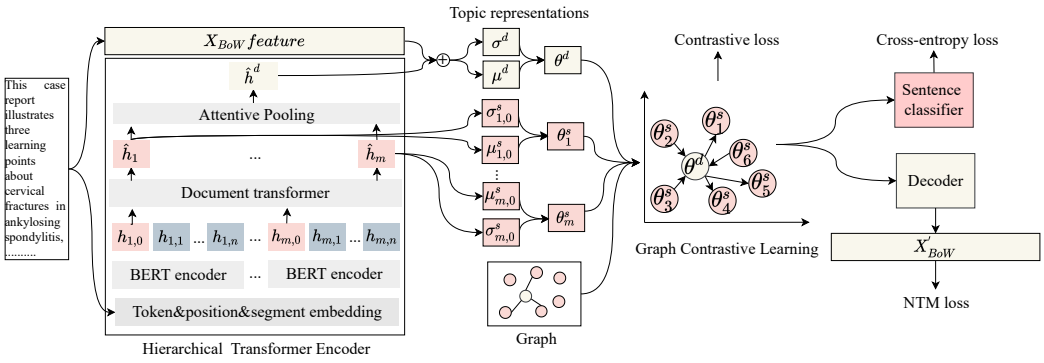

Figure 1: The model architecture of GRETEL

blocks with two modules: the block transformer encoder and the document transformer.

Block transformer encoder. We first split the document $d = \{blk_1, blk_2, \cdots, blk_m\}$ into m blocks with fixed length, in which each block $blk_l = \{x_{l,0}, \cdots, x_{l,n}\}(l \in \{1, \cdots, m\})$ has n tokens. Subsequently, each token $x_{l,p}(p \in \{1, \cdots, n\})$ at block blk_l is represented by the vector $E_{l,p}$, which is the sum of the token embedding, the block embedding, and the position embedding. Take E_l as the input embedding, the contextual representations of tokens in each block blk_l can be learned by the PLMs based transformer encoder: $h_l = BERT(E_l)$. Following previous studies (Liu and Lapata, 2019), we insert the [CLS] and [SEP] tokens at the start and end of each sentence in the block. We consider the representations of [CLS] tokens as the contextual representations of their corresponding sentences: $h^s = \{h^s_1, \cdots, h^s_u\}$, which capture the local contextual semantic in each block.

Document transformer encoder. To further model the correlations among intra-block, we stack the document transformer encoder on h^s to yield the document context-aware sentence representations: $\tilde{h}^s = Transformer(h^s)$. To denote the position of each block, we add the block position embedding (Vaswani et al., 2017) to h^s. Finally, we use the pooling layer to generate the document representation h^d based on $\tilde{h}^s$.

4.2 Graph Contrastive Topic Model

Next, we introduce the graph contrastive topic model, to capture global semantics empowered by the semantic information from HTE and the gold summary. It consists of a probabilistic topic encoder with HTE and supervised graph contrastive learning and a probabilistic decoder.

4.2.1 Probabilistic Topic Encoder Enhanced with HTE

We assume that θ^s and θ^d refer to sentence topic distributions and document topic distribution, β represents the topics (topic word distributions in the vocabulary), and X_i is the BoW feature of document i. Different from existing methods (Wang et al., 2020b; Aralikatte et al., 2021; Fu et al., 2020) considering only the BoW features, we further employ the representations from HTE to leverage the semantic and syntactic dependencies among words to generate more coherent topics and topic distributions for documents and sentences.

For document topic distribution, we sample it from the logistic normal distribution [2]. We first generate the mean and covariance of a multinomial distribution variable and then use the softmax activation function to convert it into the logistic normal distribution variable. Based on the contextual hidden representations from HTE and BoW features, for each document i, we have:

$$\begin{aligned} \tilde{h}^d_i &= f_X(X_i) + \tilde{h}^d_i \\ \mu^d_i &= f^d_\mu(\tilde{h}^d_i), \sigma^d = diag(f^d_\sigma(\tilde{h}^d_i)) \\ \theta^d_i &= softmax(\mu^d_i + (\sigma^d_i)^{\frac{1}{2}} \epsilon^d_i) \end{aligned} \tag{1}$$

where $\epsilon^d_i \in N(0, I)$ is the sampled noise variable, f^d_μ and f^d_σ are the feed-forward neural networks which takes input as the BoW feature X_i and the contextual hidden representations $\tilde{h}^d_i$ of document i from HTE respectively.

