Extractive Text Summarization with Latent Topics using Heterogeneous Graph Neural Network

Anonymous ACL submission

Abstract

This paper presents a heterogeneous graph neural network (HeterGNN) model for extractive text summarization (ETS) by using latent topics to capture the important content of input documents. Specifically, topical information has been widely used as global information for sentence selection. However, most of the recent approaches use neural models, which lead the training models more complex and difficult for extensibility. In this regard, this study presents a novel graph-based ETS by adding a new node of latent topics into HeterGN for the summarization (TopicHeterGraphSum). Specifically, TopicHeterGraphSum includes three types of semantic nodes (i.e., topic-word-sentence) in order to enrich the cross-sentence relations. Furthermore, an extended version of TopicHeterGraphSum for multi documents extraction is also taken into account to emphasize the advantage of the proposed method. Experiments on benchmark datasets such as CNN/DailyMail and Multi-News show the promising results of our method compared with state-of-the-art models.

1 Introduction

ETS is an important task of Natural Language Processing (NLP) in terms of extracting several relevant sentences from the original documents while keeping main information. The traditional methods for ETS are TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004), which focus on calculating the similarity between sentences scores. Sequentially, the rapid development of Deep Learning (DL) has brought breakthrough records by modeling a document as a sequence of sequences in order to deal with long-range inter-sentence relationships for summarization (Cheng and Lapata, 2016; Cohan et al., 2018). However, cross-sentence relations is still a challenge of this research field (Liu and Lapata, 2019).

Recent works focus on Graph Neural Networks (GNNs) (e.g., Graph Convolutional Network (GCN) (Kipf and Welling, 2017) or Graph Attention Network (GAT) (Velickovic et al., 2018)) to explore the cross-sentence relationships for summarization task. The core idea is to represent inter-sentential graphs and using message passing to extract the complex relationship in the input documents. For instance, Yasunaga et al. (2017) and Xu et al. (2020) adopt discourse analysis to build document graphs. Jia et al. (2020) and Wang et al. (2020) built a bi-partite graph between words and sentences, which is referred as heterogeneous graph neural network. Moreover, modeling global information is also taken into account for sentence selection by using pretrained models (Liu and Lapata, 2019; Zhang et al., 2019). Sequentially, Cui et. al. Cui et al. (2020) utilized pre-trained BERT to learn contextual sentence representations and train jointly with latent topics using neural topic model (NTM). Nguyen et al. (2021) presents an extend version using NTM for abstractive text summarization indicates the capability of enriching the global information for the summarization.

Although the existing methods have provided remarkable results, there are several open research issues that need to take into account: i) the high performance mainly depends on pre-trained models for learning sentence representations, which is difficult for the extensibility, especially for low research languages; ii) the current external information (e.g, latent topics) are extracted by neural models, which requires more complex configurations of the training process. Furthermore, the model might be suffered by bias problem, especially in terms of small datasets; iii) multi document summarization is still an open research issue, which requires a comprehensive summary for covering an event and avoiding redundancy. In this regard, this study proposes a new HeterGNN model for EST problem by adding latent topic node into graph structure, in which the initialized topic features are extracted by well-known clustering methods such as K-mean and Gaussian Mixture Models (GMM). The core idea is to investigate the impact of topical information for the EDS problem in terms of both single and multiple document extraction. To the best of our knowledge, this paper is the first study to adopt topical information for multi documents summarization. More details of the proposed model is described in the following sections.

2 Background

The proposed model based on the concept of a HeterGNN model, which is proposed by Wang et al. (2020), for enriching the relationships between sentences by adding nodes with semantic features. Particularly, the model includes three main components such as initialized graph structure, graph layer, and sentence selection module. Graph structure is initialized by the set of word node, which is encoded using Glove (Pennington et al., 2014) as the addition node, and sentence features, which are calculated by combining CNN for extracting the local n-gram feature of each sentence and bidirectional Long Short-Term Memory (BiLSTM) for extracting the sentence-level feature, respectively. In this regard, the feature of the sentence s_j can be obtained as follows:

$$X_{s_j} = CNN(x_{1:p}) \oplus BiLSTM(x_{1:p}) \quad (1)$$

where p denotes number of word in the sentence. Moreover, TFIDF is adopted for further approval information of the relationships between word and sentence. Sequentially, the graph layer is updated using GAT (Velickovic et al., 2018), with a modification for heterogeneous graph. Specifically, the updated node representation with modified GAT can be formulated as follows:

$$z_{ij} = LeakyReLU(W_a[W_q h_i; W_e h_j; \overline{e}_{ij}]) \quad (2)$$

where $\overline{e}_{ij}$ denotes the multi-dimensional embedding space ($\overline{e}_{ij} \in \mathbb{R}^{d_e}$), which is mapped from edge weight e_{ij}. Thereby, the sentences with their neighbor word nodes are updated via modified-GAT and Position-Wise Feed-Forward (FFN) layer, which can be sequentially formulated as follows:

$$\begin{aligned} U^1_{s\leftarrow w} &= GAT(H^0_s, H^0_w, H^0_w) \\ H^1_s &= FFN(U^1_{s\leftarrow w} + H^0_s) \end{aligned} \quad (3)$$

where H^0_w and H^0_s are the node features of word X_w ($X_w \in \mathbb{R}^{m\times d_w}$) and sentences X_s ($X_s \in \mathbb{R}^{n\times d_s}$), respectively. Therefore, the new representations of word node can be obtained using the updated sentence nodes and further updated sentences or query nodes, iteratively. Each iteration contains a sentence-to-word and a word-to-sentence update process, which can be demonstrated as follows:

$$\begin{aligned} U^{t+1}_{w\leftarrow s} &= GAT(H^t_w, H^t_s, H^t_s) \\ H^{t+1}_w &= FFN(U^{t+1}_{w\leftarrow s} + H^t_w) \\ U^{t+1}_{s\leftarrow w} &= GAT(H^t_s, H^{t+1}_w, H^{t+1}_w) \\ H^{t+1}_s &= FFN(U^{t+1}_{s\leftarrow w} + H^t_s) \end{aligned} \quad (4)$$

The output of the new sentence representation is input into a sentence classier, which use cross-entropy loss, for ranking the classification.

3 Methodology

In this study, our model is proposed for single document summarization (SDS), however, it can be extend for multi documents with minor modifications. The methods for two aforementioned problems are described in following sections.

3.1 Single Document Summarization

Given an arbitrary document $d = \{s_1, .., s_n\}$, which includes n sentences, the objective of EDS for single document problem is to predict a set of binary label $\{y_1, .., y_n\}$ ($y_j \in [0, 1]$), which determine that the sentence in the summary or not. Figure 1 illustrates the structure of the proposed HeterGNN model. Specifically, comparing with previous works, the main idea of the proposed model is to enrich the global information. Accordingly, instead of using neural models for generating

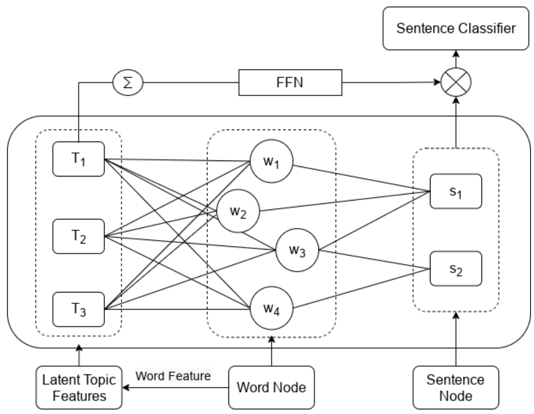

Figure 1: Overview of TopicHeterGraphSum for single document summarization. The initialized word node and sentence node features are processed following the work in Wang et al. (2020). Furthermore, we provide latent topics as addition nodes into the Hetergraph.

latent topics, we first extract the initialized topic feature of each document using simple clustering methods (e.g., K-mean and GMM) of pretrained word embeddings (Sia et al., 2020). In particular, the initialized topic feature is calculated as follows:

$$X_T = argmin \sum \begin{cases} \| c^{(i)} - x_j \|, Kmean \\ \theta_i f(x_j | c^{(i)}, \Sigma_i), GMM \end{cases} \quad (5)$$

where θ_i denotes topic proportions. $c^{(i)}$ and x_j represent the cluster center and word vector, respectively. Sequentially, the latent topics are put into graph layer for extracting semantic information. Similar to sentence representation calculation in Eq. 3, the topic representation can be updated via modified GAT as follows:

$$\begin{aligned} U^1_{t \leftarrow w} &= GAT(H^0_t, H^0_w, H^0_w) \\ H^1_t &= FFN(U^1_{t \leftarrow w} + H^0_t) \end{aligned} \quad (6)$$

