A Finer-grain Universal Dialogue Semantic Structures based Model For Abstractive Dialogue Summarization

Yuejie Lei[1]* **Fujia Zheng**[1]* **Yuanmeng Yan**[1] **Keqing He**[2] **Weiran Xu**[1]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Meituan Inc., Beijing, China
{2019110830,fujia_zheng,yanyuanmeng,xuweiran}@bupt.edu.cn
hekeqing@meituan.com

Abstract

Although abstractive summarization models have achieved impressive results on document summarization tasks, their performance on dialogue modeling is much less satisfactory due to the crude and straight methods for dialogue encoding. To address this question, we propose a novel end-to-end Transformer-based model **FinDS** for abstractive dialogue summarization that leverages **Fin**er-grain universal **D**ialogue semantic **S**tructures to model dialogue and generates better summaries. Experiments on the SAMsum dataset show that FinDS outperforms various dialogue summarization approaches and achieves new state-of-the-art (SOTA) ROUGE results. Finally, we apply FinDS to a more complex scenario, showing the robustness of our model. We also release our source code[1]

1 Introduction

The field of abstractive summarization has recently seen impressive progress in document scenarios, while less attention has been paid to dialogue summarization. Previous research on dialogue summarization is based on successful document summarization models (Nallapati et al., 2016; See et al., 2017; Nikolov et al., 2018; Liu et al., 2018) which model the dialogue in a crude and straight manner. Taking the example of Table 1, the truth in this dialogue is that *Mark* lied to *Anne*, and that *passport* belongs to *Mark*, but the summary generated by Pointer-Generator Network (PGN) makes some factual error, which is denoted by (the comparison between) red and blue text in Table 1. Moreover, the predicted summary omits the critical information in the dialogue, as shown in the green text.

Such factual errors indicate that it is not suitable to transfer the document summarization model to the dialogues summarization model. This is mainly because, unlike the document, the dialogue serves the purpose of information exchange. It naturally contains more than one participants (Zhang et al., 2019) and multiple topics in many turns of utterances, (Xiao and Carenini, 2019) and hence the core information is distributed randomly. Besides, every speaker talks in a first-person perspective, which brings referral and coreference due to human language habit (Lei et al., 2021; Chen and Yang, 2021a). Straightly concatenating and sequentially understanding the dialogue might capture some erroneous and redundant semantic relationships between speakers and utterances. (Gao et al., 2020) Therefore, the dialogue summarization task is facing different challenges from document summarization:

Dialogue Scripts
Anne: You were right, he was lying to me :/. Irene: Oh no, what happened? Jane: Who? That Mark guy? Anne: Yeah, he told me he's 30, today I saw his passport - he's 40. Irene: You sure it's so important? Anne: He lied to me Irene.
Ground-Truth Summary: Mark lied to Anne about his age. Mark is 40.
Pointer-Generator Prediction: Anne was lying today. Anne saw her passport today.

Table 1: A dialogue example from SAMsum (Gliwa et al., 2019) with a ground-truth summary and a summary predicted by Pointer Generator Networks.

- Compared with the structural and logical writting style of document, dialogue is always unstructured, informal, and complex. Core information is randomly distributed in the whole dialogue. Sequential encoding is difficult to capture key information correctly.

- There are naturally multiple speakers in the dialogue, and how to capture the dependency between different speakers and utterances are

*The first two authors contributed equally. Weiran Xu is the corresponding author.

[1]https://github.com/apexmeister/FINDS

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1354–1364
November 7–11, 2021.

important for the understanding of dialogue.

Based on the understanding of these potential risks of dialogue scenario, recent work focuses on developing methods suitable for this dialogue summarization: Shang et al. (2018) developes an unsupervised multi-sentence compression algorithm, while Zhao et al. (2019) proposes a self-adaptive learning model to learn the segmentation strategy of utterances and topics. These methods are modified based on document summarization methods. Others also introduces some models designing specially for dialogue summarization: Liu et al. (2019b); Li et al. (2019); Zou et al. (2020) leverages the topic information that flows in the dialogue to help generate topic-aware summaries, Goo and Chen (2018); Liu et al. (2019a) manually annotates the dialogue to construct some prior structural knowledge which helps the model obtain a more informative and accurate context. Chen and Yang (2020) introduces two model-annotated dialogue structural views to help encode the utterances. Unfortunately, these jobs remain at a coarse level that can not correctly capture the relationships between speakers and utterances and topics. Some of them are time and labor-consuming or contain error superposition because of some handcrafted or model-based label.

