Improving Abstractive Dialogue Summarization with Hierarchical Pretraining and Topic Segment

MengNan Qi*, Hao Liu, YuZhuo Fu*, Ting Liu
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University,Shanghai,China
{qmn1998,liuh236,yzfu,louisa_liu}@sjtu.edu.cn

Abstract

With the increasing abundance of meeting transcripts, meeting summary has attracted more and more attention from researchers. The unsupervised pre-training method based on transformer structure combined with fine-tuning of downstream tasks has achieved great success in the field of text summarization. However, the semantic structure and style of meeting transcripts are quite different from that of articles. In this work, we propose a hierarchical transformer encoder-decoder network with multi-task pre-training. Specifically, we mask key sentences at the word-level encoder and generate them at the decoder. Besides, we randomly mask some of the role alignments in the input text and force the model to recover the original role tags to complete the alignments. In addition, we introduce a topic segmentation mechanism to further improve the quality of the generated summaries. The experimental results show that our model is superior to the previous methods in meeting summary datasets AMI and ICSI.

1 Introduction

Meeting is a common activity for people to discuss, exchange views and obtain information around specific topics. The widespread application of speech transcription technology has brought about the rapid expansion of meeting corpus. Therefore, automatic meeting summaries are valuable to people and society by providing quick access to important content of the information.

The recently successful sequence-to-sequence (Sutskever et al., 2014) based architecture has greatly inspired the existing meeting summary methods. Specifically, earlier studies use RNNs (Chung et al., 2014) structures such as LSTM (Hochreiter and Schmidhuber, 1997) to capture the local composition of documents and learn the semantic representation of documents. Unfortunately, RNNs lack global modeling capability and is difficult to deal with long-term dependency. To overcome this limitation, more and more researchers introduce convolution or transformer (Vaswani et al., 2017) model. These methods are easy to modify to capture more global information. However, recent studies indicate that they may not be sufficient to build long-term dependency models, which makes them significantly less effective in the context of long-term multi-human dialogue. Therefore, constructing hierarchical encoding structure (Li et al., 2015) to capture the content information of each speaker and the high-level semantic information hidden among utterances has become the mainstream method in the field of meeting summary.

Different from news texts, utterances are often turned from different interlocutors, which leads to the topic drifts, and lower information density. These problems need to be overcome by introducing external high-level semantic information, such as conversation behavior, topic mining and so on. (Goo and Chen, 2018) proposed to use the dialogue act signals in a neural summarization model. (Li et al., 2019) introduced Visual Focus Of Attention (VFOA),which represents the common concerns of all conference participants in each time stamp, to keep the keep meeting summaries on topic. (Zhao et al., 2020) improved abstractive dialogue summarization with Graph Structures and Topic Words. These studies have proved that the introduction of external high-level semantic information has positive feedback on the results of meeting summary.

Meanwhile, the use of carefully designed unsupervised pre-training tasks and large scale pretraining corpus has achieved great success in the field of document summary and dialogue understanding. BART (Lewis et al., 2019) corrupted text with an arbitrary noising function and learned to reconstruct the original text. Pegasus (Zhang et al., 2020) masks the key sentences in the original text and requires the model to generate those

*Corresponding authors: YuZhuo Fu, MengNan Qi

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1121–1130
November 7–11, 2021.

designated sentences at the decoder. Besides, DialogBERT (Gu et al., 2020) applied Next Utterance Generation, Masked Utterance Regression and Distributed Utterance Order Ranking tasks to capture discourse-level coherence among utterances. Combined with different downstream tasks, only using a small number of labeled training sets in the field for supervised fine tuning can get quite good results. Unlike the context of many people's short conversations or texts in the form of news or papers, each participant's speech contains not only a complete fragment of their own views, but also a common discussion and exchange of views with other speakers. We think that the meeting summary task can combine the document summary and dialogue understanding to get a better result.

Therefore, we propose a hierarchical transformer encoder-decoder network with auxiliary multi-task learning. We mainly follow the model structure of HMNet (Zhu et al., 2020) and construct new pre-training tasks on different levels of encoder. Our contributions are as follows:

(1) In word-level encoder, we construct the GSG pre-training task proposed by Pegasus, and extract the key sentences for every utterance, and then generate them in the decoder. The difference is that we improve the meeting summary results by using TextRank (Mihalcea and Tarau, 2004) and MMR (Carbonell and Goldstein, 1998) algorithm to extract key sentences from the original text.