The sentence topic distribution is sampled with only the document context-aware hidden represen-

[2] Following the previous method (Srivastava and Sutton, 2017), we use the logistic normal distribution to approximate the Dirichlet distribution.

tations of sentences from HTE:

$$\mu^s_{i,j} = f^s_\mu(\tilde{h}^s_{i,j}), \sigma^s_{i,j} = diag(f^s_\sigma(\tilde{h}^s_{i,j}))$$
$$\theta^s_{i,j} = \mu^s_{i,j} + (\sigma^s_{i,j})^{\frac{1}{2}} \epsilon^s_{i,j} \quad (2)$$

where $\epsilon^s_{i,j} \in N(0, I)$ is the sampled noise variable, f^s_μ and f^s_σ are the feed-forward neural networks which takes the same input as the representation from HTE for the sentence j in the document i. Notice that the BoW features for sentences are not considered since they would be too sparse to introduce external noise.

4.2.2 Supervised Graph Contrastive Learning

Although the posterior topic distributions can now utilize the sequential dependencies of words from HTE, they still cannot distinguish between important and redundant topics without the semantic information from the gold summary. Thus, to fill the gap between the posterior topic distributions from NTMs and the key semantics in the gold summary, we propose the supervised graph contrastive learning to explicitly guide the document topic and sentence topic representations with the gold summary.

Graph Construction. For each document, we first build the graph $G = \{V, E\}$ with nodes V as the document and all its sentences. Edges E can be represented by the adjacency matrix A, in which the edge between two nodes (i, j) is defined as:

$$A_{i,j} = \begin{cases} 1, & \textit{i is the document node, j is the} \\ & \textit{sentence node, and } j \in S^+ \\ 1, & i = j \\ 0, & otherwise \end{cases} \quad (3)$$

where S^+ is the oracle summary of each document which has the maximum semantic similarity with the gold summary. Notice that the graph is a bipartite graph with only connections between the document and sentences.

Graph Contrastive Representation Learning. Based on the bipartite graph embedded with the supervision information, we argue that the representation of the document should be similar to representations of informative sentences in the oracle summary and dissimilar to representations of redundant sentences that are not mentioned. Therefore, we design the following loss for the graph contrastive representation learning:

$$\mathcal{L}_{con} = -\frac{1}{|V|} \sum_{i=1}^{|V|} log(\frac{\sum_{0<A_{i,j}} -A_{i,j} cos(x_i, x_j)}{\sum_{A_{i,j}=0} cos(x_i, x_j)}) \quad (4)$$

where $|V|$ is the number of nodes, cos denotes the cosine similarity, and $x_i \in \theta^d_i, \theta^s_i$ means the features of node i. This loss explicitly pushes close the topic distributions of the nodes with connections in the graph, i.e., the document node and the sentence node in the oracle summary, and pulls away otherwise. It guides the model to learn more discriminative document and sentence distributions that are semantically related to the gold summary.

4.3 Probabilistic Decoder

Based on sampled sentences and document representations, we use the probabilistic decoder to generate the observed words and predict the labels of sentences in each document. For each document i, we assume the v-th word $w^d_{i,v}$ is generated from the multinomial distribution based on the dot product of the document representations and topics:

$$p(w^d_{i,v}|(\theta^d_i, \beta); \Phi) = Mult([\theta^d_i \cdot \beta]) \quad (5)$$

where Φ is the parameter set of the probabilistic decoder. The topics β are randomly initialized. We assume the j-th sentence label $\tilde{y}_{i,j}$ of document i is generated from the feed-forward neural network f_y with the sigmoid activation function, based on the sentence representation $\theta^s_{i,j}$:

$$p(\tilde{y}_{i,j}|\theta^s_{i,j}; \Phi) = f_y(\theta^s_{i,j}) \quad (6)$$

Since we fill the gap between the approximation and true global semantics with supervised contrastive learning based on both the contextual representations from HTE and BoW features, our method allows us to directly use the sentence topic representations to predict the labels of sentences, without any further distillation or fusion.