Accordingly, we propose a novel end-to-end Transformer-based (Vaswani et al., 2017) model **FinDS** equipped with four **Fin**er-grain universal **D**ialogue semantic **S**tructures. To meet the first challenge, we propose **Inner Utterance semantic Structure (IUS)** and **Global Topic semantic Structure (GTS)** that helps the understanding of the dialogue from utterance-level to topic-level: The IUS only focuses on the information inside each utterance, as Figure 1(a) shows. The GTS connects utterances according to the topic that they are talking about, as Figure 1(b) shows. In response to the second challenge, we introduce **Inner Speaker semantic Structure (InSS)** and **Inter Speaker semantic Structure (ItSS)** to help model clarify the correct relationships between speakers and their topics: The InSS only focuses on the information from the same speaker, as Figure 1(c) shows. The ItSS interacts with the information from one speaker to other speakers except for himself, as Figure 1(d) shows. All these structures are constructed based on the universal characteristic of dialogue previously in an automatic method. With the help of these finer-grain universal dialogue semantic structures, our FinDS model performs effectively and robustly for dialogue summarization. Our contributions are three-fold:

(1) We develop a novel end-to-end Transformer-based model FinDS for abstractive dialogue summarization which models the dialogue with four pre-constructed universal dialogue semantic structures.

(2) We propose to construct four kinds of dialogue semantic structures in an automatic method to assist FinDS for better dialogue summarization: IUS focuses on the information inside each utterance; GTS connects utterances with the same topic; InSS focuses on the information from the same speaker; ItSS interacts the information from one speaker to other speakers except for himself.

(3) Extensive experiments on the SAMsum dataset present a comparable result compared with many strong baselines. The further analysis presents that FinDS performs robust and effective when the dialogue scenario getting complex.

2 Related Works

2.1 Document Summarization

Document summarization has received extensive attention in recent years, on which a lot of works have been done, and has achieved many successes. Rush et al. (2015) proposes an abstractive text summarization method by using sequence-to-sequence models originally. To address the out-of-vocabulary problem, See et al. (2017) introduces a pointer-generator network to allow the model to copy tokens from the source document. Paulus et al. (2017); Chen and Bansal (2018) achieves the goal of generating summarization by selecting appropriate content in the original document as summary sentences on the reinforcement learning framework. The performance of document summarization has also been further improved by using large-scale pre-trained language models proposed by Liu and Lapata (2019b); Raffel et al. (2019); Lewis et al. (2019), and Zhang et al. (2020) designs a new pre-training task for document summarization and achieved remarkable success.

2.2 Dialogue Summarization

While document summarization gains such great success, intensive research on dialogue summarization is also underway. Shang et al. (2018) introduces Multi-Sentences Compression Graph (MSCG) for meeting summarization, by choos-

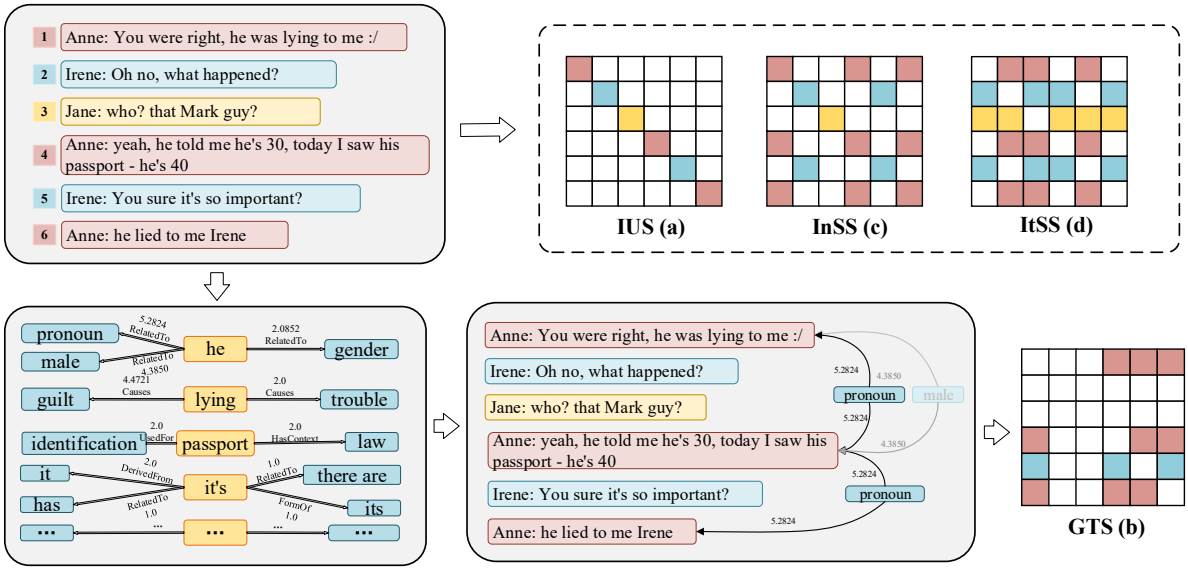

Figure 1: The construction of four universal dialogue semantic structures including IUS, GTS, INSS, and ITSS.

ing the correct path to compress sentences. Zhao et al. (2019); Zhu et al. (2020) proposes hierarchical models to obtain multi hierarchical-grain semantic representations to identify the turns, or utilizes role vectors to capture dialogue information. A few research also focus on utilizing external knowledge as features of the conversation. Goo and Chen (2018) captures the dialogue states changing during the dialogue by recording the dialogue acts. Other features like key points sequence (Liu et al., 2019a) and topics sequence (Liu et al., 2019b; Li et al., 2019) are also applied in dialogue summarization methods. However, such external knowledge is a human-annotated or model-based label which might be time and labor-consuming or include extra errors.