(2) In utterance-level encoder, our model integrates role representations into the underlying layers of the semantic module based on the alignments between text and roles. Besides, for the better fusion of textual and role features, we design a new pre-training objective by randomly mask some of the role alignments in the input text and asking the model to recover the original role tags to complete the alignments. Unlike the existing pre-trained language representation models only utilizing local context to predict tokens, our objectives require models to aggregate both context and role tags for predicting roles.

(3) We also introduce the topic segmentation information for assisting model to generate better summaries. Specifically, We add a topic segmentation embedding to the input of the utterance lever encoder. Besides, we limit the attention of turn level encoder to different topics, which further improves the results of meeting summary.

To evaluate our model, we employ the widely used AMI (Carletta et al., 2006) and ICSI (Janin et al., 2003) meeting corpus. Results show that our model significantly outperforms previous meeting summarization methods. We then conduct ablation studies to verify the effectiveness of different components in our model.

2 Related Work

Meeting Summarization. The early works of meeting summary often focused on the use of unsupervised extraction algorithm to obtain the key information in the conversation. (Nihei et al., 2016) propose a multimodal fusion model, which combines audio, video, motion and language. The model is trained by convolutional neural network method, and can identify important words that should be included in the summary of group discussion. Furthermore, many researchers have focused on improving the abstractive meeting summarization model. (Liu et al., 2019) used the pointer generation network, which can sense the topic transfer of conversation, integrates the external topic information to improve the quality of summary generation. In the work of HMNet, a hierarchical conference summary network is proposed, which is pre-trained with news datasets, and obtained good results in AMI and ICSI.

Pre-trained Language Models. BERT (Devlin et al., 2019) introduces Masked Language Modelling and Next Sentence Prediction, which leads to the upsurge of pre-training research in NLP field. However, BERT does not perform well in the field of text generation due to the feature of auto-encoding model. The pre-training task for text generation task is designed based on MASS (Song et al., 2019) and BART. In MASS, an input sequence with a masked span of tokens is mapped to a sequence consisting of the missing tokens, while BART is trained to reconstruct the original text from corrupted input with some masked tokens. Furthermore, Pegasus build GSG task for text summary scenario, it masks the key sentences in the original text and requires the model to generate those designated sentences at the decoder. In order to make the model fully learn the high-level semantic information hidden between dialogues, (Mehri and Eskenazi, 2019) proposed a transformer based hierarchical model and various unsupervised goals for the pre-training of the context semantics of dialogue discourse. DialogBERT (Gu et al., 2020) applied Next Utterance Generation, Masked Utter-

ance Regression and Distributed Utterance Order Ranking tasks to capture discourse-level coherence among utterances.

3 Models

We mainly follow the hierarchical meeting summarization network model structure and make some improvements on the utterance-level encoder, which its fusion blocks come from ERNIE (Zhang et al., 2019). The problem of meeting summarization can be formalized as follows. The input meeting transcripts X contain some of meeting participants and the corresponding speech content. Each meeting transcript consists of multiple utterances U, where each utterance belongs to a specific topic T. The input meeting transcripts $X = \{(u_1, r_1, t_1), (u_2, r_2, t_1), \ldots, (u_m, p_n, t_l)\}$, where $u_j, 1 \leq j \leq m$ is an utterance, $t_j, 1 \leq j \leq l$ is a topic and $r_j, 1 \leq j \leq k$ is the role tag of participants. The golden summary Y written by human beings is a sequence of tokens. an utterance is made up with $w_1, w_2, \ldots, w_n$, where w_i means the word in an utterance. So the goal of the model is to generate meeting summary $Y = (y_1, \ldots, y_s)$ from the meeting transcripts $X = \{(u_1, r_1, t_1), (u_2, r_2, t_1), \ldots, (u_m, p_k, t_l)\}$.