Each iteration contains word-to-sentence, sentence-to-word and word-to-topic for the update process, which can be formulated as follows:

$$\begin{aligned} U^{t+1}_{w \leftarrow s} &= GAT(H^t_w, H^t_s, H^t_s) \\ U^{t+1}_{w \leftarrow T} &= GAT(H^t_w, H^t_T, H^t_T) \\ U^{t+1}_{w \leftarrow s,T} &= \sigma(U^{t+1}_{w \leftarrow s} + U^{t+1}_{w \leftarrow T}) \\ H^{t+1}_w &= FFN(U^{t+1}_{w \leftarrow s,T} + H^t_w) \\ U^{t+1}_{s \leftarrow w} &= GAT(H^t_s, H^{t+1}_w, H^{t+1}_w) \\ H^{t+1}_s &= FFN(U^{t+1}_{s \leftarrow w} + H^t_s) \\ U^{t+1}_{T \leftarrow w}, A^{t+1}_{T \leftarrow w} &= GAT(H^t_T, H^{t+1}_w, H^{t+1}_w) \\ H^{t+1}_T &= FFN(U^{t+1}_{T \leftarrow w} + H^t_T) \end{aligned} \quad (7)$$

where $A_{T \leftarrow w}$ denotes the attention matrix from word node to topic node. Subsequently, the topic representation of the input document is calculated by combining all topic features, which are learned using GAT as follows:

$$\begin{aligned} \alpha_i &= \frac{\sum_{n=1}^{N_d} c(w_n) * A_{i,n}}{\sum_{j=1}^{K} \sum_{n=1}^{N_d} c(w_n) * A_{j,n}} \\ H_{T_d} &= \sum_{i=1}^{K} \alpha_i * H_{T_i} \end{aligned} \quad (8)$$

where $A_{i,j}$ indicates the amount of information word j contribute to topic i. $c(w_n)$ is frequency of w_n in the document, K is the number of topics and α_i refers the level dominant of topic ith to total document-topic. Sequentially, each sentence hidden state is integrated with the above topic vector to capture sentence-topic representation as follows:

$$H_{s_i,T_d} = FFN(H_{T_d}) \oplus H_{s_i} \quad (9)$$

Finally, the output sentence-topic representation is used for sentences classification by using cross-entropy loss as the training objective:

$$\mathcal{L} = \sum_{i=1}^{n} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \quad (10)$$

3.2 Multi Documents Summarization

Currently, there is not much studies for multi document summarization (MDS). The main challenge of MDS is that the input documents may different in terms of main focus and point of view (Fabbri et al., 2019). Intuitively, enriching global information is able to improve the performance of MDS problem in which latent topics, extracting from word node, are considered for whole sentences in the multi documents. Therefore, in this paper, we take MDS into account by extending our proposed HeterGNN model. In particular, comparing with the original model for SDS, there are several minor modifications. Firstly, the word node and sentence node are generated by a set of relevant documents, therefore, the relationship between words and sentences are more complicated. Secondly, latent topics are generated for covering the topics of whole relevant documents. In this regard, instead of combining all topic features for the topic representation, we keep each topic feature representation separately to maintain the information. Specifically, supporting $D = \{d_1, d_2, ..., d_n\}$ denotes the set of each input multi documents, the output sentence-topic representation $s_i$is re-calculated as follows:

$$H_{s_i,T_D} = \sigma(FFN(FFN(H_{T_D}) \oplus H_{s_i})) \quad (11)$$

Sequentially, the output matrix is transformed to vector by a flatten layer for the final classification.

4 Experiment

4.1 Experimental Setting

Datasets: Two benchmark datasets are considered for the evaluation such as CNN/DailyMail (Nallapati et al., 2016) (single document dataset) and Multi-News (Fabbri et al., 2019) (multi documents dataset). For the data processing, we use the same split as the work in Wang et al. (2020).

Hyperparameter Setting: Regarding the word node generation, the vocabulary is limited to

50,000. The tokens are initialized with 100 dimensions using Glove embedding (Pennington et al., 2014). The multi head of GAT layer for word-to-sentence and word-to-topic are set to 4 and 1, respectively. The maximum number of sentences in each document is set to 100. The initialized dimensions of sentence embedding and topic embedding are set to 128 and 100, respectively. The dimension of final output representation of all models is set to 64. Regarding decoder process, we select top-3 for CNN/DailyMail and top-11 sentences for Multi-News following the performance of validation set. Furthermore, n-gram Blocking (Liu and Lapata, 2019) is also taken into account to improve the performance. Specifically, we vary the values of n-gram from 3 to 6 in order to determine the best results. The number of latent topics is set to 5 both single and multi documents, respectively. Our source code is available[1] for the reproducibility.