3 Method

To capture the core topics and build up a correct dependency between speakers and utterances at a finer-grain level, we propose to model the complex dependencies in the dialogue with the following procedures: (1) Constructing IUS, GTS, InSS, and ItSS in an automatic way (Section 3.2). (2) Encoding the dialogue by modifying self-attention processing with four dialogue semantic structures (Section 3.3). (3) The decoder receives the context from the encoder to predict a summary.

3.1 Motivation of Semantic Structures

To understand a dialogue, firstly, we must tell the model what each utterance is telling exactly. Because each speaker is talking sequentially and they are talking about a different topic sometimes. So we build up IUS to model the single utterance first. Once the model is able to understand the dialogue utterance by utterance, we can go further to teach the model to distinguish the topic of each speaker and the relationships between topics by building up InSS and ItSS. However, the topics flow in different speakers is sometimes facing interrupting and jumping. We need to construct closer relationships between topics and utterances. So we leverage the ConceptNet to build up the GTS to captures those relationships.

3.2 Semantic Structures Construction

This section describes the automatic constructing process of our four universal dialogue structures. Formally, for a given dialogue $D = \{w^0_{0,0}, w^0_{1,0}, ..., w^m_{l,n}\}$ with l words in total, we can figure out the speaker of each utterance according to the first word of each utterance, which is the speaker's name. So we denote $w^m_{l,n}$ as the l-th word in the n-th utterance from the m-th speaker. Then we take words as the Elementary Discourse Units (EDUs) to construct four dialogue semantic structures, G^{IUS}(Section 3.2.1), G^{GTS}(Section 3.2.2), G^{InSS}(Section 3.2.3), and G^{ItSS}(Section 3.2.4):

3.2.1 Inner Utterance Semantic Structures

Utterances in dialogue are not organized sequentially as documents due to the repetition and interruption, which also explains why core contents of the same speaker randomly distributed in the dialogue. And there is naturally more than one speaker in the dialogue, which makes it harder to

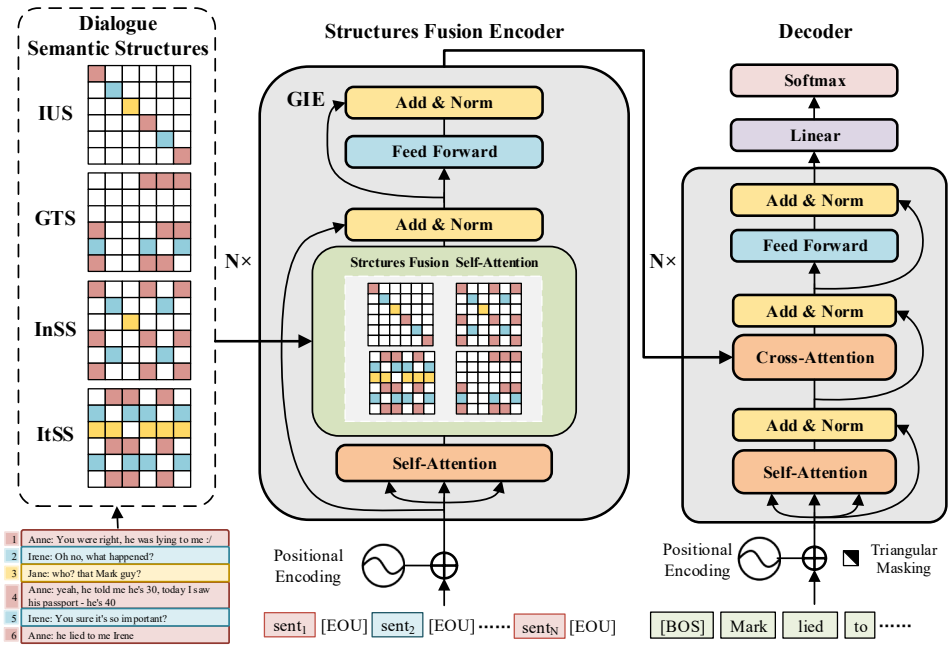

Figure 2: The overall architecture of FinDS is demonstrated and it is enhanced by four different universal dialogue semantic structures IUS, InSS, ItSS, and GTS.

capture the correct dependencies between speakers and their topics. For example, speaker **A** claims "*I like to eat apple*", while the other speaker **B** says "*I prefer banana*". In this situation, If we model the dialogue in a document summarization way, the attention might confuse speaker A's and speaker B's favorite fruits. Hence, before building up the relationships between speakers and their topics, we capture the local dependencies (Liu and Lapata, 2019a; Jin et al., 2020) inside each utterance by constructing Inner Utterance semantic Structure (IUS) as a graph $G^{IUS} = (D, E^{IUS})$, where D is the set of nodes that represent EDUs and E^{IUS} is the adjacent matrix that describes the connection of each node inside the same utterance as Figure 1(a) shows.