Word-level Encoder. The word-level encoder (`W-Encoder`) is designed to extract the semantic information of a single utterance in meeting transcripts. We encode each token in one utterance using glove embeddings from spacy library. Since the parallelization mechanism of transformer can not obtain the position information of the sequence, the positional encodings are added to the input vector.There are standard transformer encoder modules on the embedded layer, which is stacked by the same block with a multi-head attention layer and a feed-forward layer. To incorporate syntactic and semantic information, we also train two embedding matrices to represent the part-of-speech (POS) and entity (ENT) tags. We directly take the output of the last hidden layer and do the average pooling operation to get the semantic representation of the turn. So we denote the output of the word-level transformer:

$$\{u_1, \ldots, u_m\} = \texttt{W-Encoder}(\{w_1^1, \ldots, w_n^1\}, \ldots, \{w_1^m, \ldots, w_n^m\}) \quad (1)$$

Utterance-level Encoder. The utterance-level encoder (`U-Encoder`) processes the word-level outputs of all utterances in a meeting and gets the high-level semantic information hidden among utterances. Each meeting participant has a different role, such as project manager and industrial designer. The speaker's information should be considered when generating the summary of the model. To be specific, we represent speaker identities with fixed-length vector called role vector. Then, both role embedding and utterance embedding are fed into utterance-level encoder for fusing heterogeneous information and computing final output embeddings. The utterance-level encoder consists of stacked fusions, which are designed for encoding both tokens and entities as well as fusing their heterogeneous features. In each fuser block, the input utterance embeddings and role embeddings from the preceding aggregator are fed into two multi-head self-attentions (MH-ATTs) respectively. Next, the fusion block adopts an information fusion layer for the mutual integration of the utterance embedding and role embedding, and computes the output embedding for each utterance and role. For an utterance vector u_j and its aligned role vector r_k, the process of information fusion is as follows:

$$h_j = \sigma(\tilde{W}_u^{(i)} \tilde{w}_j^{(i)} + \tilde{W}_r^{(i)} \tilde{r}_k^{(i)} + \tilde{b}^{(i)}) \quad (2)$$

$$u_j^{(i)} = \sigma(W_u^{(i)} h_j + b_u^{(i)}) \quad (3)$$

$$r_k^{(i)} = \sigma(W_r^{(i)} h_j + b_e^{(i)}) \quad (4)$$

where h_j is the inner hidden state integrating the information of both the utterance and the role. The final output represents two embedding of utterance semantic information and role tag information respectively:

$$\{r_1^o, \ldots, r_k^o\}, \{u_1^o, \ldots, u_m^o\} = \texttt{U-Encoder}(\{r_1, \ldots, r_k\}, \{u_1, \ldots, u_m\}) \quad (5)$$

The utterance output embedding will be input to the decoder to participate in the summary generation, and the role output embedding will be used in dRA pre-training task.

Decoder. The decoder receives the output of word-level encoder and utterance-level encoder, and generates the corresponding summary according to the semantic information of the meeting transcripts. Based on the structure of transformer decoder, the

