A Lite Hierarchical Model for Dialogue Summarization with Multi-Granularity Decoder

Tong Zheng
Osaka Prefecture University

Ryosuke Saga
Osaka Metropolitan University
r.saga@omu.ac.jp

Abstract

Abstract dialogue summarization generation has recently attracted considerable research attention, especially in using hierarchical models to accomplish abstract dialogue summarization tasks successfully. However, problems in recent studies often include an excessive amount of model parameters and long training time mainly because existing dialogue summaries of hierarchical models are typically generated by adding extra encoders and attention layers in the decoder to enhance learning and summarization generation ability of the model. Hence, designing an increasingly lightweight hierarchical model is necessary. This paper proposes a lightweight hierarchical model named ALH-BART to generate high-accuracy dialogue summaries rapidly. The proposed hierarchical model includes word and turn encoders, which enhance the ability of the model to understand dialogue. A multigranularity decoder in the model is also proposed to decode word- and turn-level information in the decoder at the same time. Encoder parameters in multihead self-attention are provided for each corresponding multihead self-attention to reduce the number of model parameters and improve the speed of model learning effectively. Finally, the effectiveness of the model is verified on SAMSum and DialogSum datasets.

1. Introduction

Text chat social software, ex. Whatsapp, LINE, Messenger, WeChat, Telegram etc., have become an integral part of daily life in recent years. Users may join multiple chat groups to communicate with many users simultaneously and accumulate hundreds or thousands of unread messages in group chats while using the software. However, this process is inefficient if someone wants to know what other users are talking about in a group chat because it requires a considerable amount of time to read each message [1]. Therefore, a means for translating large daily interactions into natural, concise, and informative text, such as abstract conversation summaries, is important.

Table 1. Example of Dialogue data and Summarization Reference

Dialogue Transcript(15 Turns)
(1)Hannah: Hey, do you have Betty's number? (2)Amanda: Lemme check (3)Hannah: (file_gif) (4)Amanda: Sorry, can't find it. (5)Amanda: Ask Larry (6)Amanda: He called her last time we were at the park together . . . . . . (12)Amanda: Just text him(emoji_font) (13)Hannah: Urgh.. Alright (14)Hannah: Bye (15)Amanda: Bye bye
Reference:
Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.

Although existing text summarization generation models perform properly for document-level text summarization, such as news, achieving the same results with dialogue text is difficult [2, 3] because dialogue text is different from document-level text. Dialogue text presents the following characteristics (Table 1). (1) Frequent speaker interruptions, such as repetitions, pseudostarts, and hesitations, exist and important information is located in different parts of the conversation. These unstructured features pose a challenge for summary generation models [4]. (2) Emojis may be used in dialogue text. (3) Multiple participants may be involved at the same time, with each contributor speaking in a different way. (4) Multiple topics are discussed at the same time.

Research on summarizing unstructured and complex dialogue text has progressed considerably. For example, summary sentences are generated by constructing

URI: https://hdl.handle.net/10125/102901
978-0-9981331-6-4
(CC BY-NC-ND 4.0)

a co-occurrence word graph, weighting edges with a combination of several scoring methods based on CoreRank, and solving budgeted submodular maximization [5]. Some hierarchical models are used to generate summarizations by building a graph of relationships between dialogue texts [4, 6, 7, 8]. Existing models enhance their learning ability by increasing the number of layers in the model to understand increasingly complex dialogue text and produce accurate generated summaries [4, 9, 10, 6]. However, the number of model parameters and length of model computation time are difficult to apply and challenges persist when reducing weight while maintaining the accuracy of the generated model. Therefore, an effective dialogue summary generation model that can minimize the consumption of computational cost while increasing the accuracy of the summary results is necessary.

On this basis, we propose a hierarchical model that can effectively reduce the number of model parameters according to the BART model [11] and share parameters of the encoder at each stage of the hierarchical model through the concept of ALBERT. In addition, word and turn encoders are constructed to ensure the accuracy of the generated summary.

This study presents the following contributions. (1) A lightweight hierarchical model for generating conversational sentence summaries is proposed and (2) satisfactory summarization accuracy is achieved while reducing the number of model parameters.