3.2.2 Global Topics Semantic Structures

As mentioned above, the topics of different speakers are distributed randomly in the dialogue. and the meaning of each utterance is not isolated (Qin et al., 2017). Therefore we follow (Feng et al., 2020) to build up the Global Topic semantic Structures (GTS) as a graph $G^{GTS} = (D, E^{GTS})$, where D is the set of nodes that represent EDUs and E^{GTS} is the adjacent matrix that describes the connection of each node according to the topic information. The topic information was collected by the commonsense knowledge graph ConceptNet (Speer and Havasi, 2012). For any subject s in the ConceptNet, it will have an object o and the relationship r between them with a confidential weight w. They will form a concept tuple like $c = (s, r, o, w)$. We input all words in the dialogue except real names and stopwords into ConceptNet and get concept tuple sets $C = \{c_{1,1}, c_{2,1}, ..., c_{i,k}, ..., c_{m,l}\}$, where $c_{i,k}$ represents a concept tuple obtained by using the word $w^j_{i,k}$ as the query for ConceptNet. And if any two concept tuples from different utterances $c_{i,j}$, $c_{p,q}$ has the same object o, we consider that the utterances they belong to are talking about the same topic. For example, if speaker **A** says "*I don't have his number*" while speaker **B** says "*I called him yesterday*", we can search the same object "*phone calling*" by matching the query word "*number* and *called* from the ConceptNet. Then we believe they are talking about the same topic "*phone calling*. According to such topic information, we can preconstruct the GTS for capturing the dependencies between topics and utterances as Figure 1(b) shows.

3.2.3 Inner Speaker Semantic Structures

Building up the dependencies between speakers and utterances (Murray et al., 2006) are of the same importance as the dependencies between utterances and topics. Because, a topic might have multiple participants, and the utterances from different speakers are usually unstructured and illogical because of the alternation and informality (Jackson and Moulinier, 2007) of utterances. So, we proposed to regroup the utterances and initially understand the main ideas of every speaker. We construct

the Inner Speaker semantic Structure (InSS) as a graph $G^{InSS} = (D, E^{InSS})$, where D is the set of nodes that represent EDUs and E^{InSS} is the adjacent matrix that describes the connection of each node from the same speaker as Figure 1(c) shows.

3.2.4 Inter Speaker Semantic Structures

The ideas of every speaker are not narrated isolatedly. Because dialogue carries the function of information exchange between people. To capture the dependencies between different speakers, we construct the Inter Speaker semantic Structures (ItSS) as a graph $G^{ItSS} = (D, E^{ItSS})$, where D is the set of nodes that represent EDUs and E^{ItSS} is the adjacent matrix that describes that every node connects to those nodes coming from other speakers as Figure 1(d) shows.

3.3 Encoder

Given a dialogue and its pre-constructed dialogue semantic structures, we propose a Structures Fusion Encoder (SFE) to obtain a structure-aware dialogue hidden representation by combining and interacting dialogue with four structures as Figure 2 shows.

3.3.1 Structures Fusion Encoder

We initialize our Structures Fusion Encoder $F_{SFE}(.)$ with a pre-trained encoder, i.e., BART-large (Lewis et al., 2019), and incorporate four structures into the self-attention calculation processing to encode all words $D = \{w^0_{0,0}, w^0_{1,0}, ..., w^m_{l,n}\}$ in dialogue into its hidden representation. To do so, we regard four pre-constructed structures as four mask matrixes M^{IUS},M^{GTS},M^{InSS},M^{ItSS} that have the same shape with the similarity matrix calculated by the Cartesian product by query and key. Then, we combine this similarity matrix and four mask matrix to influence the final attention weights and the hidden representation:

$$\{h^0_{0,0}, ..., h^m_{l,n}\} = F_{SFE}(D, M^{IUS}, M^{GTS}, M^{InSS}, M^{ItSS}) \quad (1)$$

Then, we introduce the Structures Fusion(SFA) Self-Attention to fuse the dialogue hidden representation with four structure mask matrixes:

Structures Fusion Self-Attention(SFA) The SFA module follows standard multi-head attention (MHA) to calculate four different attention results by superposing different structure mask matrixes directly to the original attention weights to modify original attention scores, and finally obtain a structure-aware hidden representation:

$$SFA = Concat\{heads^{M_0}, ..., heads^{M_j}\}W^L \quad (2)$$

$$heads^{M_j} = \{head^j_1, head^j_2, ..., head^j_i\} \quad (3)$$

$$M_j \in \{M^{IUS}, M^{GTS}, M^{InSS}, M^{ItSS}\} \quad (4)$$

$$head^j_i = \text{Softmax}\frac{(QW^Q_i)(KW^K_i)^T \cdot M_j}{\sqrt{d_K}} VW^V_i \quad (5)$$

where, W^Q, W^K, W^W, W^L are trainable parameters, Q,K,V are query, key, value in the self-attention calculation process.