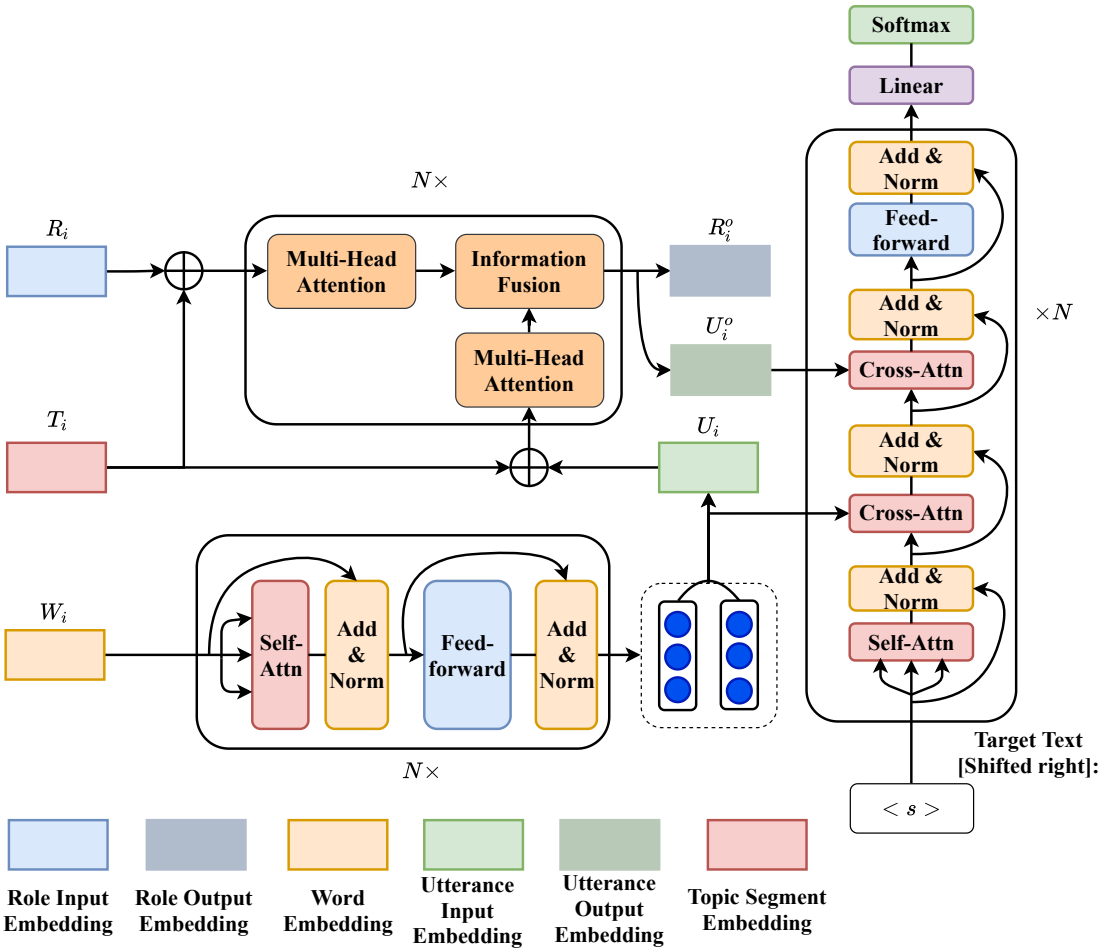

Figure 1: The overall structure of hierarchical encoder and decoder network. The word-level encoder structure processes the semantic information of each round of conversation. The utterance-level encoder structure fuses the information with role tags and topic segmentation information. The decoder receives the output of different levels of encoders and finally generates the summary.

transformer block includes two cross-attention layers. The decoder input embedding first passes through masked self-attention layer, and then performs the attention operation with word-level output and utterance-level output successively. This makes the model pay attention to the hierarchical semantic information in each inference step. The output of the decoder transformer is denoted as:

$$y_1^o, \ldots, y_s^o = \texttt{Decoder}(\\ u_1^o, \ldots, u_m^o, u_1, \ldots, u_n, y_1, \ldots, y_s) \quad (6)$$

We illustrate the whole model network in Fig.1.

4 Pretraining

We expect that our model can fully extract the semantic information of different levels in the hierarchical encoder-decoder structure through carefully designed pre-training tasks. The following three sections describe our tasks built on a hierarchical network.

Gap Sentences Generation (GSG). This pre-training task is proposed in Pegasus for the first time. It is based on the assumption that the model can achieve better and faster fine-tuning performance when the pre-training target is very similar to the downstream task. The principle is to mask the whole key sentences from the document and concatenate the gap-sentences into a pseudo-summary. In order to obtain the key sentences in the original text unsupervised, the researchers select top-m scored sentences according to importance. As a proxy for importance they compute ROUGE (Lin, 2004) between the sentence and the rest of the document.

Different from Pegasus, we try the graph based sorting algorithms TextRank and Maximum Margin Relevance(MMR) to get the key sentences in the original text.

TextRank regards each sentence in the text as a node V. If two sentences are similar, it is considered that there is a non-directional weighted edge between the corresponding nodes of the two sentences. According to the similarity calculation formula, the algorithm circularly calculates the similarity between any two nodes, sets a threshold to remove the edge connection between the two nodes with low similarity, constructs a node connection graph G(V,E), and then iteratively calculates the TextRank value of each node. After sorting, the sentences corresponding to the nodes with the highest TextRank value are selected as the key sentences. The iterative formula of TextRank is as follows:

$$WS(V_i) = (1-d) + d \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (7)$$

Where w_{ji} is the weight of the edge from V_i to V_j. d is the damping coefficient, representing the probability of pointing from one node to any other node in the graph, generally 0.85. $In(V_i)$ and $out(V_i)$ are the node set pointing to node V_i and the node set pointing from the edge of node V_i.