2. Related Work

2.1. Transformer

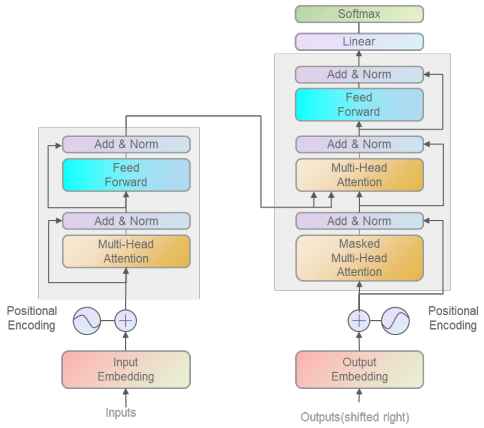

Figure 1. Transformer Model architecture.[12]

Vaswani et al. proposed a new neural network Transformer [12] based on the self-attention mechanism (Figure 1). Transformer consists of Encoder and Decoder.

2.1.1. Encoder The encoder obtains the hidden representation of each input token of the model and consists of six layers, with each layer composed of two sublayers, namely, multihead attention and feedforward networks. The output of each sublayer is normalized through residual connection and normalization.

$$sub\_layer\_out = LayerNorm\left(x + \left(SubLayer\left(x\right)\right)\right)$$

where x is the input of each sublayer and $SubLayer\left(x\right)$ is the output of each sublayer.

The output dimension of the embedding layer and the two sublayers must be the same to achieve residual connection. The overall output of the encoder can be expressed as follows

$$Encoder\_out = Encoder\left(x\right)$$

where x is the input of the encoder.

2.1.2. Decoder The decoder processes the output of the encoder and generates each element of the model output one after the other. Each layer consists of three sublayers: masked multihead attention, multihead attention, and feedforward network. Notably, masked multihead self attention is similar to the encoder's attention but adds a mask, which can provide invisibility to something that follows a word . Similar to the encoder, the decoder can also normalize through residual connection and normalization.

2.1.3. Multihead Attention Multihead attention helps the model depict existing interdependencies between words, such as referential and modifier relationships, and maps the embedding vector of the word to three new vectors (query, key, and value) with the same dimension. A similarity vector of dimensions equal to the number of words in the current sentence is obtained when all words have been compared. Note that each element of the vector represents the similarity between the word in the corresponding position and the word that sent the query. Finally, values for each word are weighted and their sum is obtained to compute the output.

$$attention\_out = Attention\left(Query, Key, Value\right)$$

2.1.4. Feedforward Networks Feedforward networks are used to map each word vector. The transformer uses two feedforward networks and ReLU activation function between them. The feedforward network can be expressed as follows:

$$FFN\left(x\right) = max\left(0, xW_1 + b_1\right)W_2 + b_2$$

where x is the input to the feedforward network, W_1 and W_2 are weights, and b_1 and b_2 are biases

2.2. Bidirectional and Auto-Regressive Transformers (BART)

Lewis et al. [11] proposed the transformer-based encoder– decoder model called BART that uses specific pretraining methods to (1) corrupt text with an arbitrary noise function and (2) restore the original text by training the model. Also, BART evaluated various noisy methods, randomly corrupted the order of the original text, and explored the optimal performance with a novel text-filling method by replacing text segments with individual mask tokens.

BART has been successfully used on a variety of tasks, including text generation and comprehension; performance comparable to RoBERTa [13] on the GLUE[14] and SQuAD datasets [15]; obtained the best results on tasks such as abstract dialogue, QA, and text summarization; and performed properly on the XSum dataset [16] with ROUGE [17] performance improvement using comparable training resources. BART improved the BLEU [18] by 1.1% compared with the back-translation system [19] through target language pretraining in the machine translation task.

Therefore, we built our model on the basis of BART because it maintains high performance in the text summarization task.

2.3. Hierarchical model

Hierarchical models have been increasingly used in language generation tasks in recent years. Zhu et al. proposed HMNet [9], which builds a hierarchical structure from word- and turn-level information and uses news summary data to overcome the lack of training data (Figure 2). An approach to summarization that takes into account hierarchical relationships among sentences and words is hierarchical-BART (Hie-BART) [10] (Figure 3). Specifically, input documents are divided into word- and sentence-level information and both word- and sentence-level information types are computed in the encoder of BART. A summary that considers both word- and sentence-level information types can then be generated by combining word- and sentence-level multigranularity outputs.