4.4 Optimization

We optimize the loss function from both graph contrastive topic modeling and extractive summarization to support joint inference. The final loss of GRETEL is the sum of the evidence lower bound (ELBO) and the graph contrastive loss:

$$\mathcal{L} = \mathcal{L}(\Theta, \Phi; X_i) + \eta \mathcal{L}_{con} \quad (7)$$

where η is the parameter to control the sensitivity of the contrastive normalization, $\mathcal{L}_{con}$ is the contrastive loss, and $\mathcal{L}(\Theta, \Phi; X_i)$ is:

$$\begin{aligned} \mathcal{L}(\Theta, \Phi; X_i) = & \mathbb{E}_{q_\Theta(\theta^s_i|\tilde{h}^s_i)}[logP_\Phi(y_i|\theta^s_i)] \\ & + \mathbb{E}_{q_\Theta(\theta^d_i|X_i, \tilde{h}^d_i)}[logP_\Phi(w_i|\theta^d_i, \beta)] \\ & - D_{KL}[q_\Theta(\theta^d_i|X_i, \tilde{h}^d_i)||P_\Phi(\theta^d_i)] \end{aligned} \quad (8)$$

where y_i is the ground truth labels of sentences in document i. The ELBO is composed of three terms, including the sentence label prediction loss for the extractive summarization, the word reconstruction loss of the neural topic modeling, and the KL divergence between the variational posterior and the prior of θ^d, which uses the prior $P_\Phi(\theta^d|\alpha)$ to normalize the document topic representations.

5 Experimentation Details

In this section, we present the details of the datasets used, evaluation metrics, and different baselines.

Datasets. To evaluate the effectiveness of GRETEL, we conducted experiments on four benchmark datasets and two biomedical domain-specific long document datasets: 1) CNN/DM (Hermann et al., 2015): a commonly used news dataset; 2) Arxiv (Cohan et al., 2018): a dataset containing long scientific documents from the Arxiv website; 3) PubMed-Long (Cohan et al., 2018): a dataset containing long scientific documents from biomedicine; 4) PubMed-Short (Zhong et al., 2020): adapted PubMed-Long to use only the introduction of the document as input and filter noisy documents; 5) CORD-19 (Wang et al., 2020a): an openly released dataset including long biomedical scientific papers related to COVID-19. We use the version of the dataset which was released on 2020-06-28 (Bishop et al., 2022; Xie et al., 2022); 6) S2ORC (Lo et al., 2020): a publicly released dataset that includes long scientific papers from several domains. We sample a random subset of articles from only the biomedical domain (Bishop et al., 2022; Xie et al., 2022). We show the statistics of the datasets in Table 4. We use abstracts of documents as the gold summary. For CNN/DM and Arxiv, we extract 3 and 7 sentences respectively to formulate the final summary, following previous methods (Zhong et al., 2020). For the remaining datasets, we extract 6 sentences to formulate the summary (Bishop et al., 2022; Xie et al., 2022).

Dataset	Train	Valid	Test	Avg len	Ext
CNN/DM	287,226	13,368	11,490	757	3
ArXiv	203,037	6,436	6,440	5,038	7
PubMed-Long	119,924	6,633	6,658	3,235	6
PubMed-Short	83,233	4,676	5,025	444	6
CORD-19	31,505	6,299	4,200	3,324	6
S2ORC	47,782	9,556	6,371	2,631	6

Table 4: Statistics of datasets. Ext denotes the number of sentences extracted in the final summary.

5.1 Implementation Details

Our method is implemented by Pytorch and Huggingface (Wolf et al., 2020). We investigated the RoBERTa (Liu et al., 2019) implemented in Huggingface as the encoder. We use the base size of it. We set the learning rate to 2e-3, dropout rate to 0.0, warmup steps to 5000, topic number between$\{100, 200, 300, 400, 500\}$, the parameter to control the negative samples of the contrastive loss γ to 1, and the weight parameter η to 0.5. We set the hidden size of the transformer in HTE to 768. Due to the memory limitations of GPU, we set the max tokens of input documents as 6000. We train the model with 50000 steps and save the model checkpoint at every 1000 steps. We select the best checkpoint according to the loss in the validation and report the results in the test. To extract the sentence label for training the model, we use the greedy search algorithm (Nallapati et al., 2017) to select the oracle summary of each document, via maximizing the ROUGE-2 score against the gold summary. we use the pyrouge[3] to calculate the ROUGE (Lin, 2004) metric.