At the beginning of the design, MMR is used to calculate the similarity between the query text and the searched document for ranking the documents. The algorithm formula is as follows:

$$MMR(Q,C,R) = Arg \underset{d_i in, C}{\overset{k}{max}} [\lambda sim(Q, d_i) - (1-\lambda) \underset{d_j ik}{max}(sim(d_i, d_j))] \quad (8)$$

Where Q and C represent the whole document, R is an initial set which has been obtained based on the correlation, d_i represents a sentence in the document. The physical meaning of the first term in the formula refers to the similarity between the sentences to be extracted and the whole document, while the latter term refers to the similarity between the sentences to be extracted and the key sentences obtained. The key sentences extracted by MMR algorithm can not only express the meaning of the whole document, but also have diversity.

denoising Role Auto-encoder (dRA). In order to inject role information into utterance embedding by informative roles, we propose a new pre-training task, which randomly masks some utterance-role alignments and then requires the model to predict all corresponding role embedding based on aligned utterance embedding. As our task is similar to training a denoising auto-encoder (Bengio et al., 2013), we refer to this procedure as denoising role auto-encoder (dRA). Given the utterance sequence $\{u_1, \ldots, u_m\}$ and its corresponding role embedding $\{r_1, \ldots, r_k\}$, we define the aligned role distribution for the utterance embedding u_i as follows:

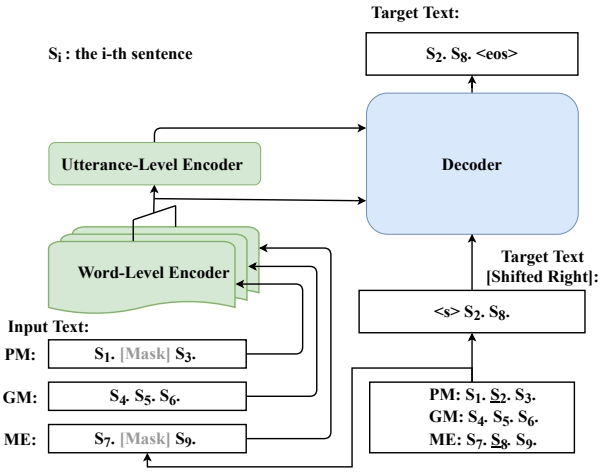

Figure 2: The procedure of Gap Sentences Generation pretraining task. Select the key sentences in the word-level encoder and replace them with [Mask]. The extracted sentences are stitched together and used as target generation text.

$$P(r_j|u_i) = \frac{exp(linear(u_i^o) \cdot r_j)}{\sum_{t=1}^{k} exp(linear(u_i^o) \cdot r_t)} \quad (9)$$

$P(r_j|u_i)$ will be used to compute the cross-entropy loss function for dRA. Figure 3 shows the implementation process of the dRA task.

Similar to BERT, we perform the following operations for dRA: (1) In 5% of the time, we replace the role embedding with another random role embedding, which aims to train our model to correct the errors that the turn is aligned with a wrong role; (2) In 15% of the time, we mask turn-role alignments and predict the role embedding; (3) In the rest of the time, we keep turn-role alignments unchanged, which aims to encourage our model to integrate the role information into turn embedding for better language understanding.

Topic Segmentation. We add a special symbol [TSEP] of topic segmentation to further improve the summary effect. Specifically, in addition to the position embedding, we also add the topic segmentation embedding in the utterance lever