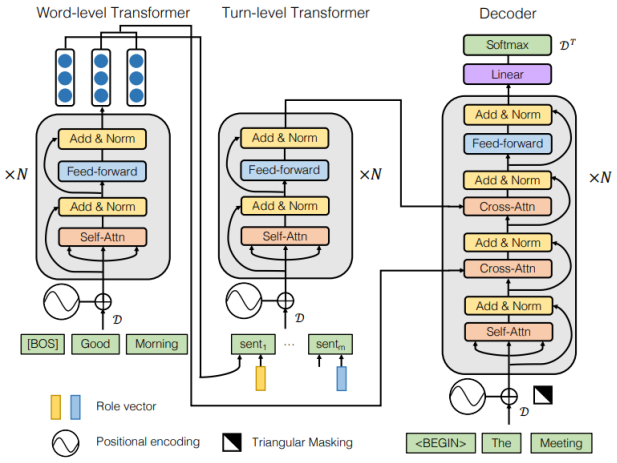

Figure 2. HMNet Model architecture.[9]

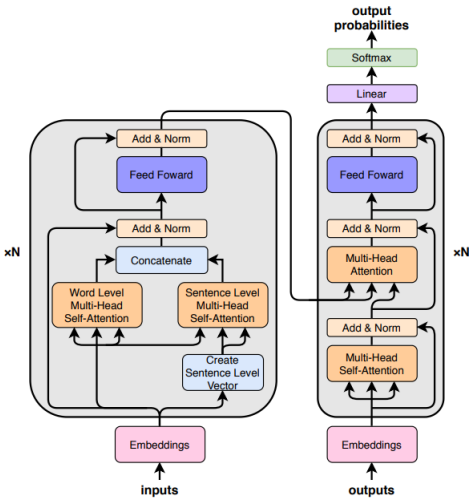

Figure 3. Hie-BART Model architecture.[10]

Feng et al. proposed DDAMS [6], which is a meeting text summarization generation model that can understand multispeaker dialogues and solve the problem of ignoring different relations between discourses in HMNet and Hie-BART by constructing a relational graph with relations between different discourses. The researchers also put forward a conversation-aware data augmentation method for constructing a summary corpus that can increase the data available for model training by a factor of 20 over the original dataset.

Chen et al. presented S-BART[4], which is a text summarization generation model that can perceive discourse relations between conversations (Figure 4). The researchers designed discourse relation and active graphs to describe the behavior in a dialogue. These

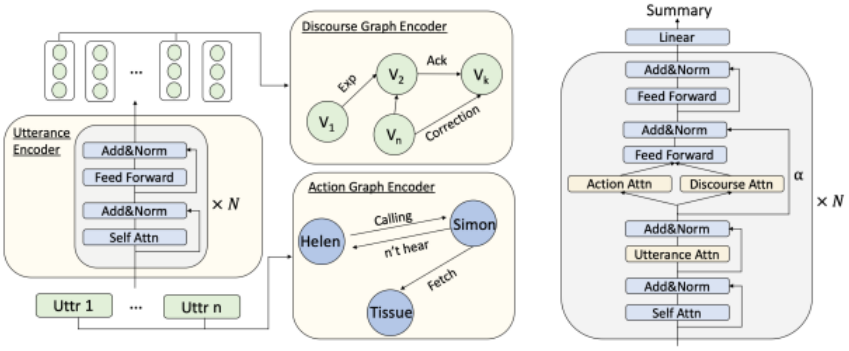

Figure 4. S-BART Model architecture.[4]

graphs are encoded together with dialogue data, and different levels of information are integrated via a multigranularity decoder to summarize the given dialogue data at the decoder.

However, notably, additional layers of the encoder and decoder are computationally expensive for model parameters although these hierarchical models improve the accuracy of summary generation. For example, BART, Hie-BART, HMNet, and S-bart presents about 121M (according to BART-base settings), 140M (according to BART-base settings), 160M (according to the settings of the transformer six-layer model), and 171M (according to BART-base settings) model parameters, respectively. This feature limits not only the depth of layers that can be added to these models but also increases training time and hardware requirements, such as GPU and memory of the computer, thereby increasing the difficulty in real-life applications.

2.4. A Lite BERT (ALBERT)

Lan et al. designed "A Lite BERT" (ALBERT) [20] with remarkably fewer parameters than the traditional BERT architecture.

There are two methods of parameter reduction that are worth mentioning in ALBERT, and some of these ideas are also used in this study. The first technique is a decomposition of the embedding parameterization. A large lexical embedding matrix is decomposed into two smaller matrices to separate sizes of hidden and embedding layers. This separation facilitates the increase in the number of hidden layers without significantly increasing the vocabulary embedding parameters. A one-shot vector is first mapped to the low-dimensional lexical embedding space E and then to the hidden space instead of directly mapping a one-shot vector onto a hidden space of size H. This decomposition approach reduces word embedding parameters from $O(V \times H)$ to $O(V \times E + E \times H)$ and significantly reduces the number of parameters when H is considerably larger than E.