SAMsum	Train	Validation	Test
Sizes	14732	818	819
Max.Speakers	4	12	9
Max.Turns	46	30	27
Avg.Speakers	2.40	2.39	2.36
Avg.Turns	11.17	10.83	11.25
Most.Speakers	2(10723)	2(605)	2(624)
Most.Turns	6(1309)	6(87)	6(86)

Table 2: Details of SAMsum

3.4 Decoding and Training

At decoding stage, FinDS follows standard transformer decoding approach. The decoder F_D receives the $l-1$ previous generated tokens $t_1, t_2, ..., t_{l-1}$ and predicts the l-th token with the finer-grain structure-aware context from SFE:

$$c_l = F_D(\{t_1, t_2, ..., t_{l-1}\}, F_{SFE}(D)) \quad (6)$$

$$P(\hat{t}_l | t_{<l}, c_l) = \text{Softmax}(W_p c_l) \quad (7)$$

where, W_p is a parameter to be learned.

And the training objective is to minimize the cross entropy loss:

$$L = -\sum \log P(\hat{t}_l | t_{<l}, c_l) \quad (8)$$

Additionally, we also apply the teacher forcing strategy: When training, the inputs of decoder are previous tokens from the ground truth summary. And ,at test time, the inputs are previous tokens predicted by the decoder.

4 Experiments

4.1 Experiment Settings

We evaluate our FinDS on a dialogue summarization dataset SAMsum (Gliwa et al., 2019)[2] which is written by langugage experts. Details of the dataset conditions are shown in Table 2. We load[3] the pre-trained sequence-to-sequence model "BART-large"[4] (Lewis et al., 2019) as our baseline, and modify the encoder as our Graph-Interactive Encoder. Normally, We use the Sharpening Interaction(SI) to involving four commonsense semantic graphs. Our model consists of 12 layers in total, 768 model dimensions, 12 heads. And fine-tune it with $3e^{-5}$ learning rate, 4 batch size, 512 max sequence length, and 15 max training epoch. All of our experiments are running on an Ubuntu 18.04 platform with two NVIDIA GeForce GTX 2080Ti GPUs. At testing stage, we follow (Chen and Yang, 2020) use the pltrdy-rouge[5] tool to calculate the ROUGE (Lin, 2004) scores. The baselines our model compares with are describing in the Appendix.

4.2 Experiment Baselines

- **Pointer Generator** (See et al., 2017): We input each utterances of the dialogue as division into the model, following (Gliwa et al., 2019). Through pointer mechanism, we generate the summary by generating or copying tokens from origin dialogue.

- **Fast Abs RL** (Chen and Bansal, 2018): This method first select important sentences from origin text and then rewrite these sentences to an abstractive pattern with sentence-level policy gradient methods. We also follow (Gliwa et al., 2019) to concatenate all utterances into one block.

- **Transformer** (Vaswani et al., 2017): This model utilizes the self-attention mechanism to parallelize the input text to generate summaries, and has achieved great results on the text summarization task. We use fully visible self-attention on this model, that is, do not make any changes to the original mask matrix.

- **LightConv** (Wu et al., 2019): To address the problem of the limited ability of self-attention to process long-span sentences, this model proposes a lightweight convolution module. We regard this model as one of our baseline models testing on SAMsum dataset.

- **DynamicConv** (Wu et al., 2019): Different from lightweight convolution module, the dynamic convolution module only changes in the weight parameters of the convolution. The weight parameters of the former are fixed on each feature map, and the weight of the latter needs to be the dot product based on the fixed value of the former and the feature point of the current position, and its outputs is used as the new wight.

- **Multi-View BART** (Chen and Yang, 2020): This is the first attempt on modeling dialogue with some dialogue structure information. Specifically, it introduces two extra relatively complicated dialogue-views to model the topics and stages in the dialogue and reach a State-Of-The-Art result on SAMsum so far.

- **S-BART** (Chen and Yang, 2021b): This work leverages the discourses relationships and speakers' actions to build up two graph explicitly. Combining them into the dialogue encoding and summary predicting procedure, which is the first job to modeling the dependencies between discourses and speakers.

4.3 Experiments Results

We evaluate FinDS on the SAMsum test set with ROUGE metrics (Lin and Och, 2004; Lin, 2004). As the Table 3 shows, Either PGN (See et al., 2017) or Transformer (Vaswani et al., 2017) performs disappointingly when facing dialogue summarization. The PGN gets the highest scores among those demonstrated traditional document summarization models. When testing on the pre-trained model BART-large, all scores improve averagely 10 points than those document models that prove the strong performance from pre-training. Based on BART, Chen and Yang (2020) introduces Multi-view BART that reached the previous SOTA ROUGE scores on the SAMsum dataset.