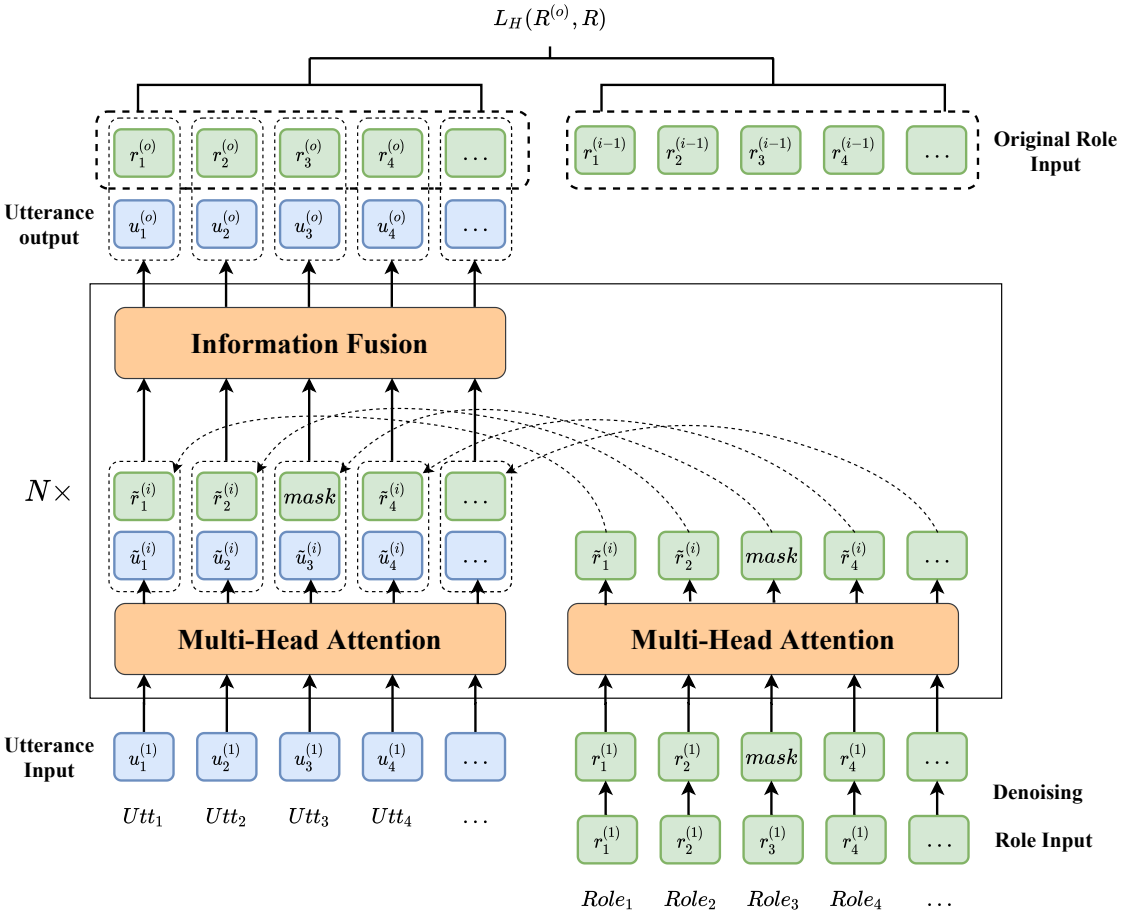

Figure 3: The procedure of denoising Role Auto-encoder pretraining task. In the input role sequence, 15% of the role tags are masked, and the original role sequence is restored after the model merges role embedding and utterance embedding.

encoder input. It is a kind of interval segment embedding to distinguish multiple topics in meeting transcripts. For example, for utterance $(u_1, u_2, [TSEP], u_3, u_4, [TSEP], u_5...)$, where every two utterances belong to the same topic. we would assign the topic segmentation embedding $(t_a, t_a, t_a, t_b, t_b, t_b, t_a...)$.

Besides, we restrict the scope of attention computation to different topics, it can alleviate the noise impact caused by long-distance dependencies. In each Transformer block, multiple self-attention heads are used to aggregate the output embeddings of the previous layer. The attention score is calculated as follows:

$$A = softmax(\frac{QK^T}{\sqrt{d_k}} + O)V$$
$$O_{ij} = \begin{cases} 0 & the\ same\ topic \\ -\infty & other\ topics \end{cases} \quad (10)$$

We introduce the mask matrix O determines whether a pair of tokens can be attended to each other. Each utterance can only focus on the other utterances under the same topic, as illustrated in Figure 4.