The second technique involves parameter sharing between layers that prevents the increase of the number of parameters as the network depth increases. Although ALBERT and BERT-large present similar configuration, the former contains 1/18 of the number of parameters of the latter and exhibits a learning speed 1.7 times faster than the latter. Notably, experiments have shown that reducing the number of parameters of the model is accompanied by constant performance degradation. Lan et al. also show that sharing parameters has the effect of stabilizing the network parameters by making the output of each layer less oscillatory than the original model.

The model encoder in this study was constructed according to ALBERT. Sharing parameters of the attention layer of the encoder can reduce the number of parameters of the model and the overall model weight.

3. A Lite Hierarchical BART (ALH-BART)

3.1. Overview

Before proceeding with the discussion, we would like to make a few definitions. A series of utterances made by a speaker is generally called a Turn [21]. This turn is assumed identical until the next utterance by another speaker. A dialogue consists of a number of turns, and Zhu et al. described that if a model can understand relationships among individual turns, then it will exert a positive impact on improving model comprehension and generating summary accuracy.

The model we propose in this study is based on the BART model, called A Light Hierarchical BART (ALH-BART). We first encode the word- and turn-level information in data using word and turn encoders and then combine the two levels of information using the multigranularity decoder to reduce the consumption of

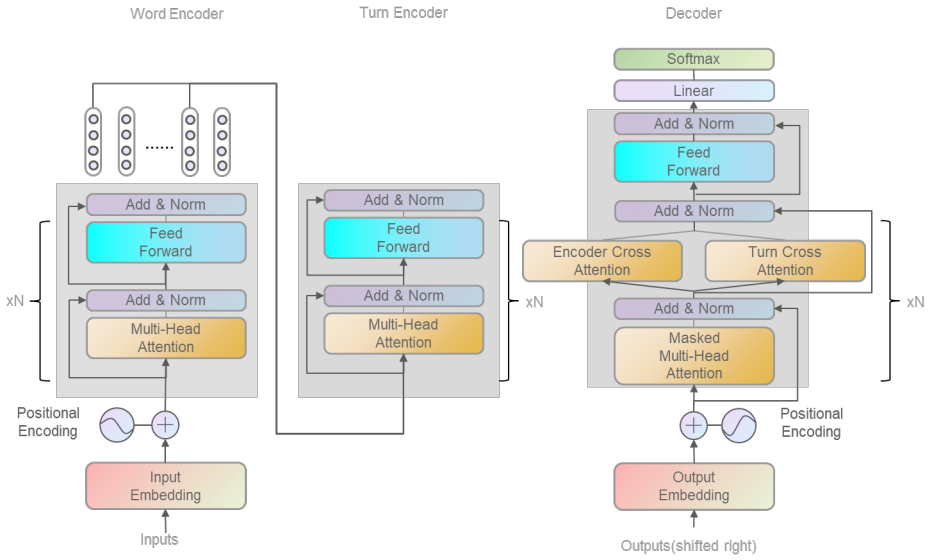

Figure 5. Proposed Model:ALH-BART architecture.

computational cost while generating summaries with sufficient accuracy. Accordingly, word- and turn-level information can be decoded to produce an increasingly accurate summary. And cross-layer parameter sharing reduces the number of model parameters and minimizes computation cost consumption. The overall architecture is shown in Figure 5.

3.2. Encoder

ALH-BART utilizes word and turn encoders to find the hidden representation.

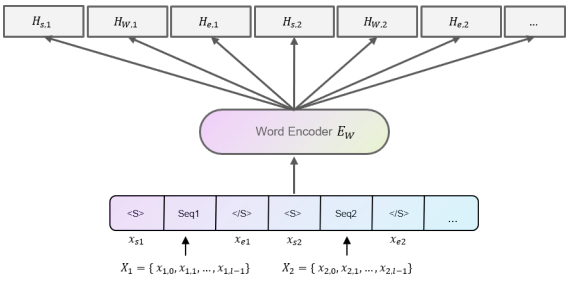

Figure 6. Example of Word Encoder process.

The word encoder is the layer that obtains the hidden representation of each input token of the model. On the basis of the BART model, the word encoder (E_W) is initialized in this study, and the corresponding hidden representation $H_{W,i} = \{h_{i0}, h_{i1}, \ldots, h_{il-1}\}$ is generated for the input sentence i of length l, $X_i = \{x_{i0}, x_{i1}, \ldots, x_{il-1}\}$.

$$H_{W,i} = E_W\left(X_i\right)$$

In addition, input sentences are divided by adding the special token <s> at the beginning of each input sentence and </s> at the end. Each embedding and hidden representation is generated by the following equations;

$$H_{s,i} = E_W\left(x_{si}\right)$$

$$H_{e,i} = E_W\left(x_{ei}\right)$$

where x_{si} and $H_{s,i}$ are embedded and hidden representations of <s>, x_{ei} and $H_{e,i}$ are embedded and hidden representations of </s>(Figure 6).