Compared with previous baselines, FinDS achieves new SOTA ROUGE results by 52.23

[2] https://www.tensorflow.org/datasets/catalog/samsum

[3] https://github.com/huggingface/transformers

[4] https://huggingface.co/facebook/bart-base

[5] https://github.com/pltrdy/rouge

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
Pointer Generator (See et al., 2017)	40.10	-	-	15.28	-	-	36.63	-	-
Fast Abs RL (Chen and Bansal, 2018)	40.96	-	-	17.18	-	-	39.05	-	-
Transformer (Vaswani et al., 2017)	37.27	-	-	10.76	-	-	32.73	-	-
LightConv (Wu et al., 2019)	33.19	-	-	11.14	-	-	30.34	-	-
DynamicConv (Wu et al., 2019)	33.79	-	-	11.79	-	-	30.41	-	-
BART (Lewis et al., 2019)	48.20	49.30	54.00	24.50	25.10	26.40	46.60	47.50	49.50
Multi-view BART (Chen and Yang, 2020)	49.30	51.10	52.20	25.60	26.50	**27.40**	47.70	49.30	49.90
S-BART (Chen and Yang, 2021b)	46.07	51.13	46.24	22.60	25.11	22.81	45.00	49.82	44.47
FinDS	**52.23***	**54.74***	**55.06***	**25.91***	**27.39***	27.11	**50.87***	**52.66***	**53.15***
FinDS w/o IUS	51.60	53.92	54.18	24.97	26.23	26.08	49.84	52.96	51.89
FinDS w/o GTS	50.57	54.07	52.54	24.78	26.61	25.70	49.04	51.63	50.61
FinDS w/o InSS	51.22	54.66	53.61	25.09	26.73	25.97	49.78	51.96	51.47
FinDS w/o ItSS	51.62	54.20	54.47	25.70	26.94	27.00	50.12	51.94	52.33

Table 3: ROUGE-1, ROUGE-2, ROUGE-L scores that different models perform on SAMsum test set. The numbers with * indicate the significant improvement over all baselines with $p < 0.05$ under t-test.

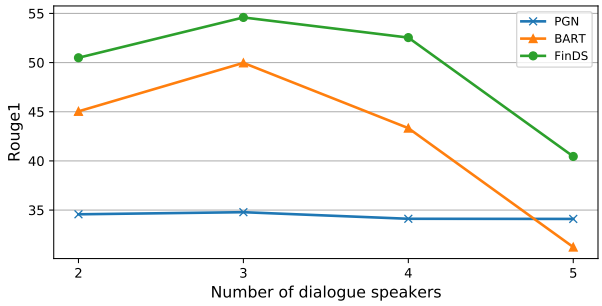

Figure 3: The changing ROUGE-1 F1 scores as speaker numbers increasing.

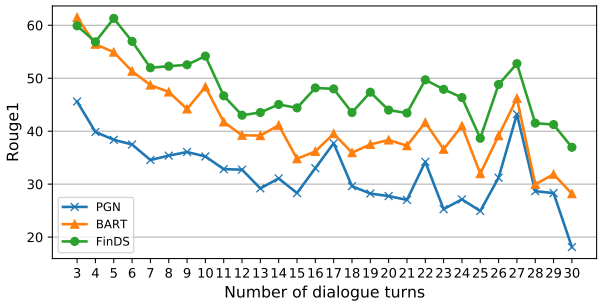

Figure 4: The changing ROUGE-1 F1 scores as dialogue turns increasing.

for ROUGE-1-F score, 25.91 for ROUGE-2-F, and 50.87 for ROUGE-L-F. Analyzing the results, FinDS gets nearly 3 points higher than the previous SOTA at 49.30 for ROUGE-1-F and 47.70 for ROUGE-L-F. The ROUGE-2-F score gains 0.3 points higher than the previous SOTA ROUGE-2-F result at 25.60. These results prove that our model can effectively capture those keywords as 1-grams that are distributed randomly in the dialogue, which is contributed by the IUS and GTS for constructing the local context dependencies inside the utterance and the global topic dependencies throughout the dialogue. And because the reference summaries are written by language experts manually that have high-level attractiveness. Therefore, it is difficult for content compression and synonymous rewriting, and neither Multi-view BART nor FinDS can achieve great improvement on the ROUGE-2-F score which represents the ability of a model to capture the core 2-grams in the dialogue for summarization.

4.4 Ablation Experiment

We also conducted ablation experiments on FinDS. In cases of removing any semantic structure, the ROUGE scores of FinDS are reduced but they are still higher than our baseline BART-large, as Table 3 shows. This phenomenon shows that each semantic structure contributes to the improvement of FinDS.

According to the results, it is obvious that GTS and InSS contribute more to the improvement of the model effect, especially GTS. The ROUGE scores of FinDS suffer the highest level reduction while removing InSS structure or GTS structure. And when IUS or ItSS is removed, FinDS suffers less damage on the performance. There are two intuitive explanations for this phenomenon. Firstly, GTS enhances the model's global understanding of dependencies between topics and utterances by introducing external knowledge. Then, by focusing on the dialogue content of each speaker, InSS allows the model to understand the characteristics and core topic of each speaker's discourses respectively, which brings more valuable information for dialogue summarization than capturing the information exchanges between different speakers by ItSS.

5 Analysis

5.1 Effect of Speaker Numbers

Figure 3 further shows the performance of FinDS when facing increasing speaker numbers from 2

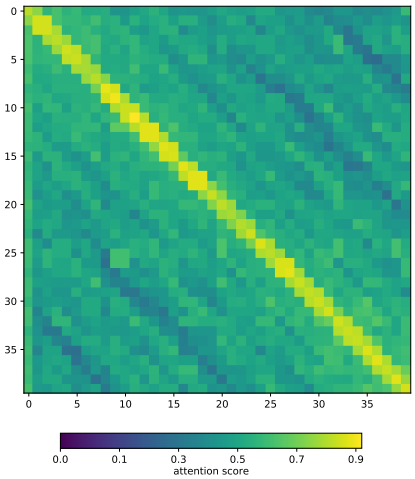

Figure 5: The attention heatmap of BART when encoding dialogue.

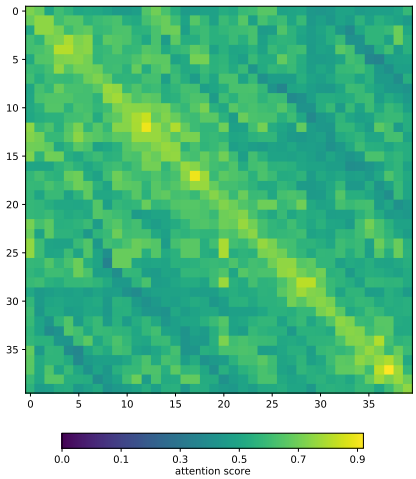

Figure 6: The attention heatmap of FinDS when encoding dialogue.

speakers to 5 speakers. We compare FinDS with the PGN model and the BART-large model on the ROUGE-1-F. With the increasing number of speakers, the performance of all models first have an upward trend and obtain a maximum score when reaching 3 speakers, then they all show a downward trend. The performance of the BART model drops sharply as the number of speakers increasing. When the speaker number reaches 5, the performance of the BART model is even worse than that of the PGN model, and FinDS outperforms others stably with an averagely score higher than 45. And the performance gap between FinDS and BART is getting larger when speakers increasing, which proves that FinDS still performs robustly and effectively when facing such a complex dialogue scenario. And it also testifies the InSS and ItSS are efficient for capture the information and modeling the dependencies inside and across the speakers.

5.2 Effect of Dialogue Turns

Figure 4 shows the performance of FinDS when facing increasing dialogue turns from 3 turns dialogue to 30 turns dialogue. Similar to Section 5.1, we compare FinDS with the PGN model and the BART model on the ROUGE-1-F. The performance of all models experiences an overall downward trend. When dialogue has few turns, BART and FinDS perform much better than PGN. As the speaker number increases, the performance of the PGN model and the BART model approach gradually and experience fluctuating downward. Though FinDS receive some damage on performance as well, it still outperforms enormously the other two all the time, which also proves the robustness and effectiveness of FinDS when facing a complex dialogue scenario. Furthermore, these phenomenons also evidence that IUS and GTS capture the strong dependencies inside and across the utterances even when dialogue has more than 20 turns.

5.3 Attention Heatmap Analysis

We randomly choose a dialogue sample and draw the attention heat-map when encoding it. The original BART-large pays more attention to the diagonal region which means there is an insufficiency for capturing the global information and remote semantic dependencies when modeling the dialogue as Figure 5 shows. This phenomenon directly evidences the fact that traditional document encoding approaches and the original self-attention are limited and implicit when modeling dialogue. On the contrary, FinDS incorporates four universal dialogue semantic structures to calibrate the direction of self-attention explicitly by capturing finer-grain and remote semantic dependencies as Figure 6 shows. Essentially, the model is forced to attend to the core contents purposefully and has more chance to learn useful information and relationships to help dialogue summarization.

5.4 Human Evaluation

To verify the improvement of FinDS beyond the ROUGE scores, we randomly choose some predicted summaries to conduct the human evaluation on 5 different model settings. We randomly invite 10 annotators to participate in the human evaluation and sample 10% examples generated by 5 model settings respectively. Given a prediction by FinDS with a specific model setting, a prediction by BART-large, a PGN result, and a ground-truth summary in each evaluation round, we provide all

Model	origin(10%)	w/o IUS(10%)	w/o GTS(10%)	w/o InSS(10%)	w/o ItSS(10%)
FinDS	+25%	+22%	+10%	+17%	+23%
BART	1.00	1.00	1.00	1.00	1.00
PGN	-52%	-55%	-53%	-50%	-49%

Table 4: The human evaluation result of the ablation FinDS performances and Vanilla PGN compared to the BART.

annotators the following guidelines:

(1) You will not be able to know the given three predictions are predicted by which models. They are all shuffled.

(2) Firstly, you should score all the summaries according to the **completeness** from 0 to 2. If a summary is incomplete, you should give it a 0 score which means this summary is unreadable and nonsensical.

(3) Secondly, you should score all the summaries according to the informativeness from 0 to 2. If a summary you score 0, it means the summary contains irrelevant and unimportant messages from the dialogue compared to the ground-truth summary.

