Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056877

ARTICLE

Multi-Head Encoder Shared Model Integrating Intent and Emotion for Dialogue Summarization

Xinlai Xing, Junliang Chen*, Xiaochuan Zhang, Shuran Zhou and Runqing Zhang

School of Artificial Intelligence, Chongqing University of Technology, Chongqing, 401135, China

*Corresponding Author: Junliang Chen. Email: youchen0124@stu.cqut.edu.cn

Received: 01 August 2024 Accepted: 13 November 2024 Published: 17 February 2025

ABSTRACT

In task-oriented dialogue systems, intent, emotion, and actions are crucial elements of user activity. Analyzing the relationships among these elements to control and manage task-oriented dialogue systems is a challenging task. However, previous work has primarily focused on the independent recognition of user intent and emotion, making it difficult to simultaneously track both aspects in the dialogue tracking module and to effectively utilize user emotions in subsequent dialogue strategies. We propose a Multi-Head Encoder Shared Model (MESM) that dynamically integrates features from emotion and intent encoders through a feature fusioner. Addressing the scarcity of datasets containing both emotion and intent labels, we designed a multi-dataset learning approach enabling the model to generate dialogue summaries encompassing both user intent and emotion. Experiments conducted on the MultiWoZ and MELD datasets demonstrate that our model effectively captures user intent and emotion, achieving extremely competitive results in dialogue state tracking tasks.

KEYWORDS

Dialogue summaries; dialogue state tracking; emotion recognition; task-oriented dialogue system; pre-trained language model

1 Introduction

Task-oriented dialogue systems (TOD) are a significant type of dialogue system designed to fulfill specific user goals, such as hotel reservations and route planning [1–3]. In recent years, the application of deep learning in TOD research has achieved remarkable progress, attracting considerable interest from both academia and industry. Dialogue state tracking (DST) is a fundamental component of such task-oriented dialogue systems [4,5]. The primary goal of DST is to understand the needs expressed by the user during the dialogue based on a given schema or ontology.

However, most dialogue state tracking (DST) tasks focus primarily on tracking user intent, often neglecting the emotional states of users. The user's emotional state can provide valuable feedback to enhance the component's ability to accurately recognize user intent. As is well-known, emotions are psychological behaviors generated by the joint stimulation of internal and external worlds. In other words, emotions are determined by one's intentions and external actions. Positive emotions are often

displayed when actions satisfy intentions. This emotional awareness can help dialogue AI generate responses that are more emotionally and semantically appropriate [6].

In task-oriented dialogue systems, accurately extracting the user's departure location, destination, and departure time, while simultaneously perceiving changes in the user's emotional state and adjusting the dialogue strategy in real-time, is crucial for creating a positive user experience, as illustrated in Fig. 1.

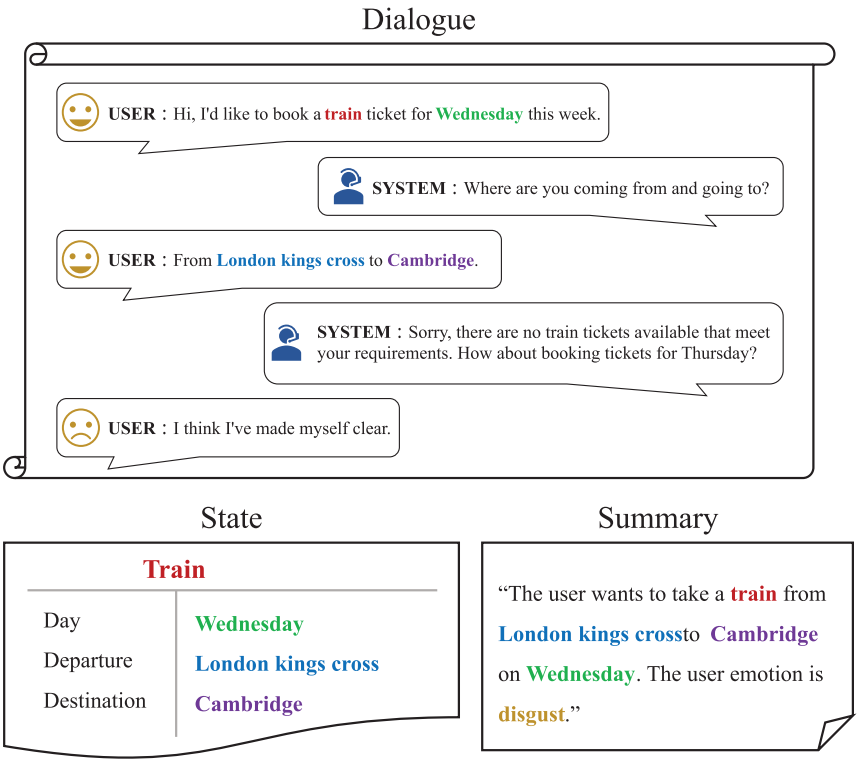

Figure 1: An example of a dialogue in the train domain and the summary created based on the state and user emotion

Nevertheless, the acquisition of turn-level dialogue state annotations is highly resource-intensive, necessitating considerable effort from domain experts in terms of design and mediation. Typically, this annotation process employs the Wizard-of-Oz (WoZ) technique [7], wherein two individuals engage in dialogue and annotate each turn's state. In a notable study [8], researchers utilized crowdsourcing to compile MultiWoZ 2.0, creating one of the most extensive publicly available multi-domain task-oriented dialogue datasets. Many works on task-oriented dialogue systems (TOD) are based on MultiWoZ 2.0 [8] and MultiWoZ 2.1 [9].

So far, MultiWoZ is undoubtedly the most suitable dataset for guiding models in intent recognition, but it does not contain emotion labels. However, recognizing emotions is equally important in task-oriented dialogues. To address this issue, Reference [10] conducted extensive manual emotion annotations based on MultiWoZ, creating the EmoWOZ dataset. Unfortunately, the proportion of neutral emotion labels in EmoWOZ remains high, which is insufficient to fully guide models in dialogue emotion recognition.