The turn encoder is the layer that processes all turns in a conversation to obtain their hidden representation. We will use the embedded output of the special token <s>from the word encoder $H_{s,i}$, which represents the turns as input to the turn encoder. The input containing k turns is $X_T = \{H_{s1}, H_{s2}, \ldots\ldots, H_{sk}\}$ and the corresponding embedded output is $H_T = \left\{H_{s1}^T, H_{s2}^T, H_{s3}^T, \ldots\ldots, H_{sk}^T\right\}$.

$$H_T = E_T\left(X_T\right)$$

3.3. Multi-Granularity Attention

We designed the model according to the framework of Chen et al.[4] to improve the incorporation of the turn-level information into the generated summary. A multigranularity decoder that includes multigranularity attention, which is a combination of the Encoder Cross Attention layer and the Turn Cross Attention compared with the BART model, and adds a turn cross-attention layer to the cross-attention layer is built in this study.

The multigranularity decoder computes word- and turn-level hidden representations from word and turn encoders as multigranularity attention to obtain the representations of A_W and A_T containing word- and turn-level information, respectively.

$$A_W = Attention\,(Q_M, K_W, V_W)$$

$$A_T = Attention\,(Q_M, K_T, V_T)$$

where K_W and V_W are the key and value mapped by the output of the word encoder; K_T and V_T are the key and value mapped by the output of the turn encoder, respectively; Q_M is the query mapped by the output of decoder's masked multihead attention. Here, the initial parameters of the turn cross attention may adversely affect the pretrained BART decoder in that stage because these parameters are randomly initialized. Therefore, we apply a one trainable parameter α called ReZero (residual with zero initialization)[22] to the residual connections after the multigranularity attention layer in the decoder layer to perform normalization shown in the below equation.

$$A_m = LayerNorm\,(A_W + \alpha A_T)$$

where A_m is the normalized output from residual connection with ReZero. Finally, it is passed through the feedforward network to become the new representation A_M.

$$A_M = FFN\,(A_m)$$

3.4. Cross-layer parameter sharing

Lan et al. proposed three methods of parameter sharing in ALBERT: attention-related parameter sharing only, feedforward network-related parameter sharing only, and all parameter sharing[20]. Although sharing all and feedforward network-related parameters can significantly reduce the number of model parameters, it seriously impacts the model performance. Sharing only parameters related to the attention layer can reduce the number of parameters while exerting a slight impact on the model performance.

Therefore, only attention-related parameters are shared in the model to ensure the accuracy of model generation in this study. Each layer shares parameters of the first layer of multihead attention, including the multihead attention of word and turn encoders, to ensure that only the multihead attention of the first layer is learned when training the model. In addition, only the multihead attention of the last layer of each encoder is trained to maintain the learning capability of the model based on the idea of universal transformer[23].

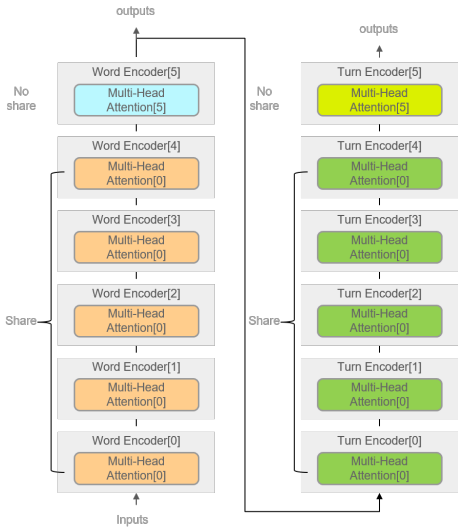

Figure 7. Example of Cross-layer parameter sharing in 6-layer Encoder.

Figure 7 is an illustration of the concept. As shown in this figure, the same Multihead-Attention parameter (Multihead-Attention[0]) is shared from layer 0 to layer 4 of the Word/Turn Encoder, and only the final layer holds it as a different parameter. By using this idea, the model avoids the process of computing separate multihead attention parameters for each layer.