Baseline: For the SDS problem, we divide the baseline models into two approach. The first approach includes recent models that use pretrained models (e.g., BERT and RoBERTa) for sentence representation such as BERTSUM (Liu and Lapata, 2019), DISCOBERT (Xu et al., 2020), MATCHSUM (Zhong et al., 2020), Topic GraphSum (Cui et al., 2020), and HAHSum (Jia et al., 2020). The second approach is non-pretrained models such as BANDDITSUM (Dong et al., 2018), JECS (Xu and Durrett, 2019), HER (Luo et al., 2019), Topic GraphSum (non-pretrained version) (Cui et al., 2020), HSG (Wang et al., 2020), Multi GraS (Jing et al., 2021), including our models. Regarding the MDS problem, most recent state of the art methods using pretrained model are proposed for abstractive summarization (Xiao et al., 2021). Consequentially, we follow the reports in Wang et al. (2020) to present the comparison. The proposed model, TopicHeterGraphSum (THGS) is executed with two versions, by adopting two clustering algorithms for initialized latent topic feature, such as K-Mean (THGS-KMean) and GMM (THGS-GMM).

4.2 Results Analysis

Single Document Summurization: Table 1 shows the results of our evaluation on the CNN/DailyMail dataset. As results, our model outperforms the models of non-pretrained approach and are comparable with pretrained approach. Especially, comparing with the method using Neural Topic Modeling (NTM) (Cui et al., 2020), our model are better in terms of non-pretrained version (using Bi-GRU).

Model	R-1	R-2	R-L
BERTSUM	43.25	20.24	39.63
DISCOBERT	43.77	20.85	40.67
MATCHSUM	44.41	20.86	40.55
Topic-GraphSum	44.02	20.81	40.55
HAHSum	**44.68**	**21.30**	**40.75**
BANDITSUM	41.50	18.70	37.60
JECS	41.70	18.50	37.90
HER	42.30	18.90	37.90
Topic-GraphSum	41.93	19.15	38.22
HSG	42.95	19.76	39.23
Multi-GraS	43.16	20.14	39.49
THGS-Kmean (ours)	43.25	20.20	39.62
THGS-GMM (ours)	**43.28**	**20.31**	**39.67**

Table 1: Results on CNN/DailyMail dataset. Report results are obtained from respective papers. Bold font indicates the best results of pretrained-based models and non-pretrained models, separately.

Multi Document Summurization: Table 2 shows the results on the Multi-News dataset for MDS problem. Specifically, the results indicate that enriching global information by using latent topics is able to improve the performance of MDS problem.

Model	R-1	R-2	R-L
TextRank	41.95	13.86	38.07
LexRank	41.77	13.81	37.87
PG-BRNN	45.27	15.32	41.38
Hi-MAP	45.21	16.29	41.39
HDSG	46.05	16.35	42.08
THGS-Kmean (ours)	46.60	16.81	42.63
THGS-GMM (ours)	**46.66**	**16.90**	**42.73**

Table 2: Results on Multi-News dataset. Report results are obtained from Wang et al. (2020).

5 Conclusion and Future Work

We introduce a new method for EDS problem by enriching global information using latent topics. Specifically, we first generate the latent topics using well-known clustering algorithms and put into a proposed HeterGNN for learning feature representing. A major drawback of this study is that we use the same latent topic aggregation method for both SDS and MDS problems. Therefore, Further exploitation of topic aggregation is considered for our future work regarding this study.

[1] https://github.com/anonymous

References

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5360–5371. International Committee on Computational Linguistics.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3739–3748. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3622–3631. Association for Computational Linguistics.

Baoyu Jing, Zeyu You, Tao Yang, Wei Fan, and Hanghang Tong. 2021. Multiplex graph neural network for extractive text summarization. *CoRR*, abs/2108.12870.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. Association for Computational Linguistics.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: human reading inspired extractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3031–3041. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *CoRR*, abs/2109.10616.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1728–1736. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of* *the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6209–6219. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. PRIMER: pyramid-based masked sentence pre-training for multi-document summarization. *CoRR*, abs/2110.08499.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3290–3301. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 5021–5031. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir R. Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 452–462. Association for Computational Linguistics.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6197–6208. Association for Computational Linguistics.