Baselines and Metrics. We compare our model with SOTA extractive summarization methods including: 1) PLMs based methods: BERTSum (Liu and Lapata, 2019) and MatchSum (Zhong et al., 2020); 2) PLMs based models for long documents: HIBERT (Zhang et al., 2019), ETCSum (Narayan et al., 2020), SSN-DM (Cui and Hu, 2021a), GBT-EXTSUM (Grail et al., 2021), Longformer-Ext (Beltagy et al., 2020), Reformer-Ext (Kitaev et al., 2019), BERTSum+SW (Liu and Lapata, 2019) which uses the BERTSum to sequentially encode the full context with the sliding window; 3) topic enhanced transformer method: TopicGraphSum (Cui and Hu, 2021b), which is the only PLMs-based model with topic modeling for extractive summarization. Following (Liu and Lapata, 2019), we report the unigram (ROUGE-1), bigram F1 (ROUGE-2), and the longest common subsequence (ROUGE-L) between the generated summary and the gold summary.

6 Results and Analysis

A series of experiments were conducted to demonstrate the efficacy of the proposed method.

[3]https://github.com/andersjo/pyrouge.git

Datasets	CNN/DM			Arxiv			PubMed-Long			PubMed-Short			CORD-19			S2ORC		
Metrics	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
LEAD	40.11	17.54	36.32	33.66	8.94	22.19	36.19	11.82	32.96	37.58	12.22	33.44	32.40	8.97	29.30	36.62	16.57	33.11
ORACLE	56.22	33.74	52.19	53.88	23.05	44.90	50.26	28.32	46.33	45.12	20.33	40.19	46.20	22.86	42.08	58.34	34.48	54.36
BERTSum	43.25	20.24	39.63	41.24	13.01	36.10	41.09	15.51	36.85	41.05	14.88	36.57	36.25	10.83	32.85	40.53	16.31	37.50
MatchSum	44.41	20.86	40.55	-	-	-	-	-	-	41.21	14.91	36.75	-	-	-	-	-	-
Topic-GraphSum	44.02	20.81	40.55	44.03	18.52	32.41	45.95	20.81	33.97	-	-	-	-	-	-	-	-	-
HiBERT	42.37	19.95	38.83	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ETCSum	43.84	20.80	39.77	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Longformer-Ext	43.00	20.20	39.30	45.24	16.88	40.03	43.75	17.37	39.71	42.03	16.08	38.01	43.61	16.27	39.39	47.73	22.67	44.03
Reformer-Ext	38.85	16.46	35.16	43.26	14.68	38.10	42.32	15.91	38.26	41.67	15.78	37.88	42.32	16.11	38.87	46.12	21.55	43.21
HETFORMER	44.55	20.82	40.37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SSN-DM	-	-	-	45.03	19.03	32.58	46.73	21.00	34.10	-	-	-	-	-	-	-	-	-
BERTSum+SW	43.78	20.65	39.67	47.86	19.17	42.50	46.36	19.67	42.49	42.07	15.10	37.29	42.51	15.72	38.58	46.21	19.73	43.01
GBT-EXTSUM	42.93	19.81	39.20	48.08	19.21	42.68	46.87	20.19	42.68	-	-	-	-	-	-	-	-	-
GRETEL	**44.62**†	**20.96**†	**40.69**†	**48.17**†	**20.31**†	**42.84**†	**48.20**†	**21.20**†	**43.16**†	**42.53**†	**16.55**†	**38.61**†	**43.91**†	**16.54**†	**40.01**†	**48.24**†	**23.34**†	**44.55**†

Table 5: ROUGE F1 results of different models on CNN/DM, Arxiv, PubMed-Long, PubMed-Short, CORD-19, and S2ORC under 5 times running. † means outperform the existing model with best performance significantly ($p < 0.05$). Part results are from (Grail et al., 2021; Zhong et al., 2020; Cui and Hu, 2021a).