The change in the number of model parameters after sharing and the associated training time are shown in Table 2. The number of parameters for BART and S-BART are also included for reference. The number of parameters is larger than that of BART because it is an extended model of BART, but it can be executed with fewer parameters than S-BART. The parameter sharing reduces the number of parameters by about 14M, which in turn reduces the occupancy of GPU memory and the actual training time.

4. Experiment

4.1. Goal and Environment

Experiments were conducted to demonstrate that our model is providing highly accurate summary results. To evaluate our model, ROUGE benchmark [17] and human evaluation are used in this study.

The ROUGE Benchmark is an automatic evaluation metric that evaluates the results of all models and compares the quality of system-generated summaries and translations with those generated by humans. ROUGE-1 is used to evaluate the co-occurrence of 1-grams between system and correct summaries.

Table 2. Model parameters, GPU memory occupancy and training time on GeForce RTX 3060 (12 GB).

Model	Para.	GPU Memory	Time/epoch
BART	121M	42.98%	838s
S-BART	171M	58.51%	1860s
ALH-BART(no shared)	151M	54.46%	1486s
ALH-BART(shared)	130M	44.24%	860s

Table 3. SAMsum Corpus and DialogSum Corpus.

Corpus	Train/dev/test	Tokens/dialog	Tokens/turn
SAMsum	14732/818/819	94	8.4
DialogSum	12460/500/500	131	13.8

ROUGE-2 is utilized to examine the co-occurrence of 2-grams between system and correct summaries. ROUGE-L is applied to evaluate the number of words that co-occur between system and correct summaries along the sentence order. We used the implementation by [24] for evaluating ROUGE.

Human evaluation is also used to evaluate the model's ability to generate summaries qualitatively. This evaluation method is based on the method used by S-BART. Specifically, five graduate students in the English department were asked to select and score 100 summaries generated from BART, S-BART, and the model randomly. The quality of generated summaries was evaluated using a five-point Likert scale, with 1 being the worst and 5 as the best, on the basis of grammar (whether significant grammatical errors exist), succinctness (whether redundant information is included), informativeness (whether maximally important content is covered), and factualness (whether the speaker and his/her actions are correctly integrated)[4]. To improve score quality, each set of evaluation was rated by three or more people.

SAMSum[25] and DialogSum [26] datasets were used in this study. The SAMsum dataset contains 16,369 conversations, providing one-on-one conversations about everyday topics such as scheduling meetings, discussing news events, etc. We also tested the model in multi-person conversations using the DialogSum dataset, a dataset of conversations between two or more people, containing a total of 13,460 conversations. The data statistics for the two datasets are shown in Table 3.

The model in this study is compared with the following representative model approaches:

- Pointer Generator [27]: This model is a generative summary model based on sequence to sequence, improved in two ways. Firstly, this model copies words in the source text with the pointer to retain the generator's ability to generate new words and copy information accurately. Secondly, coverage mechanisms are used to keep track of what has been summarized and prevent duplication.
- Transformer [12]: Transformer is a deep learning model based on the encoder–decoder structure that has been successfully used in many areas, such as machine translation. OpenNMT library [28] is utilized to train the transformer model.
- D-HGN [8]: This is a multiperson dialogue summary generation model that facilitates dialogue comprehension and summary generation by constructing a graph of discourse nodes and knowledge nodes and integrating common sense knowledge. We apply this model according to the setting of the original paper.
- BART [11]: BART is more suitable for text generation than BERT because it incorporates features of both the bidirectional encoder of BERT and the left-to-right decoder of GPT [29] and is built on the basis of the standard transformer model. BART also presents more bidirectional contextual information than GPT, which is configured according to the BART-base model.
- Multi-View Seq2seq [30]: This model allows conversations to be viewed from several different perspectives, thereby resulting in different discourse structures. The encoder in this model encodes the text from multiple viewpoints, and its decoder combines the information from multiple viewpoints to generate a summary of the conversation. The BART-base model is applied as a base for comparison.
- S-BART [4]:The model is a text summary generation model that senses discourse relations between conversations. The model based on the original setting of S-BART was applied.

Our models utilize the pretrained BART-base model for initial parameters, with a learning rate of 3e-5 and

Table 4. ROUGE benchmark results for SAMsum Corpus test set. F:F-measure, P: Precision, R: Recall

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
Pointer Generator	40.08	-	-	15.28	-	-	36.63	-	-
Transformer	37.27	-	-	10.76	-	-	32.73	-	-
D-HGN	42.03	-	-	18.07	-	-	39.56	-	-
Multi-View Seq2seq	45.56	52.13	44.68	22.30	25.58	22.03	44.70	50.82	43.29
BART	45.15	49.58	45.97	21.66	23.95	22.16	44.96	48.92	44.26
S-BART	46.07	51.13	46.24	22.60	25.11	22.81	45.00	49.82	44.47
ALH-BART(no shared)	46.37	52.60	44.89	20.99	24.24	20.58	43.25	49.09	41.85
ALH-BART(shared)	46.29	52.64	44.66	20.76	23.84	20.30	42.99	48.85	41.50

a warm-up step of 120, for all experiments. The encoder and decoder were generally set according to the BART-base model. Specifically, the hidden layer, attention head, and dropout were set to 768, 12, and 0.1, respectively, for word and turn encoders of a six-layer model. Note that a six-layer model was also established for the decoder. All experiments were performed on a GeForce RTX 3060 (12 GB memory).

4.2. Results

Table 4 and Table 5 show the results of ROUGE benchmark. In each table, "no shared" indicates no parameter sharing, and "shared" indicates parameter sharing. Also, F-value(F), Precision(P), and Recall(R) for methods other than ours (BART, S-BART, etc.) are taken from the reference values in the S-BART paper[4], since they were obtained using the same dataset.

Now, on the SAMSum dataset, our models (ALH-BART(no shared) and ALH-BART(shared)) outperform all models in F-measure for ROUGE-1. On the other hand, we found that ROUGE-2 and ROUGE-L differ from BART, S-BART, and Multiview Seq2seq. This is strange because, in general, when the F value of ROUGE-1 improves, ROUGE-2 and ROUGE-L often do correspondingly better. Therefore, to analyze this in more detail, we focused on Precision and Recall. We find that Recall is lower for our method than for other methods. The reason for this can be seen from the summary examples shown in Table 6; ALH-BARTs tend to produce shorter summary sentences than other models. This may be the reason why the overall number of recalls is lower. (For example, if the summary sentence is shorter than the ground truth, the Recall cannot be 1 in any way, indicating that the length of the summary sentence is important.)

Human Evaluation was conducted on the results of these analyses. Table 7 shows the result of Human Evaluation. The results showed that our model outperforms BART in terms of grammatical accuracy and conciseness likely because our model can combine word- and turn-level information. Moreover, the model in this study was inferior to S-BART in terms of grammatical accuracy, inclusion of important information, and factuality of the generated summary.

As summary, The results of the ROUGE benchmark and human evaluation revealed that our model is generally superior to BART but marginally inferior in the inclusion of important information and factuality of the generated summary. Although our model is inferior to S-BART, it contains fewer parameters, occupies less GPU memory, and learns faster. Therefore, our model demonstrated that it can achieve satisfactory generation accuracy while effectively reducing the number of model parameters.

5. Conclusion

The hierarchical model ALH-BART is proposed in this study to sense in both word and turn levels. The number of model parameters can be effectively reduced by sharing model parameters. On the one hand, experiments on the SAMSum dataset showed that the proposed model can effectively reduce the number of model parameters and achieve high-quality summaries. On the other hand, experiments using the DialogSum dataset indicated that the model can effectively handle the task of generating summaries of conversations between multiple people.

Future issues to be addressed are related to the current contribution, but also to accuracy and number of parameters, as well as applicability. Regarding accuracy, we plan to first examine ways to improve the summarization accuracy of the model (ROUGE-2 and ROUGE-L) while also considering indices not based on co-occurrence, such as ROUGE-S/SU. Also, for parameter reduction, we used the efficiency improvement by parameter sharing in this study, but it is also necessary to investigate the impact of reduction methods with improved embedding, such as those

Table 5. ROUGE benchmark results for DialogSum Corpus test set.(F-value)

	ROUGE-1	ROUGE-2	ROUGE-L
Transformer	35.91	8.74	33.50
BART	42.64	18.18	41.40
ALH-BART(shared)	43.22	16.48	39.54

Table 6. Summarization comparison

Model	Example 1	Example 2	Example 3
BART	Amanda can't find Betty's number. Larry called her last time they were at the park together. Hannah would rather Amanda texted him.	Eric and Rob are watching a Russian stand-up.	Bob will help Lenny with picking the outfit for her.
S-BART	Amanda can't find Betty's number. Larry called her last time they were at the park together. Hannah doesn't know Larry well. Amanda will text him.	Eric and Rob are watching a Russian comedian's stand-up.	Lenny is looking for a outfit for her birthday. Bob has four black pairs. Lenny will buy the first pair of purple trousers.
ALH-BART (no shared)	Amanda can't find Betty's number. Hannah doesn't know Larry. Hannah suggests Amanda texting him.	Eric and Rob are watching a funny piece of funny comedy.	Bob will buy Lenny two purple trousers.
ALHBART (shared)	Amanda can't find Betty's number. She will text Larry.	Eric and Rob are discussing their favourite stand-up comedy.	Bob will help Lenny choose the best pair of purple trousers.
Ground Truth	Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.	Eric and Rob are going to watch a stand-up on youtube.	Lenny can't decide which trousers to buy. Bob advised Lenny on that topic. Lenny goes with Bob's advice to pick the trousers that are of best quality.

used in ALBERT. Regarding applicability, the current dialogue summary generation model will be extended to various domains, such as meetings, discussions, and other multi-topic dialogues that require long corpora and additional participants. We hope to deepen the discussion in dialogue summarization while taking these considerations into account.

References

[1] S. Gao, X. Chen, Z. Ren, D. Zhao, and R. Yan, "From standard summarization to new tasks and beyond: Summarization with manifold information," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[2] Z. Zhao, H. Pan, C. Fan, Y. Liu, L. Li, M. Yang, and D. Cai, "Abstractive meeting summarization via hierarchical adaptive segmental network learning," in *The World Wide Web Conference*, p. 3455–3461, Association for Computing Machinery, 2019.

[3] Y. Lei, Y. Yan, Z. Zeng, K. He, X. Zhang, and W. Xu, "Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization," in *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7823–7827, 2021.

[4] J. Chen and D. Yang, "Structure-aware abstractive conversation summarization via discourse and action graphs," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1380–1391, Association for Computational Linguistics, June 2021.

[5] G. Shang, W. Ding, Z. Zhang, A. Tixier, P. Meladianos, M. Vazirgiannis, and J.-P. Lorré, "Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 664–674, Association for Computational Linguistics, July 2018.

[6] X. Feng, X. Feng, B. Qin, and X. Geng, "Dialogue discourse-aware graph model and data augmentation for meeting summarization," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3808–3814, International Joint Conferences on Artificial Intelligence Organization, Aug. 2021.

[7] L. Huang, L. Wu, and L. Wang, "Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5094–5107, Association for Computational Linguistics, July 2020.

[8] X. Feng, X. Feng, and B. Qin, "Incorporating

Table 7. Human Evaluation Results.

	Grammar	Factualness	Succinctness	informativeness
BART	4.43	4.01	4.10	3.84
S-BART	4.52	4.12	4.36	3.99
ALH-BART(shared)	4.44	3.97	4.37	3.74

commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks," in *China National Conference on Chinese Computational Linguistics*, pp. 127–142, Springer, 2021.

[9] C. Zhu, R. Xu, M. Zeng, and X. Huang, "A hierarchical network for abstractive meeting summarization with cross-domain pretraining," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 194–203, Association for Computational Linguistics, Nov. 2020.

[10] K. Akiyama, A. Tamura, and T. Ninomiya, "Hie-bart: Document summarization with hierarchical bart," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 159–165, 2021.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Association for Computational Linguistics, July 2020.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, (Brussels, Belgium), pp. 353–355, Association for Computational Linguistics, Nov. 2018.

[15] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Austin, Texas), pp. 2383–2392, Association for Computational Linguistics, Nov. 2016.

[16] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 1797–1807, Association for Computational Linguistics, Oct.-Nov. 2018.

[17] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[19] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Association for Computational Linguistics, Oct.-Nov. 2018.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020.

[21] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the organization of conversational interaction*, pp. 7–55, Elsevier, 1978.

[22] T. Bachlechner, B. P. Majumder, H. Mao, G. Cottrell, and J. McAuley, "Rezero is all you need: Fast convergence at large depth," in *Uncertainty in Artificial Intelligence*, pp. 1352–1361, PMLR, 2021.

[23] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser, "Universal transformers," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[24] pltrdy, "A full python implementation of the rouge metric (not a wrapper)," 2022. `https://github.com/pltrdy/rouge`.

[25] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Association for Computational Linguistics, Nov. 2019.

[26] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "DialogSum: A real-life scenario dialogue summarization dataset," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5062–5074, Association for Computational Linguistics, Aug. 2021.

[27] A. See, P. Liu, and C. Manning, "Get to the point: Summarization with pointer-generator networks," in *Association for Computational Linguistics*, 2017.

[28] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, (Vancouver, Canada), pp. 67–72, Association for Computational Linguistics, July 2017.

[29] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," 2018.

[30] J. Chen and D. Yang, "Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4106–4118, Association for Computational Linguistics, Nov. 2020.