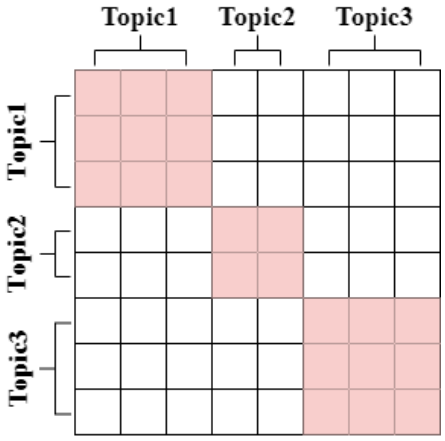

Figure 4: The attention mask matrix guided by topic segmentation.

5 Experiment

Datasets. We tested our model on the widely used AMI and ICSI meeting summary datasets. Both the AMI and ICSI meeting corpus are multi-modal datasets which consist hundreds of hours of meeting recordings. Each meeting participant has a role tag, such as marketing manager, industrial designer, professor and so on. Moreover, the datasets also have a wide range of annotations, including conversation behavior, topic segmentation, extractive and abstract summaries. In AMI, there are an average of 4757 words in the transcripts, 289 circles, 4 speakers and 322 words in the abstract. In ICSI, the average conference record was 10189 words, 464 laps, 6.2 speakers, and 534 words in the abstract. We roughly divide datasets into training, valid and test part according to the ratio of 8:1:1.

Except using some news domain pre-training data in hmnet, we also introduce dialog domain datasets MediaSum (Zhu et al., 2021) and TV4Dialog (Leilan Zhang, 2019) as the pre-training datasets. MediaSum is a large scale modular media interview dataset composed of 463.6k texts and abstracts. The dataset is mainly from the interview records of NPR and CNN, with an average of 30 rounds per conversation, six to seven participants, and a total of 1554 words. The dataset has the characteristics of large data scale, multi-party dialogue in multiple fields and clear theme. TV4Dialog is a multi round dialogue corpus extracted from the subtitles of American TV series. It contains about 260000 utterances with speaker tags. In order to make the dataset suitable for our pre-training task, we do the following three aspects of preprocessing. Firstly, the original dataset is cleaned, and with the help of Spacy library, the POS and ENT tags are added. Secondly, we also randomly stitch several different dialogues on the cleaned data to simulate the change of conversation topic in real conference scene. Finally, aiming at the problem that the label quality of speakers in pre-training datasets are uneven, we combine some of the similar role tags.

Metrics. We evaluated performance with the widely used ROUGE-1, ROUGE-2 and ROUGE-SU4 metrics in automatic summarization. More specifically, it focuses on measuring the number of overlapping units such as n-gram between the generated summary and the reference summary. When matching the reference summary and the summary to be evaluated, ROUGE-SU4 does not require that the gram must be continuous, and can "skip" some words. Through the above three metrics, we can measure the generated summary quality from many aspects.

6 Results

Ablation Study. We conduct ablation experiments to analyze the impact of each part of the model on the final results.

In table 1, we compare the quality of summary generated by the model without pre-training, adding GSG pre-training, adding dRA pre-training and introducing topic segmentation. It can be seen that GSG task is the most obvious to improve the effect of the model, followed by the topic segmentation task, and finally the dRA task. Many words in the human annotation of AMI and ICSI are obtained directly from the original text. GSG task simulates human to extract key information from the original text of the meeting, so it has a significant impact on the final summary generation. The addition of the dRA task did not change the results much. The reason may be that there are few participants in the meeting scene, and the communication order between speakers is repeated, so the model can easily predict the masked roles.

We also compare the effect of different unsupervised key sentence extraction methods in GSG task on the final summary. Table 1 shows that the MMR and TextRank methods are better than the method of calculating the rouge score between different sentences in the original text, and the method of using MMR gets the best results in meeting summary.

Automatic Evaluation. We compare our proposed method with previous methods for the problem of abstractive meeting summarization.