6.1 Main Results

We first present the ROUGE F1 results of different models on all datasets in Table 5, which shows that our method GRETEL outperforms all existing baseline methods in all datasets. It demonstrates the superiority of our method GRETEL to other methods, via capturing better global semantics with the guidance of the gold summary and the leverage of contextual information and word features simultaneously. Our methods and Topic-GraphSum both present superior performance over methods without the topic information, such as BERTSum and MatchSum, indicating the importance of modeling the global semantic information with the approximation of latent topics. When comparing with Topic-GraphSum incorporating latent topics, our method yields better performance on all datasets. This proves the benefit of the supervision from the gold summary and the integration of contextual representations to exploit the better global semantics. It is also proved from the improvement of our method when compared with methods that encode full document contents but ignore the topic information, such as Longformer-Ext, SSN-DM et al.

Moreover, our methods and other PLMs-based methods that address the truncation issue to encode full contents, such as Longformer-Ext, SSN-DM et al, achieve better performance on all long document datasets, when comparing methods with the input length limit, such as BERTSum, and MatchSum. It shows that the content loss can inhibit their ability to model the contextual information in the document, which also limits the employment of the contextual representations from existing methods in the topic generation. On the contrary, for CNN/DM and PubMed-Short whose documents are relatively short, the improvement is insignificant, since truncating these short documents would not miss much context information.

Datasets	PubMed-Long			Arxiv		
Metrics	R-1	R-2	R-L	R-1	R-2	R-L
GRETEL	**48.20**†	**21.20**†	**43.16**†	**48.17**†	**20.31**†	**42.84**†
W/O HTE	45.97	20.13	40.22	45.44	16.53	40.15
W/O Topic	47.61	20.89	42.86	47.48	19.97	42.65
W/O Contras	47.65	20.96	42.92	47.95	20.02	42.76
W/O Context	48.01	20.99	43.08	47.89	20.13	42.77
W/O Document	47.61	21.02	43.09	48.11	20.18	42.80

Table 6: ROUGE F1 results of our model under different settings on PubMed-Long and Arxiv.

6.2 Ablation Study

We further verify the attribution of each component to the performance improvement of GRETEL in this section, as shown in Table 6. It presents the results of our method including 1) W/O HTE replacing HTE with RoBERTa, 2) W/O Topic removing the loss of neural topic modeling, 3) W/O Contras removing the graph contrastive loss, 3) W/O Context without considering the contextual representations from PLMs in generating document topic representations, and 4) W/O Document without the document transformer layer to propagate information between blocks. We can observe that each component contributes to the performance of the model to a different degree. Among all the components, HTE is the most important one for improvement, which shows the importance of encoder the full contents, when introducing contextual information into the topic modeling. However, our method still outperforms Topic-GraphSum even without HTE, which demonstrates the superiority of the guidance from the gold summary during the topic generation in our method.

Furthermore, we show the impacts of different topic numbers on the performance of GRETEL in Table 7 on both the long document dataset PubMed-Long dataset and the short document

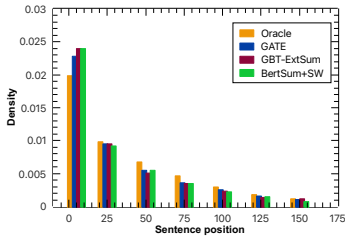

Figure 2: The position distribution of extracted sentences by different models on the PubMed-Long test set.

dataset CORD-19. It shows that the performance of our method generally increases with the growing number of topics on PubMed-Long, while it soon achieves the best of 300 topics on CORD-19 since it contains fewer topics with relatively shorter content and much fewer documents.

Datasets	**PubMed-Long**			**CORD-19**		
Metrics	R-1	R-2	R-L	R-1	R-2	R-L
K=100	47.10	20.14	42.19	43.53	15.92	38.89
K=200	47.53	20.38	42.42	43.80	16.25	39.41
K=300	47.94	20.89	42.94	**43.91**	**16.54**	**40.01**
K=400	48.15	21.12	43.24	43.85	16.42	39.94
K=500	**48.20**	**21.20**	**43.26**	43.86	16.45	40.01

Table 7: ROUGE F1 results of our model with a different number of topics on PubMed-Long and CORD-19.

6.3 Topic Analysis

To verify the quality of our generated topics, we further evaluate the NPMI (Lau et al., 2014) score in Table 8 by our methods and the classical topic model LDA. It clearly shows that our method can learn more coherent topics compared with LDA. Moreover, our method without contextual information (W/O Contextual) and the supervision (W/O Contras) all outperform LDA and underperforms our method with both features. It demonstrates that both the contextual information and the supervision from the gold summary are helpful to exploit meaningful and salient topics for a better approximation of global semantics.

Datasets	**CORD-19**		**PubMed-Long**	
Model	K=100	K=200	K=100	K=200
LDA	0.18	0.16	0.14	0.19
GRETEL	**0.25**	**0.21**	**0.23**	**0.26**
W/O Contras	0.23	0.20	0.22	0.24
W/O Contextual	0.20	0.18	0.18	0.20

Table 8: NPMI score of different models on CORD-19 and PubMed-Long using different numbers of topics.

Gold	A 53-year-old man with steroid dependent rheumatoid arthritis presented with fever and serious articular drainage. Oral antibiotics were initially prescribed. Subsequent hemodynamic instability was attributed to septic shock. Further evaluation revealed a pericardial effusion with tamponade. Pericardiocentesis of purulent fluid promptly corrected the hypotension. Proteus mirabilis was later isolated from both the infected joint and the pericardial fluid. This is the first report of combined proteus mirabilis septic arthritis and purulent pericarditis. It documents the potential for atypical transmission of gram-negative pathogens, to the pericardium, in patients with a high likelihood of preexisting pericardial disease. In immunocompromised patients, the typical signs and symptoms of pericarditis may be absent, and the clinical presentation of pericardial tamponade may be misinterpreted as one of septic shock. This case underscores the value of a careful physical examination and proper interpretation of ancillary studies. It further illustrates the importance of initial antibiotic selection and the need for definitive treatment of septic arthritis in immunocompromised patients.
Our	ID 3: We report a case of purulent pericarditis with pericardial tamponade masquerading as septic shock related to proteus mirabilis septic arthritis. ID 4: A 53-year-old man with long-standing, steroid-dependent rheumatoid arthritis complained of a painful, swollen, left elbow with purulent drainage emanating from what appeared to be a small ulceration. ID 0: Septic arthritis is a well recognized occurrence in patients with steroid dependent rheumatoid arthritis .1 treatment includes broad-spectrum antibiotics usually accompanied by surgical or needle drainage of the joint .2 while pericardial effusions are common in patients with rheumatologic disorders, the development of purulent pericarditis with pericardial tamponade is rare. ID 75: This case also underscores the importance of appropriate antibiotic selection in the initial treatment of immunocompromised patients with infected prosthetic joints. ID 30: While multiple blood cultures were negative, articular and pericardial fluid cultures grew staphylococcus epidermidis and proteus mirabilis. ID 68: In the setting of an infected joint prosthesis, fever, and immunosuppression, this patients hemodynamic instability was initially ascribed to septic shock and not to pericardial tamponade.
Baseline	ID 3: we report a case of purulent pericarditis with pericardial tamponade masquerading as septic shock related to proteus mirabilis septic arthritis. ID 4: A 53-year-old man with long-standing, steroid-dependent rheumatoid arthritis complained of a painful, swollen, left elbow with purulent drainage emanating from what appeared to be a small ulceration. ID 12: On the first postoperative day, he was transferred to the medical service for the management of presumed septic shock. ID 25: A subxiphoid pericardiocentesis yielded 500 ml of purulent fluid with prompt normalization of the blood pressure. ID 0: While multiple blood cultures were negative, articular and pericardial fluid cultures grew staphylococcus epidermidis and proteus mirabilis. ID 11: The early postoperative course, however, was remarkable for persistent fever, hypotension, and tachycardia.
Topics	T267: infected congenital patients loss blood disorders compared high fig phase T433: treatment infected severity year patients crp tract area arm isolated T446: joints dna physical risk observed tube lower examination intravenous features T153: disease type exercise vegf nerve deaths shock joints drugs lower T108: type treatment male sd well use statistically specific post mice

Table 9: Example of extractive summarization conducted by our method on the PubMed-Long dataset. The gold summary is the abstract of the document. Sentences with deep color have a higher ROUGE score. Topic words are marked with the blue color.

Sentence ID	Top-6 words
ID 3	adequate mice patients label understanding body infected report oxygen formation disease type exercise vegf nerve deaths shock joints drugs lower
ID 4	family report presented followed obesity side j macrophages high necrosis treatment infected severity year patients crp tract area arm isolated
ID 0	treatment infected severity year patients crp tract area arm isolated type treatment male sd well use statistically specific post mice
ID 75	treatment infected severity year patients crp tract area arm isolated treatment adjacent medication different motor min height stroke like rate
ID 30	et treatment lower diagnosis control observed could 7 number association infected congenital patients loss blood disorders compared high fig phase
ID 68	type treatment male sd well use statistically specific post mice effects performed revealed compared clinical observed diagnosis isolated cm family

Table 10: Top 10 words of top 2 topics in sentences, which are selected into the summary.

6.4 Case Study

In Table 9, we present the summaries of an example document generated by GRETEL and BERTSum, together with the top-5 topics (with the highest coherence) of the document from the PubMed-Long dataset. In table 10, we show the top-2 topics of selected sentences by our method for inclusion in the summary. It shows that our method generates a more coherent summary that contains more salient sentences than the summary generated by BERTSum, due to the integration of a better approximation of global semantics in our method. This is also proved by the selected sentences of our method are topically related to the captured topics about "treatment", "joints" and "infected", which are semantically similar to the meaning of the gold summary.

Moreover, the positions of our selected sentence vary in every part of the document while the sentences of BERTSum are all located in the former part of the document. This is because the employment of HTE allows our method to encode the full contents of the document without truncation. In Figure 2, we further compare the position distribution of selected sentences by different models and the oracle summary on PubMed-Long. The distribution of our method is the most similar to the oracle summary, which pays more attention to the latter sentences compared with other models.

7 Conclusion

In this paper, we propose a novel framework GRETEL for extractive summarization of long texts, that furnishes PLMs with the neural topic inference, to fully incorporate the local and global semantics. Experimental results on both general and biomedical datasets show that our model outperforms existing state-of-the-art methods, and global semantics empowered by graph contrastive learning and PLMs can yield more discriminative sentence representations to select salient sentences, that are topically similar to the gold summary. For future work, we would explore the feasibility of extending this framework to abstractive and multi-document summarization tasks.

Acknowledgments

This research is supported by the Alan Turing Institute and the Biotechnology and Biological Sciences Research Council (BBSRC), BB/P025684/1. We would like to thank Pan Du, Jennifer Bishop, and Guanghao Yang for their help and constructive comments.

References

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. *arXiv preprint arXiv:2105.11921*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. Gencomparesum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 220–240.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Jinjin Chi, Jihong Ouyang, Changchun Li, Xueyang Dong, Ximing Li, and Xinhua Wang. 2019. Topic representation: Finding more representative words in topic models. *Pattern recognition letters*, 123:53–60.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

Peng Cui and Le Hu. 2021a. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5881–5891.

Peng Cui and Le Hu. 2021b. Topic-guided abstractive multi-document summarization. *arXiv preprint arXiv:2110.11207*.

Peng Cui, Le Hu, and Yuanchao Liu. 2020. Enhancing extractive text summarization with topic-aware graph neural networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5360–5371.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Xiyan Fu, Jun Wang, Jinghan Zhang, Jinmao Wei, and Zhenglu Yang. 2020. Document summarization with vhtm: Variational hierarchical topic-aware mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7740–7747.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing bert-based transformer architectures for long document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1792–1810.

Thomas L Griffiths and Mark Steyvers. 2002. A probabilistic approach to semantic representation. In *Proceedings of the annual meeting of the cognitive science society*, volume 24.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Ye Liu, Jian-Guo Zhang, Yao Wan, Congying Xia, Lifang He, and Philip S Yu. 2021. Hetformer: Heterogeneous transformer with sparse attention for long-text extractive summarization. *arXiv preprint arXiv:2110.06388*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shashi Narayan, Joshua Maynez, Jakub Adamek, Daniele Pighin, Blaz Bratanic, and Ryan McDonald. 2020. Stepwise extractive summarization and planning with structured transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4143–4159.

Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.

Min Peng, Qianqian Xie, Yanchun Zhang, Hua Wang, Xiuzhen Jenny Zhang, Jimin Huang, and Gang Tian. 2018. Neural sparse topical coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2332–2340.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations*.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020a. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020b. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, page 109460.

Qianqian Xie, Jimin Huang, Pan Du, Min Peng, and Jian-Yun Nie. 2021. Graph topic neural network for document representation. In *Proceedings of the Web Conference 2021*, pages 3055–3065.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.