To tackle the aforementioned issues, we propose integrating dialogue intent and emotion recognition, and using multiple datasets to guide model training. We replace dialogue state with dialogue summary that incorporate user intent and emotional information, as shown in Fig. 1. In the original Dialogue State Tracking (DST), the dialogue in Fig. 1 consists of the following three states: train-day: Wednesday, train-departure: London Kings Cross, and train-destination: Cambridge. We utilize the dialogue summary: *"The user wants to take a train from London Kings Cross to Cambridge on Wednesday. The user's emotion is disgust."* to replace the original dialogue states. These summaries are generated using heuristic rules that consider dialogue state and the user's current emotion. This approach offers two main benefits: firstly, by replacing the dialogue state tracking task with the natural language generation task of dialogue summarization, and by using natural language sentence templates instead of formatted dialogue states, dialogue state tracking can more easily handle unexpected slot value information and be extended to other unknown domains. Second, we also replace user emotion classification recognition with user emotion summary generation, making the overall model's dialogue summary generation more unified and facilitating feature integration. We continue to explore the limits of dialogue summary representation and expansion on the DS2 [11] basis. Since it has been validated in DS2 that guiding a dialogue summary model to generate dialogue states is feasible. Our research questions are as follow: 1. Is it feasible to use natural language sentence templates to guide the summarization model in generating emotional summaries. 2. Can the integration of features from emotional summaries and intent summaries enhance the quality of dialogue summary generation.

Therefore, we propose a Multi-Head Encoder Shared Model (MESM) and design a multi-dataset learning method. This model can learn features from different datasets through shared encoders and generate dialogue summaries that include both user intent and emotion. Additionally, this process is highly extensible, allowing for the easy expansion of other types of summarization capabilities. Finally, experiments on the MultiWoZ and MELD datasets demonstrate that our model can generate summaries that include both user intent and emotion, achieving improvements over the DS2 baseline, thereby proving the effectiveness of our approach. The main contributions of this paper are as follows:

- We propose the use of a dialogue summary model to generate emotion summaries, providing a new approach for dialogue emotion recognition tasks.
- We designed a Multi-Head Encoder Shared Model with a feature fusion mechanism that extracts and fits features from emotion encoders and intent encoders, enabling the model to generate dialogue summaries that encompass both user intent and emotion.
- We demonstrate the effectiveness of emotion summary generation. Furthermore, when implemented within generative models with reasoning capabilities, our approach achieved extremely competitive results on the MultiWoZ dataset.

2 Related Work

Dialogue State Tracking (DST) is a recognized component within task-oriented dialogue systems [12,13]. In recent years, many works [14] have addressed the DST problem by leveraging large pre-trained language models, such as BERT [15] and T5 [16]. Large pre-trained language models enhance context understanding and provide stronger generalization capabilities, enabling them to handle diverse input data. However, they also come with increased parameter counts and longer training times. Modern advanced methods generally involve fine-tuning pre-trained language models with extensive annotated datasets [17–20]. To reduce reliance on large amounts of costly labeled training

data, many researchers have made significant explorations in Few-Shot Dialogue State Tracking [21–24]. SM2 [25], by combining meta-learning with candidate pool retrieval, has demonstrated excellent performance in Few-Shot Dialogue State Tracking. At the same time, the QA-style prompts in TransferQA [26] introduce a high time complexity for slot value decoding. To address this issue, DS2 proposes replacing traditional dialogue state tracking tasks with dia-logue summarization, achieving promising results. DualLoRA [27] achieves strong results by using two distinct Low-Rank Adaptation (LoRA) components to handle dialogue context and optimize prompts. However, the aforementioned work still overlooks the importance of tracking user emotions in multi-turn dialogues, which is just as crucial as tracking user intent. We conducted Few-Shot training in a multi-domain setup with 1%, 5%, 10%, and 100% of the data. In Section 5.1, we briefly introduce our MESM and compare it with existing models.

Emotion Recognition in Conversation, a task that has been gaining increasing attention in the NLP field [28–31], has recently seen numerous encoder-based approaches [32,33]. However, the ERC task remains underexplored, which motivates us to reframe it as a unified gen-erative paradigm, specifically by using summarization for dialogue emotion recognition. In Section 5.2, we briefly describe and compare MESM with classical models.

Dialogue Summarization has been the focus of a growing number of researchers, encompassing both datasets [34–36] and models [37–39]. In DS2, the authors proposed representing dialogue states as dialogue summaries, which offers significant flexibility and scalability. Building on this, Reference [40] integrated summarization and response generation through a shared encoder, achieving strong results in task-oriented dialogue systems. We further explored the scalability of dialogue summarization and applied it to dialogue state tracking and dialogue emotion recognition, enabling subsequent feature integration.

3 Methods

3.1 Background

In the Dialogue State Tracking (DST) task, each data point comprises a task-oriented dialogue d and and a sequence of dialogue states $\{s_t\}_{t=1}^{n}$, where t denotes the current turn index and n is the total number of turns. Here, s_t represents the dialogue state at the t-th turn. The dialogue state consists of slot-value pairs, where k represents the slot name and v denotes the corresponding slot value, defined as:

$$s_t = \{(k_1, v_1), (k_2, v_2), (k_3, v_3), ..., (k_m, v_m)\} \tag{1}$$

where the set of all possible slots k_i within the domain is predefined. For instance, Table 1 lists three slots in the "train" domain of MultiWoZ: "train-day," "train-departure," and "train-destination".

Table 1: An example of the converter between state and summary

Domain	Slot name	Slot template	Slot value
Emotion	Emotion-user	is _	Disgust
	Sentence Prefix	The user emotion	
Train	Train-day	on _	Wednesday
	Train-departure	from _	London kings cross

(Continued)

Table 1 (continued)

Domain	Slot name	Slot template	Slot value
	Train-destination	to _	Cambridge
	Sentence Prefix	The user wants to take a train	
Complete general summary		"The user wants to take a train from *London kings cross* to *Cambridge* on *Wednesday*. And the user emotion is *disgust*."	

Using this setup, DST is a task that predicts s_t given the truncated dialogue history $d_1 \sim d_t$ as input at each time t. For convenience, we simply use dh_t to represent the dialogue history $d_1 \sim d_t$. In our MESM model, the input is the current dialogue history dh_t, and the output is a dialogue summary that includes the current non-empty dialogue state. Naturally, the output dialogue summary also contains emotion summaries, dialogue summaries that include the user's emotion state.

3.2 Multi-Head Encoder Shared Model (MESM)

Incorporating user emotion recognition alongside intent detection can facilitate smoother multi-turn interactions in task-oriented dialogue systems, thereby enhancing the user experience. In this section, we describe the overall framework of our proposed Multi-Head Encoder Shared Model (MESM) for jointly generating intent and emotion summaries, as shown in Fig. 2. Our approach consists of three main steps: the first step is emotion summarization training, which primarily involves training the emotion encoder and emotion decoder; the second step is intent summarization training, which primarily involves training the intent encoder, feature fusion mechanism, and intent decoder; the third step outlines the inference process when the model is used. In our method, we treat emotion as a type of slot-value pair in dialogue state tracking, following the approach proposed in DS2, which uses dialogue summaries instead of traditional dialogue state tracking. Therefore, to convert slot-value pair labels into summaries suitable for training and to evaluate the effectiveness of the generated summaries in capturing user emotion and intent, our components also include a dialogue summarizer (φ) and a dialogue state extractor (δ).

First, we propose converting the dialogue emotion recognition task into a emotion summarization task. In Step 1, we train the emotion encoder and emotion decoder to generate emotion summaries. For a given dialogue history dh_t, we first use the dialogue summarizer to generate a emotion summary $z_e = \varphi(s)$, as shown in the emotion section of Table 1. We then train the emotion encoder and emotion decoder to predict z_e, calculating the loss between z_e and the model's predicted output z'_e during training.

Next, in the intent training step, we introduce a feature fusion mechanism (β) to integrate the output features of the emotion encoder and the intent encoder. In Step 2, we train the intent encoder, feature fusion mechanism, and intent decoder to generate intent summaries. Similar to emotion training, for a given dialogue dh_t, we first use the dialogue summarizer to generate an intent summary $z_i = \varphi(s)$, as shown in the intent section of Table 1. In the model, we input the outputs of the emotion

encoder T_e and the intent encoder T_i into the feature fusion mechanism, producing the output:

$$T_f = \beta(T_e, T_i) \tag{2}$$

which is then input into the intent decoder. In this step, we train the intent encoder, feature fusion mechanism, and intent decoder to predict z_i, calculating the cross-entropy loss between the summary z_i generated by the converter and the predicted summary z_i' from the model.

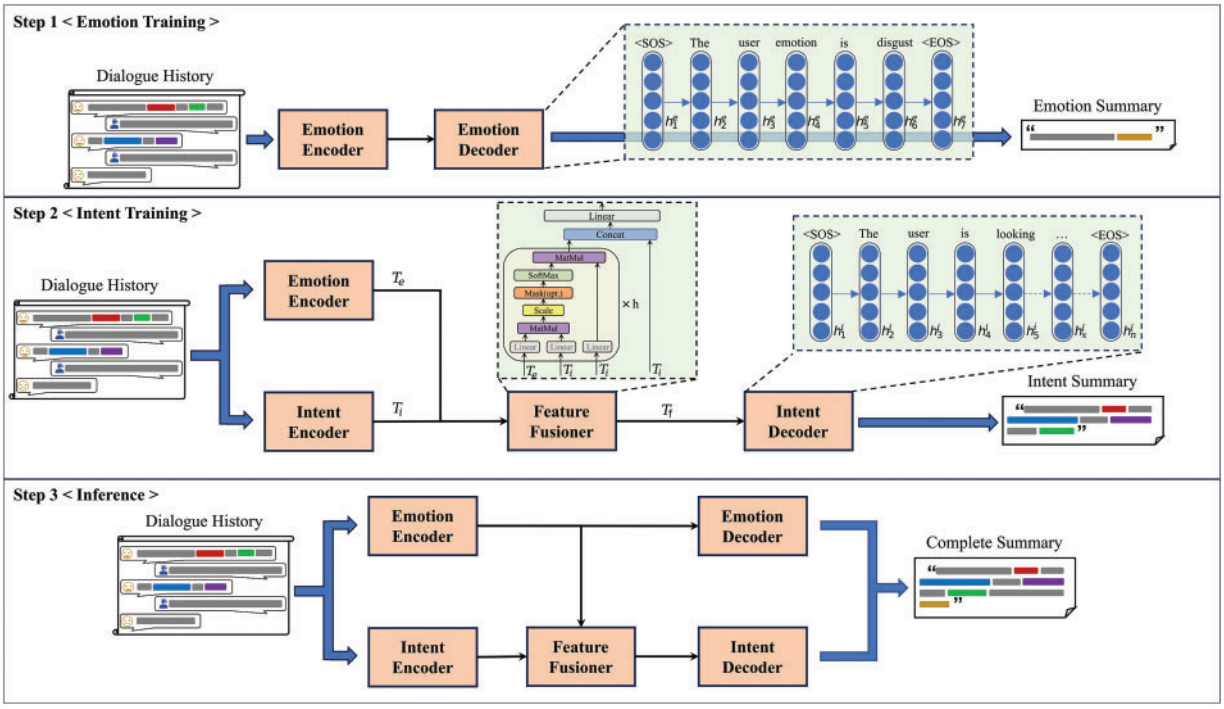

Figure 2: Overview of our Multi-Head Encoder Shared Model (MESM). In Section 3.2, we provide a detailed explanation of our model's training and inference process, dividing it into three steps

Finally, we produce a unified summary containing both emotion and intent through a unified inference process, as illustrated in Step 3 of Fig. 2. To evaluate the performance of the predicted summaries, during testing, we need to use the dialogue state extractor (δ) to parse slot values from the predicted summaries. For each dialogue state s_t, $\delta(\varphi(s_t)) = s_t$. If the predicted summary $z' = \text{MESM}(dh_t)$ is similar to the generated summary $z = \varphi(s_t)$, the testing process follows these steps:

$$\delta\big(\text{MESM}(dh_t)\big) = \delta(z') = s_t' = \delta(\text{z}) = \delta(\varphi(s_t)) = s_t \tag{3}$$

Here, $\delta(\text{MESM}(dh_t))$ functions as a traditional DST model, with the input being the current dialogue history dh_t and the output being the predicted dialogue state s_t'. Generally, integrating features from different tasks into deep learning models and making them effective is challenging. Therefore, we hypothesize that the feature fusioner (β) is the key factor in the performance of our model.

3.3 Converter between State and Summary

Building on the foundation established by DS2, we have enhanced the dialogue summarizer (φ) and the dialogue state extractor (δ). In our approach, emotions are treated as a domain within the

dialogue, similar to other domains in MultiWoZ, and emotional labels are also converted into slot-value pairs. In each domain of the dialogue, the dialogue summarizer (φ) can convert the dialogue state s_t into a dialogue summary z. During testing, the dialogue state extractor (δ) can convert the predicted dialogue summary z' back into the predicted dialogue state s'_t. The dialogue summarizer (φ) and the dialogue state extractor (δ) form a pair of inverse converters, where φ is the left inverse of δ.

For a given set of dialogue states s_t, assume there are m slots in the current dialogue state, $k_1, k_2, \ldots, k_m$, each containing corresponding slot values $v_1, v_2, ..., v_m$. Slot-value pairs where v=none are removed, resulting in the slot-value pairs $s_t = \{(k_1, v_1), (k_2, v_2), ..., (k_e, v_e)\}$, where $0 \leq e \leq m$. Each slot has a corresponding slot template. Different domain slot-value pairs are converted sequentially and then combined. For instance, in Table 1, the emotion domain conversion involves combining the slot name "emotion-user" with the slot template "is _", resulting in the phrase "is disgust". As this is a single slot, the phrase can be directly added to the domain's sentence prefix, resulting in the emotion summary z_e:

"The user emotion is disgust."

For domains with multiple slots, each slot value $v_1, v_2, \ldots, v_i$ is combined with its corresponding slot template to create a set of phrases $p_1, p_2, \ldots, p_i$. These phrases are then concatenated with the domain prefix to obtain the intent summary. For example, in the train domain shown in Table 1, since it is a single domain, the train domain summary is the intent summary z_i:

"The user wants to take a train from London kings cross to Cambridge on Wednesday."

Combining the emotion summary z_e and the intent summary z_i, we get the overall summary z_c:

"The user wants to take a train from London kings cross to Cambridge on Wednesday. And the user emotion is disgust."

As mentioned earlier, the dialogue summarizer (φ) and the dialogue state extractor (δ) form a pair of inverse converters. The dialogue state extractor (δ) extracts the dialogue state from the generated summary, which is almost the reverse process of the dialogue summarizer (φ). First, the overall summary is decomposed into sentences for different domains, and then the slot values are extracted from each domain sentence.

3.4 Feature Fusioner

The design of the Feature Fusioner (β) aims to extract features from the output of the emotion encoder that are effective for intent recognition and to fuse these with the output of the intent encoder. This enables the intent encoder to receive a richer set of features. In designing the Feature Fusioner, we employed the Multi-head Attention mechanism to capture the relationships between the outputs of the emotion encoder and the intent encoder. Multi-head Attention is a powerful attention mechanism that can simultaneously focus on different parts of the input sequence, helping to capture richer semantic information.

Specifically, in our model, let the output of the emotion encoder be $T_e = \{e_1, e_2, \ldots, e_n\}$, where n denotes the number of features in the emotion encoder output, and the output of the intent encoder be $T_i = \{i_1, i_2, \ldots, i_m\}$, where m denotes the number of features in the intent encoder output. We use Multi-head Attention to fuse and relate these two feature sequences. The computation process of Multi-head Attention can be expressed as:

$$MultiheadAttention(Q, K, V) = Concat(head_1, ..., head_h) \times W^o \tag{4}$$

where Q, K and V represent the query, key, and value matrices, $head_i$ represents the attention computation result of the i-th head, h denotes the number of heads, and W^o denotes the output weight matrix. In our context, the output of the emotion encoder T_e serves as the query Q, while the output of the intent encoder T_i serves as the key K and value V.

Through this method, Multi-head Attention can automatically learn and extract the correlations between the outputs of the emotion encoder and the intent encoder, enhancing the feature representation between them. After extracting features T_a through Multi-head Attention, we concatenate the original features T_i from the intent encoder with the extracted features T_a along the last dimension and apply a residual connection layer L to obtain the final fused features. The process of the residual connection can be represented as:

$$T_f = L(Concat(T_i, T_a)) \tag{5}$$

where T_f denotes the final feature representation, i.e., the features after fusion and residual connection. This design aims to reinforce the information in the original features while combining the features extracted through Multi-head Attention to achieve better feature representation. Finally, the features T_f are passed to the intent decoder for the final intent summary generation. This process can be mathematically expressed as:

$$IntentDecoder(T_f) = Output_i \tag{6}$$

The design of feature fusion and residual connection is intended to enhance the expressiveness of the features and the accuracy of intent recognition by effectively combining feature information from different sources. We illustrate the computation process of the Feature Fusioner in Algorithm 1 using Python-style pseudo-code.

Algorithm 1: A python-style pseudo-code for Feature Fusioner

Require: The array of shape [L, B, dim] (sequence length, batch size, hidden dimension), T_e & T_i; The dimension of model, h; The num of heads, n;
Ensure: The array of shape [L, B, dim], T_f;

Initialize attention as MultiheadAttention with $embed_{dim} = h$ and $num_{heads} = n$;
Initialize w as a learnable parameter with random values;
Initialize L as a linear layer of shape $[h * 2, h]$;
$T_e = permute(T_e, (1, 0, 2))$;
$T_i = permute(T_i, (1, 0, 2))$;
$T_a, \_ = attention(T_e, T_i, T_i)$;
$T_a = T_a * sigmoid(w)$;
$T_f = L(concatenate(T_i,\ T_a,\ dim = -1))$;
$T_f = permute(T_f, (1, 0, 2))$
return T_f;

4 Experiments

4.1 Dataset

MultiWoZ is an extensive English dataset for multi-domain task-oriented dialogues, covering seven distinct domains. However, akin to the approach in [4], we focus on five domains: attraction, taxi, train, hotel and restaurant. Table 2 lists the number of dialogues in each of the five

domains within the MultiWoZ 2.1 training set. Our model is evaluated on both MultiWoZ 2.0 and MultiWoZ 2.1.

Table 2: The number of dialogues in the MultiWoZ 2.1 training dataset

MultiWoZ 2.1	Single-domain	Multi-domain
Attraction	127	2717
Taxi	325	1654
Train	275	3103
Hotel	513	3381
Restaurant	1197	3813

MELD [31] dataset is a multimodal dataset derived from the popular TV show "Friends," comprising over 1400 dialogues and more than 13,000 utterances, each annotated with emotion and emotion labels. In our scenario, we only use the textual data. Table 3 presents the distribution of emotions in the MELD dataset and the number of instances in each category. For our purposes, we primarily focus on the emotion labels.

Table 3: Emotion distribution in MELD

Categories	Train	Dev	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

4.2 Evaluation

Intent Summary Generation

In our experiments, we use intent summary generation as a replacement for Dialogue State Tracking (DST), where the generated intent summaries can be easily converted into dialogue states within DST. Therefore, the primary performance metric for our intent summary generation experiments is the Joint Goal Accuracy (JGA) in DST, as shown in Eq. (7). In each turn of dialogue, we use a dialogue state extractor (δ) to extract the dialogue state from the model-generated intent summaries. The dialogue state is considered correct only if it matches the gold standard label set exactly [5]. We report the JGA according to the evaluation settings detailed in Section 4.4, as described in [4].

$$JGA = \frac{TP + TN}{P + N} = \frac{\sum_1^n T_{turn}}{\sum_1^n turn} \tag{7}$$

Emotion Summary Generation

Similar to intent summary generation, emotion labels can also be easily extracted from the generated emotion summaries. As shown in Eq. (8), we use weighted-average (w-avg) to evaluate the model's accuracy across each emotion category.

$$ACC_{weighted} = \frac{\sum_{i=1}^{N} n_i * ACC_i}{\sum_{i=1}^{N} n_i} \tag{8}$$

4.3 Model

Recently, many large parameter models have demonstrated outstanding performance and exceptional scalability in various natural language processing (NLP) tasks. However, in multi-turn dialogue systems, the advantages of these large language models primarily lie in their ability to handle open-domain dialogue and their strong natural language generation (NLG) capabilities. Due to their extensive parameter sizes, many studies tend to utilize large language models as a whole, employing prompt-based fine-tuning methods for implementation, as seen in [41].

Therefore, for subsequent deployment in practical environments, we evaluated two pre-trained language models, BART-large and T5-large, as the encoder and decoder within the Multi-Head Encoder Shared Model (MESM) framework after careful selection. Unfortunately, the BART-large model did not perform well in emotion summary generation. Consequently, we primarily utilized the T5-large model. The pretrained weights for T5-large were sourced from [11], from which we extracted the encoder and decoder to incorporate into the MESM model.

4.4 Experimental Settings

Intent Summary Generation Settings

In our experiments, we primarily explored the Multi-Domain (MD) setting as described in [42]. For the MD experiments, the model was trained using all domains, with each slot value contributing to the overall summary and evaluation. Performance was assessed by measuring both the per-domain JGA and the overall JGA.

We conducted Few-Shot DST experiments within the MD scenario, sampling 1%, 5%, 10%, and 100% of the training data to fine-tune the model. The full dev and test datasets were used for evaluation across all settings. As outlined in Section 4.1, each setting was run on MultiWoZ 2.0 and 2.1.

Emotion Summary Generation Settings

In our experiments, we focused on recognizing various emotion labels within the MELD dataset. We preprocessed the data from MELD into multi-turn dialogues based on Dialogue_ID. For each dialogue, we input the historical dialogue context to generate emotion summaries.

Other Settings

For both intent summary generation and emotion summary generation, we used the Adam optimizer with a learning rate of 5e-5, a batch size of 2, and trained for 50 epochs. We implemented and ran our experiments using Python 3.8.18 and PyTorch 1.9.1, leveraging CUDA 11.1 for accelerated computation. The model was trained on the training dataset, with early stopping criteria based on validation set performance. After training, final evaluations were performed on the test dataset.

4.5 Baselines

Intent Summary Generation

We conducted experiments on MultiWoZ 2.0 and MultiWoZ 2.1, as most of the baseline evaluations were performed on these datasets.

DS2 [11] is an innovative dialogue system model that proposes using dialogue summaries instead of dialogue states, aiming to simultaneously handle intent recognition, slot filling, and dialogue generation tasks.

TRADE [4] is a model based on multi-turn dialogues aimed at intent recognition and slot filling in task-oriented dialogues. It utilizes a copy mechanism and slot-domain embeddings to achieve transferability and enhance dialogue system performance.

MinTL [20] is a lightweight Transformer model specifically designed for intent recognition and slot filling tasks in dialogue systems. It employs multitask learning by minimizing the distance between the target task and the language model task.

SOLOIST [43] is a self-supervised learning-based dialogue system model that leverages local information from the dialogue history for intent recognition and slot filling. It uses a method called SOLO (Self-supervised Open-dialogue Learning) for model training.

PPTOD [24] is an end-to-end Transformer-based model for intent recognition and slot prediction tasks in multi-turn dialogues. It utilizes a pre-trained Transformer model for dialogue state tracking and generation.

SM2 [25] achieves extremely competitive results by combining a meta-learning scheme with an improved training retrieval mechanism.

DualLoRA [27] enhances model performance by processing and optimizing dialogue context and prompts through two distinct Low-Rank Adaptation (LoRA) components, all without introducing additional inference latency.

Emotion Summary Generation

In this experiment, we proposed using emotion summary generation to identify emotions in dialogues. Most recent works utilize multimodal data from the MELD dataset, while we used only text from MELD for emotion summary generation. Therefore, our primary objective here is to validate the effectiveness of using emotion summary generation and compare it with classical models.

text-CNN [28] is a convolutional neural network-based model for emotion summary generation tasks, aiming to capture local features of emotional information through convolution operations on dialogue text.

bcLSTM [29] is a bidirectional contextual long short-term memory network that considers both contextual and current input information simultaneously and captures long-term dependencies in dialogues through LSTM units.

DialogueRNN [30] is a recurrent neural network model that captures the evolution of emotions and contextual information in dialogues by recursively modeling the dialogue history.

5 Experimental Results

5.1 Intent Summary Generation

We present the performance of our method, MESM, in different Few-Shot settings compared to the baselines mentioned in Section 4.4 in Table 4. We evaluated our model on the MultiWOZ 2.0 and 2.1 version. We observed that under low-parameter training settings, MESM does not exhibit significant advantages over previous methods.

Table 4: Multi-domain Few-Shot JGA assessed across all domains collectively

Model (ver.)	1%	5%	10%	100%
TRADE (2.0)	11.74	32.41	37.42	48.62
TRADE + Self-supervision (2.0)	23.0	37.82	40.65	–
MinTL (2.0)	9.25	21.28	30.32	52.10
SOLOIST (2.0)	13.21	26.53	32.42	53.20
PPTOD (2.0)	31.46	43.61	45.96	53.89
DS2-T5 (2.0)	36.15	45.14	47.61	54.78
MESM (2.0)	35.13	45.73	48.68	55.45
TRADE (2.1)	12.58	31.17	36.18	46.00
TRADE + Self-supervision (2.1)	21.90	35.13	38.12	–
DS2-BART (2.1)	28.25	37.71	40.29	46.86
DS2-T5 (2.1)	33.76	44.20	45.38	52.32
SM2-3B (2.1)	38.06	39.94	39.85	–
SM2-11B (2.1)	38.36	44.64	46.02	–
DualLoRA (2.1)*	38.72	–	–	52.82
MESM (2.1)	33.83	45.07	45.68	53.26

Note: Except for *, all other data is sourced from the original papers.

Consequently, we retrained DS2 and DualLoRA in our environment with multi-domain full-parameter training and evaluated its specific domain JGA. As shown in Table 5, the overall performance of MESM was 55.45 (2.0) and 53.26 (2.1), achieving a notable improvement compared to DS2's 54.45 (2.0) and 52.45 (2.1). Furthermore, compared to DualLoRA, our Multi-Head Encoder Shared Model (MESM) shows significant advantages on the MultiWOZ 2.1 dataset.

Table 5: The joint goal accuracy (%) for specific domains in the MultiWOZ

Model(ver.)	Attraction	Hotel	Restaurant	Taxi	Train	Average
DS2-T5 (2.0)	76.43	57.30	69.93	80.19	83.07	54.45
MESM (2.0)	78.27	60.59	72.01	80.46	83.85	55.45
DS2-BART (2.1)	65.32	49.19	63.75	79.70	74.69	46.71
DS2-T5 (2.1)	75.71	56.06	69.79	80.09	83.71	52.45

(Continued)

Table 5 (continued)

Model(ver.)	Attraction	Hotel	Restaurant	Taxi	Train	Average
DualLoRA (2.1)	75.32	57.55	69.63	80.52	83.06	52.82
MESM (2.1)	76.00	54.33	68.38	80.07	87.45	53.26

Furthermore, we analyzed the slot accuracy for specific domains and compared MESM with DS2 on MultiWOZ 2.0 and MultiWOZ 2.1. As illustrated in Fig. 3, our model generally achieved better results compared to the baselines in most scenarios.

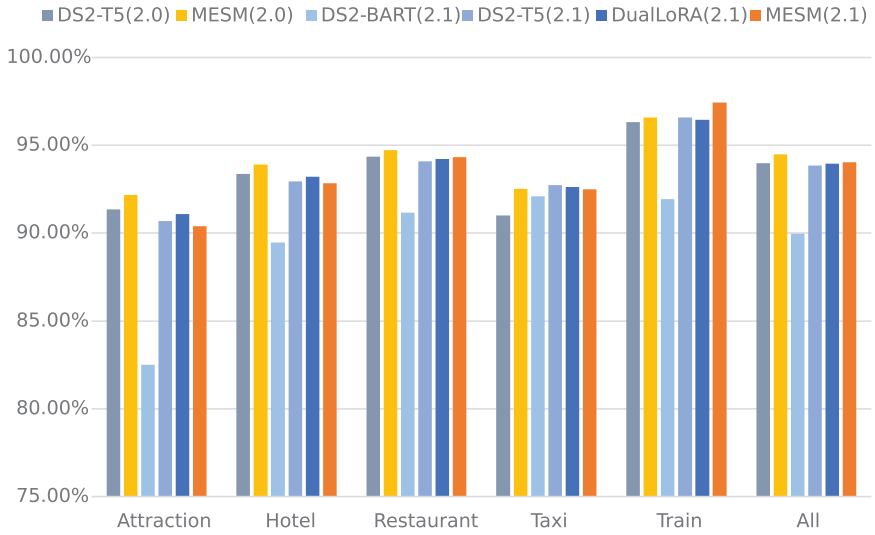

Figure 3: The slots accuracy for specific domains in the MultiWOZ

5.2 Emotion Summary Generation

In Table 6, we compare the accuracy of each emotion classification and the weighted average (w-avg) results of our model. It is important to note that our method only utilizes text for prediction. This experiment demonstrates that our proposed approach of generating emotion summaries through a summary generation model is feasible and exhibits certain advantages.

Table 6: Test-set results of emotion classification in MELD

Models		Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	W-avg
Text-CNN		34.49	8.22	3.74	49.39	74.88	21.05	45.45	55.02
cMKL	Text + audio	39.50	16.10	3.75	51.39	72.73	23.95	46.25	55.51
bcLSTM	Text	42.06	21.69	7.75	54.31	71.63	26.92	48.15	56.44
	Audio	25.85	6.06	2.90	15.74	61.86	14.71	19.34	39.08

(Continued)

Table 6 (continued)

Models		Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	W-avg
	Text + audio	43.39	23.66	9.38	54.48	76.67	24.34	51.04	59.25
DialogueRNN	Text	40.59	2.04	8.93	50.27	75.75	24.19	49.38	57.03
	Audio	35.18	5.13	5.56	13.17	65.57	14.01	20.47	41.79
	Text + audio	43.65	7.89	11.68	54.40	77.44	34.59	52.51	60.25
MESM (Ours)	Text	39.13	19.12	18.00	69.65	79.14	28.85	59.07	63.48

6 Analysis

6.1 Ablation Study

To assess the impact of the Feature Fusioner on model's performance, we conducted an ablation study. We designed three methods to assess the model's performance: (1) No Fusion between emotion features and intent features, where emotion summary generation and intent summary generation are conducted independently; (2) Simple fusion of emotion features T_e and intent features T_i, where $T_f = (1 - w) \times T_e + w \times T_i$, with w being a trainable float parameter, and T_f being the fused feature; (3) Complete fusion using the Feature Fusioner to combine emotion features and intent features. We conducted experiments using MultiWOZ 2.1, using the same train-validation-test split and maintaining consistent hyperparameter settings.

As shown in Table 7, we observed that the Feature Fusioner significantly improved performance compared to the No Fusion method, where intent summary generation was conducted independently. Due to the additional noise introduced by Simple Fusion, its performance is actually inferior to that of No Fusion. This demonstrates the effectiveness of using the Feature Fusioner to integrate emotion features T_e and intent features T_i, highlighting that the use of the Feature Fusioner is crucial for the performance of our model.

Table 7: Ablation study on MultiWOZ 2.1

Model	Joint goal accuracy
No fusion	51.21
Simple fusion	48.90
Feature Fusioner (Our full)	53.26

6.2 Cost of Experimental Engineering

In our experiments, we extended the dialogue summarizer (φ) and dialogue state extractor (δ) based on [11] to enable the conversion of emotion labels into emotion summaries and the extraction of emotion labels from these summaries. Additionally, we processed the MELD dataset to convert it into multi-turn dialogues. These tasks were relatively low-cost, requiring only two to three days for an expert to complete.

Our model comprises 1.6 billion trainable parameters, and we utilized the NVIDIA RTX A6000 48G for our experiments. For each turn in the dialogue, we input the current dialogue history and

perform inference. In terms of time complexity, the worst-case inference time complexity of our model is $O(k + \tau)$, where k denotes the number of slots and τ represents the model inference time. This demonstrates the efficiency and scalability of our approach in practical applications.

6.3 Limitations

Our experiments have several limitations. Due to the requirement of generating dialogue summaries that encompass both emotion and intent tasks, our model has a larger scale. The number of trainable parameters in our model is more than double that of other Dialogue State Tracking (DST) models, necessitating more substantial hardware resources. Additionally, the increased parameter size and the complexity of training the Feature Fusioner for feature integration result in longer training times. For instance, training on the full data for the multi-domain scenario takes approximately 75 h. These issues of large parameter size and extended training time need to be addressed in future work to improve the efficiency and practicality of our model.

7 Conclusions

To address the limitation of traditional Dialogue State Tracking (DST) modules in Task-Oriented Dialogue Systems (TOD) which fail to concurrently track user intents and emotions, we propose the Multi-Head Encoder Sharing Model (MESM). In our approach, we reframe the emotion recognition task as a summary generation task, guiding the model to produce user emotion summaries through an extended state-to-summary converter. Additionally, during intent recognition training, the Feature Fusioner in MESM integrates intent and emotion features, ultimately generating a complete summary that encompasses both user intent and emotion.

We evaluated the performance of our method on the MultiWoZ and MELD datasets. The experimental results demonstrate that MESM effectively captures user emotions and enhances the accuracy of generated user intent summaries. Our model maintains Few-Shot capabilities while achieving extremely competitive performance in multi-domain full-parameter training compared to current baselines. However, our model has a larger parameter size and longer training time compared to other DST models. These issues will be addressed in future work to improve the efficiency and practicality of our approach.

Acknowledgement: We extend our gratitude to the anonymous reviewers for their insightful feedback.

Funding Statement: This research was funded by the Science and Technology Foundation of Chongqing Education Commission (Grant No. KJQN202301153), the Scientific Research Foundation of Chongqing University of Technology (Grant No. 2021ZDZ025) and the Postgraduate Innovation Foundation of Chongqing University of Technology (Grant No. gzlcx20243524).

Author Contributions: Conceptualization, Xinlai Xing; methodology, Junliang Chen; software, Junliang Chen and Shuran Zhou; validation, Junliang Chen, Shuran Zhou and Runqing Zhang; formal analysis, Xiaochuan Zhang; investigation, Shuran Zhou; resources, Xinlai Xing; data curation, Shuran Zhou; writing original draft preparation, Junliang Chen; writing review and editing, Xinlai Xing; visualization, Shuran Zhou; supervision, Xinlai Xing; project administration, Xiaochuan Zhang; funding acquisition Xinlai Xing and Xiaochuan Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Junliang Chen, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, 2017. doi: 10.1145/3166054.3166058.

[2] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3055–3155, 2023. doi: 10.1007/s10462-022-10248-8.

[3] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 2011–2027, 2020. doi: 10.1007/s11431-020-1692-3.

[4] C. -S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," in *Proc. 57th Annual Meet.e Assoc. Comput. Linguist.*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 808–819. doi: 10.18653/v1/P19-1078.

[5] V. Balaraman, S. Sheikhalishahi, and B. Magnini, "Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey," in *Proc. 22nd Annual Meet. Special Interest Group Disc. Dial.*, 2021, pp. 239–251.

[6] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018. doi: 10.1609/aaai.v32i1.11325.

[7] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Trans. Inf. Syst.*, vol. 2, no. 1, pp. 26–41, 1984. doi: 10.1145/357417.357420.

[8] P. Budzianowski *et al.*, "MultiWOZ-A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proc. 2018 Conf. Empi. Meth. Nat. Lang. Process.*, Brussels, Belgium: Association for Computational Linguistics, Oct.–Nov. 2018, pp. 5016–5026. doi: 10.18653/v1/D18-1547.

[9] M. Eric *et al.*, "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines," in *Proc. Twelfth Lang. Res. Eval. Conf.*, Marseille, France: European Language Resources Association, May 2020, pp. 422–428.

[10] S. Feng *et al.*, "EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems," in *Proc. Thirteenth Lang. Res. Eval. Conf.*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 4096–4113.

[11] J. Shin, H. Yu, H. Moon, A. Madotto, and J. Park, "Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking," in *Find. Assoc. Comput. Linguist.: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3824–3846. doi: 10.18653/v1/2022.findings-acl.302.

[12] J. D. Williams and S. Young, "Partially observable markov decision processes for spoken dialog systems," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 393–422, 2007. doi: 10.1016/j.csl.2006.06.008.

[13] J. D. Williams, M. Henderson, A. Raux, B. Thomson, A. Black and D. Ramachandran, "The dialog state tracking challenge series," *AI Magazine*, vol. 35, no. 4, pp. 121–124, 2014. doi: 10.1609/aimag.v35i4.2558.

[14] S. Gao, S. Agarwal, D. Jin, T. Chung, and D. Hakkani-Tur, "From machine reading comprehension to dialogue state tracking: Bridging the gap," in *Proc. 2nd Workshop Nat. Lang. Process. Conversat. AI*, Association for Computational Linguistics, Jul. 2020, pp. 79–89. doi: 10.18653/v1/2020.nlp4convai-1.10.

[15] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapt. Assoc. Comput. Linguist.: Human Lang. Technol.*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[16] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[17] Y. Zhang, Z. Ou, and Z. Yu, "Task-oriented dialog systems that consider multiple appropriate responses under the same context," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 9604–9611, 2020. doi: 10.1609/aaai.v34i05.6507.

[18] C. -S. Wu, S. C. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proc. 2020 Conf. Empirical Meth. Nat. Lang. Process. (EMNLP)*, Association for Computational Linguistics, Nov. 2020, pp. 917–929. doi: 10.18653/v1/2020.emnlp-main.66.

[19] J. Zhang *et al.*, "Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking," in *Proc. Ninth Joint Conf. Lexic. Computat. Semant.*, Barcelona, Spain: Association for Computational Linguistics, Dec. 2020, pp. 154–167.

[20] Z. Lin, A. Madotto, G. I. Winata, and P. Fung, "MinTL: Minimalist transfer learning for task-oriented dialogue systems," in *Proc. 2020 Conf. Empiric. Meth. Nat. Lang. Process. (EMNLP)*, Association for Computational Linguistics, Nov. 2020, pp. 3391–3405. doi: 10.18653/v1/2020.emnlp-main.273.

[21] F. Mi, Y. Wang, and Y. Li, "CINS: Comprehensive instruction for few-shot learning in task-oriented dialog systems," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, pp. 11 076–11 084, 2022. doi: 10.1609/aaai.v36i10.21356.

[22] S. Li *et al.*, "Zero-shot generalization in dialog state tracking through generative question answering," in *Proc. 16th Conf. Eur. Chap. Assoc. Computational Linguist.*, 2021, pp. 1063–1074.

[23] G. Campagna, A. Foryciarz, M. Moradshahi, and M. Lam, "Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking," in *Proc. 58th Annual Meet. Assoc. Computat. Linguist.*, 2020, pp. 122–132.

[24] Y. Su *et al.*, "Multi-task pre-training for plug-and-play task-oriented dialogue system," in *Proc. 60th Annual Meet. Assoc. Computat. Linguist.*, 2022, pp. 4661–4676.

[25] D. Chen, K. Qian, and Z. Yu, "Stabilized in-context learning with pre-trained language models for few shot dialogue state tracking," in *Find. Assoc. Computat. Linguist.: EACL 2023*, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1551–1564. doi: 10.18653/v1/2023.findings-eacl.115.

[26] Z. Lin *et al.*, "Zero-shot dialogue state tracking via cross-task transfer," in *Proc. 2021 Conf. Empiric. Meth. Nat. Lang. Process.*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7890–7900. doi: 10.18653/v1/2021.emnlp-main.622.

[27] X. Luo, Z. Tang, J. Wang, and X. Zhang, "Zero-shot cross-domain dialogue state tracking via dual low-rank adaptation," in *Proc. 62nd Annual Meet. Assoc. Computat. Linguist.*, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5746–5765. doi: 10.18653/v1/2024.acl-long.312.

[28] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empiric. Meth. Nat. Lang. Process. (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.

[29] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh and L. -P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annual Meet. Assoc. Computat. Linguist. (Volume 1: Long Papers)*, Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 873–883. doi: 10.18653/v1/P17-1081.

[30] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh and E. Cambria, "DialogueRNN: An attentive rnn for emotion detection in conversations," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 6818–6825, 2019. doi: 10.1609/aaai.v33i01.33016818.

[31] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meet. Assoc. Computat. Linguist.*, Association for Computational Linguistics, 2019.

[32] G. Dong *et al.*, "PSSAT: A perturbed semantic structure awareness transferring method for perturbation-robust slot filling," in *Proc. 29th Int. Conf. Computati. Linguist.*, 2022, pp. 5327–5334.

[33] G. Zhao, G. Dong, Y. Shi, H. Yan, W. Xu and S. Li, "Entity-level interaction via heterogeneous graph for multimodal named entity recognition," in *Find. Assoc. Computat. Linguist.: EMNLP*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 6345–6350. doi: 10.18653/v1/2022.findings-emnlp.473.

[34] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "Mediasum: A large-scale media interview dataset for dialogue summarization," in *Proc. 2021 Conf. North American Chapt. Assoc. Computat. Linguist.: Human Lang. Technol*, Association for Computational Linguistics, 2021.

[35] M. Zhong *et al.*, "QMSum: A new benchmark for query-based multi-domain meeting summarization," in *Proc. 2021 Conf. North American Chapt. Assoc. Computati. Linguist.: Human Lang. Technol.*, 2021, pp. 5905–5921.

[36] Y. Chen, Y. Liu, and Y. Zhang, "Dialogsum challenge: Summarizing real-life scenario dialogues," in *Proc. 14th Int. Conf. Nat. Lang. Gen.*, 2021, pp. 308–313.

[37] C. Wu, L. Liu, W. Liu, P. Stenetorp, and C. Xiong, "Controllable abstractive dialogue summarization with sketch supervision," in *Find. Assoc. Computat. Linguist.: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 5108–5122. doi: 10.18653/v1/2021.findings-acl.454.

[38] X. Feng, X. Feng, L. Qin, B. Qin, and T. Liu, "Language model as an annotator: Exploring dialogpt for dialogue summarization," in *Proc. 59th Annual Meet. Assoc. Computat. Linguist. 11th Int. Joint Conf. Nat. Lang. Process.*, 2021, pp. 1479–1491. doi: 10.18653/v1/2021.acl-long.11.

[39] M. Khalifa, M. Ballesteros, and K. Mckeown, "A bag of tricks for dialogue summarization," in *Proc. 2021 Conf. Empiric. Meth. Nat. Lang. Process.*, 2021, pp. 8014–8022.

[40] M. Zhao *et al.*, "Mutually improved response generation and dialogue summarization for multi-domain task-oriented dialogue systems," *Knowl.-Based Syst.*, vol. 279, no. 2, 2023, Art. no. 110927. doi: 10.1016/j.knosys.2023.110927.

[41] I. T. Aksu, M. -Y. Kan, and N. Chen, "Prompter: Zero-shot adaptive prefixes for dialogue state tracking domain adaptation," in *Proc. 61st Annu. Meet. Assoc. Computat. Linguist. (Volume 1: Long Papers)*, Toronto, ON, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4588–4603. doi: 10.18653/v1/2023.acl-long.252.

[42] C. -S. Wu, S. C. Hoi, and C. Xiong, "Improving limited labeled dialogue state tracking with self-supervision," in *Find. Assoc. Computat. Linguist.: EMNLP 2020*, 2020, pp. 4462–4472. doi: 10.18653/v1/2020.findings-emnlp.400.

[43] B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden and J. Gao, "SOLOIST: Few-shot task-oriented dialog with a single pretrained auto-regressive model," 2020, *arXiv:2005.05298*.