(4) Thirdly, you should score all the summaries according to the information correctness from 0 to 2. If a summary gets a 0 score, it means that the information in the summary does not conform to the basic facts in the dialogue, though the information might not be relevant and important compared to the ground-truth summary.

With the pre-defined rules above, there are 3 scores range from 0 to 2 that a summary can get with the consideration of *completeness*, *informativeness*, *information correctness*. And We calculate the average score denotes the quality of the summaries from the same model setting and normalize them by the results of the BART-large model. Therefore we use the scores of the BART-large model as a baseline to evaluate the differences between it and other candidates as Table 4 shows. According to the results, we find that our best model's human-evaluating performance is 25% higher than the baseline. When removing any semantic graph, all scores reduce slightly, but still higher than the baseline. This phenomenon shows that all of our semantic graphs contribute. The removal of GTS has the greatest impact on FinDS which leads to a 15% human-evaluating performance dropping, as it introduces the global relationship of utterances into FinDS as external knowledge. Removing InSS also causes a big blow to the human-evaluating performance of the model with 8% performance dropping. And the overall human-evaluating performance of PGN is disastrously 50% lower than the BART-large. These phenomenons are conforming to the phenomenons of ablation experiments that demonstrate different extents our dialogue semantic structures contribute to dialogue summarization.

Example 1
Frank: Son, will you come home this weekend? Son: not sure yet. Something happened? Frank: Of course not . Your mother is miss you. Son: I miss her too. Frank: So will you com? Son: I will try. Frank: Good, I will tell your mother that you will come Son: oh, dad.. ok I will come.
Ground Truth Son is coming to see his parents this weekend.
PGN Pred. Son will come to Frank's mother's home.
FinDS Pred. Son will try to come home this weekend.
Example 2
Anne: You were right, he was lying to me :/. Irene: Oh no, what happened? Jane: Who? That Mark guy? Anne: Yeah, he told me he's 30, today I saw his passport - he's 40. Irene: You sure it's so important? Anne: He lied to me Irene.
Ground Truth Mark lied to Anne about his age. Mark is 40.
PGN Pred. Anne was lying today. Anne saw her passport today.
FinDS Pred. Mark lied to Anne about being 30 .Anne saw his passport today .

Table 5: Two cases to compare between the predictions from FinDS, PGN, and the Ground Truth, red words means wrong massages while green means right content and blue parts highlight the core content.

5.5 Case Study

We also present a case study with two dialogue and their relative summaries. Comparing to the traditional document summarization model, our FinDS can achieve improvement beyond the ROUGE scores, which also shows the predicted summaries are more informative and more correct. FinDS can capture all core contents in the dialogue and turn them into the right message in the summaries while the PGN is failed.

6 Conclusion

In this paper, we develop a novel end-to-end Transformer-based model FinDS for abstractive dialogue summarization that leverages finer-grain universal dialogue semantic structures to model dialogue and generates better summaries. Experiments have shown FinDS achieves new SOTA results on the ROUGE metrics. More importantly, FinDS proves its robustness and effectiveness for every structure in the complex dialogue scenario.

Acknowledgments

We thank all anonymous reviewers for their helpful comments and suggestions. This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, DOCOMO Beijing Communications Laboratories Co., Ltd, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701.

Ethical Considerations

We proposed a novel architecture to address the dialogue summarization tasks. Without extra manual annotations, our method builds up four finer-grain semantic structures from the dialogue to help model. Besides time and labor-saving, this modeling style is full of potential and is relatively not specific to a particular model or application, which will bring inspiration to later researchers to construct useful dialogue structures information and promote the development of current dialogue summarization models or other dialogue-relative tasks.

References

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*.

Jiaao Chen and Diyi Yang. 2021a. Structure-aware abstractive conversation summarization via discourse and action graphs. *ArXiv*, abs/2104.08400.

Jiaao Chen and Diyi Yang. 2021b. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.

Shen Gao, X. Chen, Z. Ren, Dongyan Zhao, and Rui Yan. 2020. From standard summarization to new tasks and beyond: Summarization with manifold information. *ArXiv*, abs/2005.04684.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

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. IEEE.

Peter Jackson and Isabelle Moulinier. 2007. *Natural language processing for online applications: Text retrieval, extraction and categorization*, volume 5. John Benjamins Publishing.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254.

Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7823–7827.

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. *arXiv preprint arXiv:1910.13461*.

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.

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

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

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

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna D Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 367–374.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. 2018. Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Kechen Qin, Lu Wang, and Joseph Kim. 2017. Joint modeling of content and discourse relations in dialogues. *arXiv preprint arXiv:1705.05039*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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. *arXiv preprint arXiv:1805.05271*.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. *arXiv preprint arXiv:1909.08089*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating character descriptions for automatic summarization of fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7476–7483.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. End-to-end abstractive summarization for meetings. *arXiv preprint arXiv:2004.02016*.

Yicheng Zou, Lujun Zhao, Yangyang Kang, J. Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2020. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. *ArXiv*, abs/2012.07311.