CoreRank (Shang et al., 2018) construct undirected weighted graph and calculate the corerank value of nodes, it is state-of-the-art extractive summarization method. PGN method is the pointer-generator network (See et al., 2017), which focuses on addressing the reproducing and repeating problem in general abstractive text summarization task. HASMR (Zheng et al., 2020) proposes a hierarchical neural encoder based on adaptive recurrent network to learn the semantic representation of conference session based on adaptive session segmentation. BertSum (Liu and Lapata, 2019) is a pretraining model with good performance in the field of text generation. MM is a multimodal model,

Model	R-1	R-2	R-SU4	R-1	R-2	R-SU4
	AMI			ICSI		
NoPretrain	49.17	16.34	22.26	40.75	10.02	18.47
GSG(MMR)	**52.06**	20.68	24.75	43.79	11.44	19.72
GSG(MMR)+dRA	51.59	21.09	24.64	44.03	11.73	19.47
GSG(MMR)+dRA+Topic	51.90	**21.16**	**25.12**	**44.41**	**11.86**	**19.95**
GSG(Rouge)+dRA+Topic	50.96	20.03	24.44	42.86	11.28	19.88
GSG(TextRank)+dRA+Topic	51.11	20.75	24.68	43.71	11.46	19.62

Table 1: The ablation experiment about GSG pretraining, dRA pretraining and topic segmentation on the results of model generation. We also compare the effect of different unsupervised key sentence extraction methods in GSG task on the final summary. Table 1 shows that the MMR and TextRank methods are better than the method of calculating the rouge score between different sentences in the original text, and the method of using MMR gets the best results in meeting summary.

Model	R-1	R-2	R-SU4	R-1	R-2	R-SU4
	AMI			ICSI		
CoreRank(2018)	37.86	7.84	/	29.82	4.00	/
PGN(2017)	40.77	14.87	18.68	32.00	7.70	12.46
BERTSUM(2019)	37.62	10.68	/	/	/	/
HASMR(2020)	48.64	17.45	22.13	/	/	/
MM(2019)	**53.29**	13.51	/	/	/	/
HMNet(2020)	53.02	18.57	24.85	**46.28**	10.60	19.12
Our model	51.90	**21.16**	**25.12**	44.41	**11.86**	**19.95**

Table 2: ROUGE-1, ROUGE-2, ROUGE-SU4 scores comparison of different models.

which introduces the external semantic information of topic segment and visual focus on attention (VFOA). HMNet introduces a hierarchical transformer structure for the first time and trains the model on the news data in advance.

Table 2 shows the ROUGE scores of our model and previous models on AMI and ICSI datasets. Our model performs well on different Rouge scores. Compared with HMNet, our model has a significant improvement in the score of ROUGE-2, it proves the effectiveness of the pre-training task. Bertsum model, which performs well in the field of text summarization, does not get good results in meeting transcripts. It shows that there is a big gap between the semantic distribution of meeting transcripts and the traditional news text, so it is necessary to propose a semantic extraction method for multi person long dialogue text. MM model and our model verify that external semantic information, such as topic segmentation, conversation behavior and conversation focus, have a promoting effect in meeting summary.

7 Conclusion

In this paper, we apply two kinds of pre-training tasks to hierarchical transformer network to improve the effect of meeting summary generation. We adjust the structure of utterance level encoder to better integrate the role vector of each participant. In addition, we introduce additional topic segmentation information to constrain the attention range, which is further improved the model's performance. Experimental results show that our model performs well on AMI and ICSI datasets.

In the future, we plan to add some new pre-training tasks to obtain the semantic information of discourse coherence in conference texts. we plan to utilize knowledge graph and dialog act, which can better capture salient information from the transcript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Project (Grant No. 61977045).

References

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. 2013. Generalized denoising auto-encoders as generative models.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2020. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Qiang Zhou Leilan Zhang. 2019. Automatically annotate tv series subtitles for dialogue corpus construction. Lanzhou, Gansu, China. APSIPA Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821.

Shikib Mehri and Maxine Eskenazi. 2019. Multi-granularity representations of dialog.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Fumio Nihei, Yukiko I. Nakano, and Yutaka Takase. 2016. Meeting extracts for discussion summarization based on multimodal nonverbal information. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 185–192, New York, NY, USA. Association for Computing Machinery.

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

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Jean-Pierre Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation.

Ilya Sutskever, Oriol Vinyals, and Quoc V.Le. 2014. Sequence to sequence learning with neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiyuan Zheng, Zhou Zhao, Zehan Song, Min Yang, Jun Xiao, and Xiaohui Yan. 2020. Abstractive meeting summarization by hierarchical adaptive segmental network learning with multiple revising steps. *Neurocomputing*, 378:179–188.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining.