Multimedia Systems (2025) 31:98
https://doi.org/10.1007/s00530-025-01672-1

REGULAR PAPER

Dialogue sentiment analysis based on dialogue structure pre-training

Liang Yang[1,2] · Qi Yang[1] · Jingjie Zeng[1] · Tao Peng[1] · Zhihao Yang[1] · Hongfei Lin[1]

Received: 15 October 2024 / Accepted: 13 January 2025 / Published online: 6 February 2025

Abstract
The task of dialogue sentiment analysis aims to identify the sentiment polarity of utterances in the context of a dialogue. Pre-trained models often struggle to capture the logical structure of a dialogue, making this task challenging. To address this issue, we propose a dialogue sentiment analysis framework that leverages pre-training on dialogue structure. Our proposed framework includes three sub-tasks for pre-training: utterance order sorting, sentence backbone regularization, and sentiment shift detection. These tasks are designed to improve the model's ability to mine dialogue logical relationships and sentiment interactions. By focusing on learning the logical structure of dialogues and the perception of sentiment interactions, our framework is able to improve the performance of pre-trained models on recognizing the sentiment polarity of dialogues. This is demonstrated by the convincing results obtained on the public MEISD dataset.

Keywords Dialogue structure · Sentiment analysis · Pre-training model

1 Introduction

In the digital era, the proliferation of social media platforms and various instant messaging tools has dramatically expanded the avenues through which individuals communicate. Every day, an overwhelming volume of dialogue texts are being created, circulated, and exchanged across the Internet, encapsulating a vast array of users' personal emotions and attitudes, each carrying a spectrum of sentiment polarities that hold considerable value both socially and commercially [1]. Leveraging cutting-edge technologies such as artificial intelligence and big data analytics, it has become feasible to systematically sift through these extensive datasets of dialogue texts to extract valuable sentiment information. This capability is of particular importance to government entities, providing them with the tools to closely track the evolution of public sentiment online. Such insights enable the establishment of a modernized framework for social governance, tailored to the nuances and dynamics of digital communication channels [2].

Text sentiment analysis constitutes a methodological approach aimed at discerning the emotional or attitudinal disposition embedded within textual content through a series of steps that include processing, analytical examination, inductive reasoning, and logical deduction, as detailed in the work by Liu [3]. Wu et al. [4] used hierarchical attention networks to predict user emotions towards products. Xiong et al. [5] focused on fine-grained sentiment analysis tasks, developing a module based on machine reading comprehension to accomplish this task. Hosseinalipour et al. [6] developed the Horse Herd Optimisation Algorithm and applied it in the field of sentiment analysis, achieving very good results. Jiang et al. [7] focused on methodology by combining methods such as K-means++ [8], SMOTE [9], CNN [10], and BiLSTM [11] models to perform sentiment

Communicated by Bing-kun Bao.

✉ Jingjie Zeng
jjtail@mail.dlut.edu.cn

Liang Yang
liang@dlut.edu.cn

Qi Yang
qiyang@mail.dlut.edu.cn

Tao Peng
taop@mail.dlut.edu.cn

Zhihao Yang
yangzh@dlut.edu.cn

Hongfei Lin
hflin@dlut.edu.cn

[1] School of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2] Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, Dalian, China

analysis on texts. Li et al. [12] focused on data augmentation on sentiment analysis datasets to improve the performance of the model, which achieved good results. But these methods are all for static isolated text and do not have dynamic interaction similar to dialogue. Dialogue sentiment analysis, as explored by Pérez-Rosas et al. [13], necessitates the segmentation of a conversation into discrete segments or utterances. This segmentation is guided by various indicators such as natural pauses in speech, changes in the speaker, and the structural cues of dialogue. The goal is to accurately identify and interpret the subjective sentiments conveyed by each participant within each utterance, taking into account the broader context and flow of the ongoing dialogue. This nuanced approach allows for a more refined understanding of the sentiment dynamics at play within interactive communicative processes.

Pre-trained language models, as discussed in foundational studies by Devlin et al. [14] and Yang et al. [15], having revolutionized natural language processing by providing a method to encode texts into semantic representation vectors, facilitating a wide range of applications in different fields. These models, built on the principles of self-supervised learning, have shown remarkable proficiency in capturing complex linguistic patterns and representations. However, when these models are applied to new domains that differ from their initial training data, they often require a process of fine-tuning. This fine-tuning is crucial for the models to acquire and integrate new domain-specific knowledge and semantic nuances. In the context of fine-tuning, the standard practice involves feeding serialized text into the pre-trained model to generate a comprehensive feature representation of the text. This approach is generally effective for texts that are short and have a straightforward structure. However, this strategy encounters limitations when applied to dialogues. Dialogues, by their nature, have a unique structure characterized by exchanges between participants, and simply serializing these exchanges into a single stream of text can lead to the loss of vital structural and conversational context. This loss can significantly impede the model's ability to understand and analyze the intricacies of dialogue interactions, as highlighted by Mehri et al. [16]. Therefore, adapting pre-trained models to handle dialogue texts necessitates a more nuanced approach that can preserve and interpret the inherent structure and dynamics of dialogues, ensuring that the rich contextual information is not overlooked.

To adapt pre-trained language models for the nuanced task of analyzing dialogue-structured texts, a variety of research efforts have introduced dialogue-specific pre-training tasks. For example, Li et al. [17] proposed a deep context modeling architecture (DCM) method for dialogue modeling. Xu et al. [18] constructed a multi-turn dialogue model with topic awareness ability by unsupervised segmentation and extraction of topic aware discourse. Zeng et al. [19] focused on multi turn dialogue models, incorporating dialogue topic information into the model to generate smooth conversations. These tasks are aimed at domain-adaptive pre-training specifically tailored to dialogue texts. Notably, much of this research has concentrated on narrower aspects within the broader area of dialogue comprehension, as exemplified by the studies conducted by Zhang et al. [20, 21]. However, when it comes to the specialized domain of dialogue sentiment analysis, there appears to be a gap in the literature, with no significant exploration into the application of dialogue pre-training concepts tailored to this area. To bridge this gap, the present paper introduces an innovative pre-training framework specifically designed for the task of dialogue sentiment analysis, leveraging the foundational architecture of the BERT model as described by Devlin et al. [14]. This proposed framework is structured around three meticulously crafted sub-tasks: sorting utterances by their order within a dialogue, regularizing the backbone of sentences to maintain coherence, and detecting shifts in sentiment throughout the dialogue. These sub-tasks are strategically devised to refine and augment the BERT model's capacity for discerning the intricate structural and emotional dynamics characteristic of dialogues, particularly in the context of sentiment analysis.To optimize the integration of domain-adaptive post-training with fine-tuning processes, the framework places a significant emphasis during the domain-adaptive post-training phase on deeply understanding the inherent structure of dialogues. This involves intensive training activities focused on the aforementioned sub-tasks-utterance order sorting, sentence backbone regularization, and sentiment shift detection-using dialogue-specific datasets. This comprehensive approach ensures that the model not only learns the general linguistic features from the pre-training corpus but also acquires a profound understanding of the unique patterns and emotional trajectories present in dialogue-based communications.

In summary, the main contributions of this paper are:

- Three pre-training sub-tasks, namely utterance order sorting, sentence backbone regularization, and sentiment shift detection, are designed to enhance the ability of the BERT model on perceiving dialogue structure and sentiment changes in dialogue sentiment analysis tasks.
- Special symbols are used to reinforce logical relationships among tasks, specifically, to learn the connection between the dialogue structure at the sentence granularity and the connection between the backbone of the sentence at the word granularity.
- A large number of experiments are conducted on the sentiment polarity prediction task of the MEISD dataset. The experiments demonstrate that the proposed dialogue structure pre-training framework can better solve the dia-

logue sentiment analysis task. The effectiveness of each subtask in the dialogue sentiment analysis task is also analyzed.

2 Related works

In this section, we primarily review the related research on large language models, pre-trained models, and dialogue sentiment analysis.

2.1 Large language models with sentiment

With the rapid development of large-scale language models (LLMs) and artificial intelligence generated content (AIGC) technology [22, 23], guiding these advanced technologies to produce dialogues that are closer to human emotions, and providing users with emotionally resonant interactive experiences, has gradually become an important topic in the field of artificial intelligence research. Against this backdrop, the task of dialog emotion analysis becomes particularly critical. It offers a practical method to help machines better understand and respond to human emotions by accurately detecting the emotional polarity and changes in dialogues. This field brings together cutting-edge technologies from multiple disciplines such as deep learning, natural language processing (NLP), and artificial intelligence (AI), aiming to delve into the emotional details in dialog content. It not only identifies the basic emotional tendencies of each message in a dialogue but also takes into account the subtle changes in emotions within the context of the dialogue and the emotional interaction patterns between participants. For example, researchers like Rozga [24] have successfully identified subtle changes in children's emotions through in-depth analysis of their interactive dialogues with robots, and accordingly guided the robots to adjust their behaviors and reactions, effectively fostering positive emotional exchanges. These breakthrough research achievements have not only significantly advanced the progress of dialog emotion analysis technology but also provided valuable experiences and insights for designing and implementing dialogue systems with more emotional intelligence and human-like characteristics, thereby demonstrating great application potential and value in the broad field of human-computer interaction.

Chen et al. [25] proposes an innovative solution to address personality consistency and dialogue coherence in dialogue systems. The LMEDR method proposed by Chen cleverly combines external and internal memory mechanisms, where external memory utilizes entailment text pairs from natural language inference datasets to learn potential entailment relationships, while internal memory focuses on processing utterance information in dialogues. By setting orthogonal constraints between these two memory spaces, the dialogue independence of potential entailment relationships is ensured. This dual memory mechanism design not only solves the problem of traditional methods requiring large amounts of annotated data but also breaks through the limitation of focusing only on single response generation, achieving control over the coherence of entire dialogues.

2.2 Pre-trained language models within dialogue

The foundational BERT model, as introduced by Devlin et al. [14], is built upon two core pre-training tasks: the prediction of masked words within a given context and the determination of whether one sentence logically follows another. This setup allows for the fine-tuning process, where researchers can derive vector representations of texts that are then used to train networks for specific tasks. In the realm of dialogue text analysis, Xu [26] extends the utility of the BERT model by introducing a set of four dialogue-centric self-supervised pre-training tasks specifically designed for context-response matching scenarios. These tasks include predicting the next session in a sequence, restoring an utterance from a scrambled version, identifying incoherent segments within dialogues, and discriminating between consistent and inconsistent dialogue parts. These tasks collectively aim to enhance the model's understanding of dialogue attributes such as continuity, coherence, and consistency, which are crucial for effective dialogue analysis.

Despite the targeted nature of these tasks for context-response matching, their specificity might limit their applicability to a broader range of dialogue-related tasks. Addressing this limitation, Zhang [27] proposes a more universally applicable pre-training sub-task aimed at bolstering dialogue comprehension capabilities. This includes an utterance order restoration task, designed to reinforce the model's grasp of the logical flow between individual dialogue utterances. Additionally, a sentence backbone regularization task is proposed to refine the model's understanding of the internal logical structure of sentences within utterances. This approach is reported to yield notable improvements in performance on downstream tasks focused on dialogue comprehension. Through these enhancements, the BERT model's adaptability and effectiveness in processing and analyzing dialogue texts are significantly improved, enabling more nuanced and accurate interpretations of dialogue structures and dynamics.

2.3 Dialogue sentiment analysis

Initial research in dialogue sentiment analysis predominantly leverages the capabilities of convolutional neural networks (CNNs), as discussed by Kim [10], and recurrent neural networks (RNNs), highlight in the work by Medsker and Jain [28]. Within this domain, Hsu [29] introduces the TextCNN approach to extract feature representations from individual utterances and subsequently employes a bidirectional long short-term memory network (BiLSTM) to analyze the interconnections within dialogue contexts. This CNN-BiLSTM hybrid model demonstrates commendable performance on the EmotionLines dataset, showcasing its effectiveness in capturing emotional nuances in dialogues. Complementing this approach, Shi [11] implement a hierarchical bidirectional long short-term memory network (H-BiLSTM) to address the layered structure inherent in dialogues. This method marks a substantial advancement over previous techniques like CNNs and LSTMs by providing a more nuanced understanding of dialogue dynamics.

The advent of pre-trained language models has significantly influenced the development of methodologies across a broad spectrum of tasks within natural language processing. Notably, in the context of the SemEval-2019 Task 3 [30], numerous participating teams incorporated pre-trained models such as GloVe, introduced by Pennington et al. [31], and BERT, as a foundation for further model training and fine-tuning. This approach yielded top-tier prediction accuracies on the EmoContext dataset, illustrating the potent impact of pre-trained models in enhancing sentiment analysis accuracy.

Building on this momentum, Hazarika [32] uniquely merged the BERT model with transfer learning techniques, initially training on a dialogue generation task within a source domain before transferring the learned encoder weights to a dialogue sentiment analysis task. This innovative strategy, which involves further adapting the decoder to sentiment classification, led to a notable leap in performance compared to direct applications of pre-trained models.

Further expanding the toolkit for dialogue sentiment analysis, Shen [33] introduced an enhancement to the XLNet model by integrating a memory mechanism, resulting in the DialogXL model. This model innovatively utilizes a memory module to carry forward context from one dialogue segment to the next, while employing a self-attention mechanism to refine the integration of contextual information within dialogues. This advancement underscores the ongoing evolution and refinement of methodologies aimed at understanding and analyzing sentiment in dialogue-based interactions.

3 Approach

This section mainly introduces the definition of our model, the three loss functions we defined to enable the model to better recognize dialogue emotions, and the training framework of the model.

In this study, we conducted a systematic classification and analysis of dialogue structures. Based on different patterns of dialogue interaction, we divided dialogue structures into five main types: Linear dialogue structure, Branch dialogue structure, Nested dialogue structure, Parallel dialogue structure, and Circular dialogue structure. The Linear dialogue structure appears as simple back-and-forth interactions between participants, such as two people having a sequential conversation about the weather; the Branch dialogue structure reflects that a single topic may lead to multiple subtopics, like discussing specific movies, locations, etc., from the weekend plan; the Nested dialogue structure is characterized by including sub-dialogues within the main dialogue framework, such as asking about attendance during a project discussion; the Parallel dialogue structure allows for simultaneous discussions of multiple topics, like discussing shopping and house-hunting plans at the same time; and the Circular dialogue structure manifests as certain topics recurring periodically. These different dialogue structures exhibit unique emotional expression patterns, topic transition modes, and contextual dependencies. By conducting an in-depth analysis of these structures, we can better understand the patterns of emotional flow in dialogues, providing more accurate structural information support for subsequent sentiment analysis tasks.

3.1 Task formulation

The primary objective of dialogue sentiment analysis is to meticulously evaluate and determine the sentiment polarity associated with individual utterances within the framework of a dialogue context. This involves scrutinizing a multi-turn dialogue, denoted as $C = \{U_1, U_2, ..., U_m\}$, where U_i symbolizes the utterance corresponding to the i-th turn within the dialogue, and m signifies the total number of turns. Alongside the dialogue, there exists a set of sentiment polarity labels represented as $Y = \{y_1, y_2, ..., y_m\}$, with each $y_i \in \{positive, neutral, negative\}$ serving as the designated sentiment label for the utterance U_i. The fundamental challenge lies in devising and training a predictive model, denoted as f, capable of accurately associating the given input dialogue C with the correct set of sentiment labels Y. This process requires the model to not only understand the lexical and syntactical nuances of each utterance but also to grasp the complex interplay of sentiments as the dialogue progresses from one turn to the next, thereby achieving a

nuanced understanding of the sentiment dynamics at play within conversational exchanges.

3.2 Utterance order sorting

Within the framework of dialogue analysis, the aspect of semantic coherence is paramount in the process of dialogue modeling. Zhu et al. [34] summarized two reasons for vanilla code-switched sentences often lack semantic and grammatical coherence, (1) randomly replacing code-switched tokens with equal probability and (2) disregarding token-level dependency within each language, which is also very enlightening for our task. It is essential that utterances within a dialogue adhere to a coherent relationship and maintain a logical progression, as this is fundamental to accurately conveying the intended semantics of the dialogue. Consequently, the sequential interconnection and logical order of utterances emerge as pivotal elements in the characterization of dialogue dynamics. For any given series of dialogue utterances, the employed model is tasked with assimilating the progression and coherence of these utterances. This learning process is instrumental in enabling the model to grasp the underlying logical framework of the dialogue. By doing so, not only is the model's capacity to accurately represent and interpret the dialogue structure significantly enhanced, but its overall robustness and effectiveness in dialogue modeling are also substantially bolstered. This focus on understanding and maintaining semantic coherence is thus central to improving the model's performance in capturing the essence and complexity of dialogical exchanges.

In the process of preparing the dialogue context for the BERT model, we assume an original order of utterances denoted by $C = \{U_1, U_2, ..., U_m\}$. This sequence of utterances is then meticulously arranged and concatenated with specific delimiters to form the input structure for the BERT model. The formulated input, represented as $X = \{[\texttt{CLS}]U_1[\texttt{EOU}]U_2...[\texttt{EOU}][\texttt{SEP}]U_m[\texttt{EOU}]\}$, incorporates special symbols like [CLS] and [SEP], which are pivotal in the BERT model's pre-training phase. Additionally, we introduce a novel symbol [EOU], which stands for "End Of Utterance". This unique symbol is specifically designed to demarcate the conclusion of individual utterances, ensuring a clear separation between the token sequences of consecutive utterances. Each utterance U_i in the dialogue is succeeded by an [EOU], designated as o_i, to signify its end. Upon processing the input X through the BERT model, we extract the hidden state H from the last layer of the Transformer encoder. This hidden state encapsulates the semantic attributes of every word within the dialogue context, providing a rich representation of the dialogue's semantic structure. The use of [CLS] is traditionally associated with capturing the global semantics in BERT, whereas in our approach, the [EOU] symbols play a critical role. They not only signify the end of an utterance but also facilitate the extraction of semantic features pertinent to each utterance, denoted as h_{o_i} for the segment associated with o_i. By employing this method, we are able to distill the semantic essence of individual utterances as well as the overarching sequential relationship that exists between them. This refined understanding aids the model in deciphering the logical framework and the nuanced interactions within the dialogue, enhancing its ability to interpret and respond to the context effectively. The Fig. 1 shows an example of utterance order restoration.

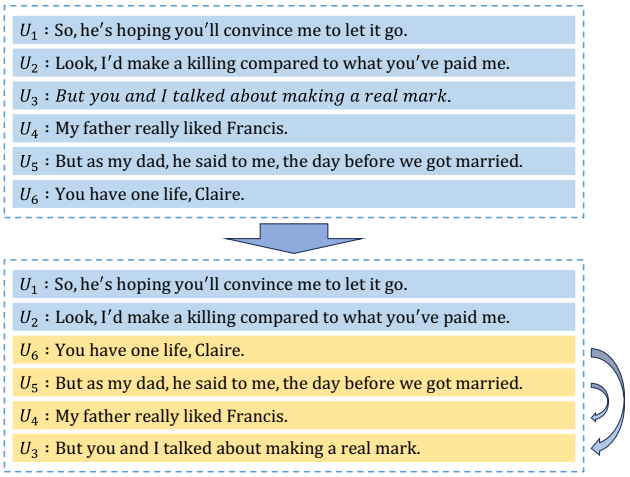

Fig. 1 Example of utterance order restoration

The loss function for the utterance order sorting task, represented by $\mathcal{L}_{\text{UOS}}$, is defined as follows (1):

$$\mathcal{L}_{\text{UOS}} = \log\left(1 + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sum_{k=1}^{m-j}\exp(\cos(h_{o_i}, h_{o_{j+k}}) - \cos(h_{o_i}, h_{o_j}))\right), \tag{1}$$

where $\cos(\cdot,\cdot)$ represents the cosine similarity between the i^{th} and j^{th} [EOU] of the utterances, m represents the length of the dialogue, and h_{o_i} and h_{o_j} are the vector representations of the i^{th} and j^{th} [EOU] of the utterances, respectively.

The goal of this loss function is to ensure that the similarity between consecutive utterances is high, thus constraining the model to learn the order relationship between dialogues. In other words, this loss function aims to minimize the difference in cosine similarity between consecutive [EOU] vectors, encouraging the model to learn a coherent dialogue structure, where utterances are in a logical sequence. By minimizing this loss function, the model can better understand the context and relationship between different utterances, and therefore improve its ability to analyze sentiment in the dialogue.

3.3 Sentence backbone regularization

Wu et al. [35] captured the rich correlation between predicates and discourse by constructing a predicate perception module. This has inspired us. Bordes et al. proposed a similar idea in [36] that the word embeddings of the head entity in relation triplets should be as close as possible to the word embeddings of the tail entity. We combine these two parts and use the BERT model to capture information about emotions in the semantic space. Specifically, it proposes that the word embedding of the subject plus the word embedding of the verb should be as close as possible to the word embedding of the object. To achieve this, the hidden layer state of the Utterance Order Sorting task is used to find the word embeddings corresponding to the subject, verb, and object, represented by h_{subject}, h_{verb}, and h_{object}, respectively. The Fig. 2 shows an example of subject-verb-object triples. The relationship between these three is then expressed in the following Eq. (2):

$$h_{o_{i-1}} + h_{\text{subject}} + h_{\text{verb}} \longrightarrow h_{o_i} + h_{\text{object}}, \tag{2}$$

where h_{o_i} represents the hidden layer vector of [EOU] for the i^{th} [EOU] of the utterances. The "$\longrightarrow$" symbol denotes that the distance between vectors should be as small as possible, and the cosine similarity is used to measure this distance in the semantic space. To implement this, the loss function of sentence backbone regularization is defined, assuming that T subject-verb-object triples are extracted from the input dialogue context. This will help to ensure that the sentence structure is logical and coherent, and thus enhance the model's ability to understand the underlying dialogue structure. Then the loss function of sentence backbone regularization can be defined as Equation (3).

$$\mathcal{L}_{\text{SBR}} = \sum_{t=1}^{m} (1 - \cos(h_{o_{i-1}} + h_{\text{subject}} + h_{\text{verb}}, h_{o_i} + h_{\text{object}})) \tag{3}$$

3.4 Sentiment shift detection

Sentiment shift detection arises from the sentiment coherence of dialogue, which aims to model the change of sentiment polarity between the current utterance and the previous utterance, and tries to establish the correlation between the change of sentiment polarity and semantics. Zhang et al. [37] focused on emotional transitions in conversations, which gave us a lot of inspiration. Given a context $C = \{U_1, U_2, ..., U_m\}$, for setting equal weight to all utterances, pseudo utterance U_0 is added to each dialogue. The text of U_0 is [SOU], and [SOU] is a special symbol designed for this task, short for "Start Of Utterance", meanwhile the sentiment polarity of U_0 is neutral. The sentiment shift label $Z = \{z_1, z_2, ..., z_m\}$ of each utterance can be calculated by Eq. (4).

$$z_i = \begin{cases} 1, & y_{i-1} \neq y_i \\ 0, & y_{i-1} = y_i \end{cases}, \forall i \in [1, m] \tag{4}$$

The input text for the sentiment shift detection task is constructed as $X_i = \{\text{[CLS]}\ U_{i-1}\ \text{[EOU][SEP]}U_i\ \text{[EOU]}\}, i \in [1, m]$, we can obtain the hidden state h_i in last hidden state of symbol [CLS]. Predicted sentiment transition probability distribution $\hat{z}_i$ is obtained by feed-forward neural network and Softmax function. By computing the cross-entropy loss function between the true sentiment shift labels and the predicted labels, the training loss for the SSD task can be calculated as following Eq. (5).

$$\mathcal{L}_{\text{SSD}} = -\sum_{k=1}^{m} z_k \log(\hat{z_k}) \tag{5}$$

3.5 Language modeling training framework

On the basis of the three dialogue pre-training sub-tasks, the masked language model (MLM) and next sentence prediction (NSP) of the two sub-tasks trained by the BERT model are also combined to model the language features at the word level and sentence level respectively. The target task is conversational sentiment analysis (CSA), which identifies the sentiment polarity of a single utterance. For the above six tasks, we design a model framework that combines domain adaptive post-training and multi-task fine-tuning. This model is named DPBERT (Dialogue Pretraining BERT).

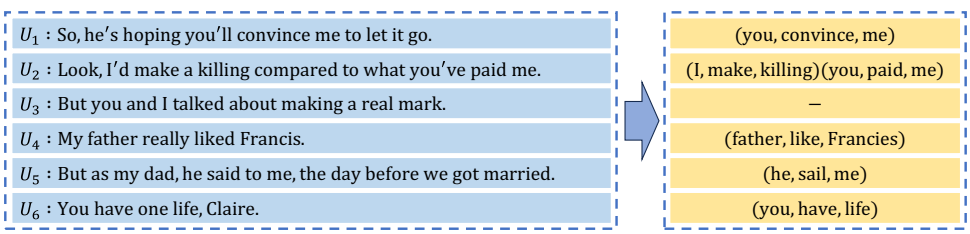

Fig. 2 Example of subject-verb-object triples

The core purpose of task is to analyze the emotional characteristics of utterances. We focus on semantic and dialogue structure modeling in the domain adaptive post-training, put masked language model (MLM), next sentence prediction (NSP), utterance order sorting (UOS), sentence backbone regularization (SBR) and sentiment shift detection(SSD) tasks in the stage of training.

Limited to the dataset itself, as long as the corpus in the same field can be put into the first stage of training, even if the dataset does not have sentiment shift labels, sentiment shift detection cannot be performed. Fine-tuning is mainly based on sentiment modeling, the target task conversational sentiment analysis (CSA) are placed in the second stage of training. The data of the training set is dominated so that it can be transferred to the test set with the same sentiment distribution during the model inference stage.

$$\mathcal{L}_{\mathrm{DAP}} = \lambda_1 \mathcal{L}_{\mathrm{MLM}} + \lambda_2 \mathcal{L}_{\mathrm{NSP}} + \lambda_3 \mathcal{L}_{\mathrm{UOS}} + \lambda_4 \mathcal{L}_{\mathrm{SBR}} + \lambda_5 \mathcal{L}_{\mathrm{SSD}} \tag{6}$$

$$\mathcal{L}_{\mathrm{MTL}} = \lambda_6 \mathcal{L}_{\mathrm{CSA}} \tag{7}$$

The loss function of domain adaptive post-training is the sum of the weights of the five tasks of masked language model (MLM), next sentence prediction (NSP), utterance order sorting (UOS), sentence backbone regularization (SBR) and sentiment shift detection (SSD), while the loss function of fine-tuning is the conversational sentiment analysis (CSA). The Fig. 3 shows the framework of how to train this language model.

The calculation of $\mathcal{L}_{\mathrm{MLM}}$ and $\mathcal{L}_{\mathrm{NSP}}$ can refer to the original BERT, and the calculation of $\mathcal{L}_{\mathrm{CSA}}$ is the cross-entropy loss function between the sentiment probability distribution predicted by the model and the real sentiment label.

In this study, we propose a dialogue structure classification system that, although it cannot cover all possible dialogue patterns, has already covered the vast majority of common dialogue interaction forms. Our experimental results show that even when faced with some dialogue samples that do not fully conform to the predefined structures, the model can still achieve relatively good sentiment analysis performance through the extraction and generalization of structural features. This robustness is mainly due to two aspects: first, the five basic dialogue structures (linear, branch, nested, parallel, and circular) we defined have strong expressive power and can be combined as basic units to express more complex dialogue patterns; second, our model, through deep learning mechanisms, can adaptively extract structural features from dialogues, and has a certain generalization ability for unseen dialogue patterns. These results demonstrate that our proposed dialogue structure classification framework has good practicality and scalability, and can effectively support sentiment analysis tasks in real-world scenarios.

Although designing complex pre-training tasks and strategies increases model training time and computational costs, experimental results show that the performance improvements brought by this complexity are significant and valuable. Specifically, complex pre-training helps models learn richer feature representations and stronger generalization capabilities, demonstrating better performance in downstream tasks. Moreover, pre-training is a one-time investment process; compared to training separately on each downstream task, the increased time cost of complex pre-training can be amortized. Therefore, when balancing training efficiency and model performance, this increase in pre-training complexity is a reasonable and practical design choice.

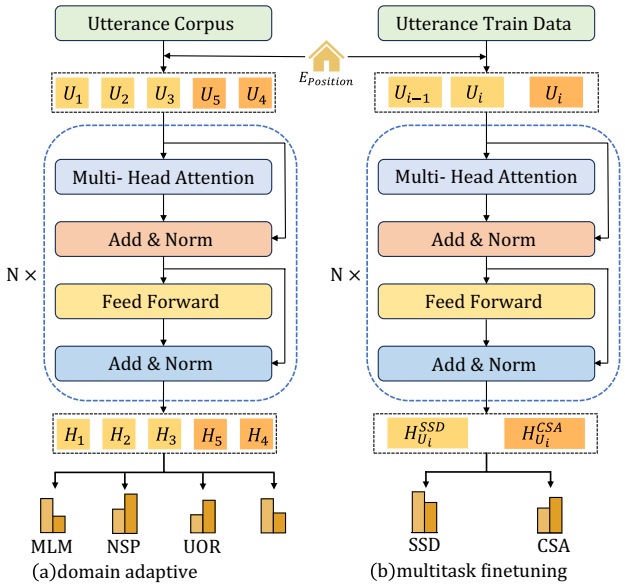

Fig. 3 Model architecture and two-stage training process, where the first stage (a)domain adaptive training primarily trains the following tasks: MLM, NSP, UOR and SBR, and the second stage (b)multitask finetuning primarily focuses on training: SSD and CSA

4 Experiment

This section is a detailed description of the experimental part.

Table 1 Sentiment label distribution in MEISD dataset

Sentiment	Train	Valid	Test
Positive	5268	691	1679
Negative	5361	770	1614
Neutral	3224	391	1019

4.1 Dataset

We conduct experiments on the dataset MEISD released by Firdaus [38]. The dialogues in the MEISD dataset come from ten different film and television works, and each utterance has its sentiment polarity label, which includes positive, negative, and neutral. Since the MEISD dataset has not been divided, we divides the entire dataset at the dialogue level. The statistics of the MEISD dataset after division are shown in Table 1. In this paper, the macro-average F1 ($F1_{macro}$) and the micro-average F1 ($F1_{micro}$) are used to measure the quality of the classification results of the model.

Our method also draws on dimensional sentiment analysis techniques [39, 40] and sentiment embeddings [41–43].

In order to analyze the model from multiple perspectives, we also test our model on the IEMOCAP [44], MELD [45], DailyDialog [46] and EmoryNLP [47] datasets.

4.2 Experiment settings

We conduct dialogue pre-training experiments on BERT with different parameter scales, and select the open-source English pre-training models $BERT_{BASE}$ and $BERT_{LARGE}$ in the Transformers library. After the dialogue pre-training designed by us, they are named $DPBERT_{BASE}$ and $DPBERT_{LARGE}$ respectively. The batch size of $DPBERT_{BASE}$ is 16, the learning rate is set to 1e-5, the batch size of $DPBERT_{LARGE}$ is 4, and the learning rate is set to 2e-6. Models are both trained for 3 epochs in the domain adaptive post-training phase, and use a warm-up learning rate of 0.1 during training. The loss function weights of the six sub-tasks are set to 1.0, i.e. $\lambda_{1-6} = 1.0$.

4.3 Baseline models

We select some baseline models for comparison with the models proposed by us:

- **TextCNN** [29]: Using pre-trained GloVe 300-dimensional word vectors, the convolution kernel size ranges from 2 to 5.
- **DPCNN** [48]: a four-layer stacked convolutional network and residual network are used to predict sentiment orientation.
- **bcLSTM** [49]: Using the $BERT_{BASE}$ with fixed model parameters to extract the semantic features at the [CLS] symbol of each segment, and then input the bidirectional LSTM and the feed-forward neural network for sentiment prediction.
- **Transformer-XL** [50]: The 18-layer Transformer-XL pre-trained language model is used for fine-tuning, and the model parameters are pre-trained with WIKITEXT-103 corpus.
- **$BERT_{BASE}$** [14]: Fine-tuned on the $BERT_{BASE}$ model.
- **$BERT_{LARGE}$** [14]: Fine-tuned on the $BERT_{LARGE}$ model.
- **XLNet** [15]: Fine-tune the model on the XLNet model.
- **DialogXL** [33]: A memory mechanism is introduced in XLNet to model dialogue context information. Since MEISD lacks dialogue subject identity information, the dialogue subject attention mechanism in the DialogXL model is replaced by a global attention mechanism.

Table 2 Experimental results of sentiment polarity prediction

Dataset	Model	Recall	Precision	F1	F1-micro
MEISD	TextCNN	43.32	44.06	46.80	43.69
	DPCNN	41.43	44.30	45.73	42.82
	bcLSTM	45.13	47.76	48.65	46.41
	Transformer-XL	46.22	47.66	49.49	46.93
	BERT-base	46.17	49.14	50.72	47.61
	BERT-large	48.26	52.10	51.97	50.11
	XLNet	51.25	48.55	51.58	49.86
	DialogXL	51.63	52.33	53.38	51.98
	DPBERT-base	52.54	54.98	55.35	53.73
	DPBERT-large	**53.70**	**55.66**	**56.86**	**54.66**
IEMOCAP	DPBERT-base	57.06	61.70	59.30	\
	DPBERT-large	63.27	61.45	62.35	\
MELD	DPBERT-base	59.18	57.05	58.10	\
	DPBERT-large	60.93	62.10	61.50	\
DailyDialog	DPBERT-base	53.68	50.44	52.01	\
	DPBERT-large	51.47	55.96	53.62	\
EmoryNLP	DPBERT-base	32.11	31.64	31.88	\
	DPBERT-large	32.23	36.36	34.17	\

4.4 Experiment results

The experimental results of the comparison between the proposed DPBERT model and the baseline model are shown in Table 2.

Due to CNN only modeling utterance information within the window without context information, TextCNN and DPCNN based on convolutional neural networks have a poor performance. The bcLSTM [49] model encodes utterance information with BERT, and also considers the flow of cross-utterance information in the context, so it performs better than the models that do not use a pre-trained model fine-tuning strategy.

Among the pre-trained baseline models, such as Transformer-XL, BERT, XLNet, and DialogXL, the Transformer-XL model performs the worst, which may be limited by the choice of the training corpus domain and language modeling tasks, and results in the poor performance of Transformer-XL. In the BERT model, $BERT_{LARGE}$ is 1.25% higher than

$BERT_{BASE}$ in $F1_{micro}$, and $F1_{macro}$ is 2.50% higher. It proves that $BERT_{LARGE}$ does have more powerful features under the premise of having a deeper model layer and parameter scale extraction capacity.

The model performance of XLNet is between $BERT_{BASE}$ and $BERT_{LARGE}$, which indicates the superiority of XLNet's autoregressive language modeling, but it is inferior to $BERT_{LARGE}$, which has a larger parameter scale. Based on the XLNet model, DialogXL considers the information transfer of the dialogue context, and it applies the memory mechanism and attention mechanism to obtain contextual information, and achieves the best experimental results among these fine-tuned pre-trained models.

Compared with the DialogXL model, the $DPBERT_{BASE}$ models are improved by 1.97% and 1.75% in $F1_{micro}$ and $F1_{macro}$, and the $DPBERT_{LARGE}$ model in $F1_{micro}$ and $F1_{macro}$ the improvement is 3.48% and 2.68%, which proves that the DPBERT model we proposed is indeed better than other baseline models, and can better transfer the BERT model to the dialogue text data.

Besides, $DPBERT_{BASE}$ has an improvement of 4.63% and 6.12% on $F1_{micro}$ and $F1_{macro}$ compared with $BERT_{BASE}$. The $DPBERT_{LARGE}$ model has an improvement of 4.89% and 4.55% on $F1_{micro}$ and $F1_{macro}$ compared with $BERT_{LARGE}$, which proves that the dialogue structure pre-training subtask proposed in this paper can indeed improve the dialogue sentiment modeling ability of the BERT model. In general, the model improvement of $BERT_{BASE}$ is more significant than that of $BERT_{LARGE}$, it proves that the smaller the parameter size, the better the dialogue structure pre-training ability.

4.5 Ablation study

The three sub-tasks of dialogue structure pre-training we proposed will be removed from DPBERT to verify their effectiveness, and the performance of $DPBERT_{BASE}$ and $DPBERT_{LARGE}$ models with different parameter scales is computed while still using the same model parameters and training framework. Results of ablation experiments are listed in Table 3.

Table 3 DPBERT ablation experiment results. (utterance order sorting (UOS), sentence backbone regularization (SBR) and sentiment shift detection(SSD))

Model	$F1_{micro}$	$F1_{macro}$
$DPBERT_{BASE}$	0.5535	0.5373
w/o UOS	0.5341	0.5138
w/o SBR	0.5413	0.5217
w/o SSD	0.5306	0.5122
$DPBERT_{LARGE}$	0.5686	0.5466
w/o UOS	0.5407	0.5243
w/o SBR	0.5478	0.5312
w/o SSD	0.5563	0.5498

As the ablation experiments shows, the UOS and SSD have a greater impact on the performance of the model, while the SBR model has less impact on the performance of the model. It indicates that in the task of dialogue sentiment analysis, the contextual logic of the paragraph and the trend of sentiment shift are the dominant factors, and the subject-verb-object relationship of the sentences taken into consideration by the SBR task is less helpful to the task. it explains that at the level of word granularity, it is difficult to say which words in the dialogue are important. But the SBR task can also slightly improve the model performance, because it contains contextual `[EOU]` information, it help understand contextual of the model.

Comparing the UOS and SSD subtasks, the SSD subtask of the $DPBERT_{BASE}$ model has a greater improvement, while the UOS subtask of the $DPBERT_{LARGE}$ model has a greater contribution to the performance improvement. This illustrates that the $BERT_{BASE}$ with a smaller parameter scale is not effective in dialogue sentiment analysis. The bottleneck is the lack of detection ability of sentiment shift, and the $BERT_{LARGE}$ model with large parameter scale is mainly restricted by the lack of utterance logic modeling ability, but the UOS task and the SMD task are both crucial for the DPBERT model.

4.6 Further analysis

In order to further study the influence of the number of dialogue utterances m on the DPBERT model, for the testset of MEISD, we divide dialogues with $m \leq 20$ and $m > 20$ into two small test sets. And the experimental results in the two test sets respectively are shown in Table 4.

The experimental results of DPBERT on the dialogue data with $m \leq 20$ are significantly better than those with $m > 20$, which shows that when the number of utterances of the dialogue is too long, the model's ability will be greatly reduced. The reason is when the dialogue logic and

Table 4 The effect of the number of dialogue utterances on the DPBERT model

Model	m	$F1_{micro}$	$F1_{macro}$
$DPBERT_{BASE}$	> 20	0.5254	0.4982
	≤ 20	0.5546	0.5318
$DPBERT_{LARGE}$	> 20	0.5356	0.5011
	≤ 20	0.5664	0.5575

sentiment shift of the super-long dialogue are more complicated, and this will result in insufficient ability of the model for understanding the dialogue text. Compared with $\text{DPBERT}_{\text{BASE}}$ and $\text{DPBERT}_{\text{LARGE}}$, $\text{DPBERT}_{\text{LARGE}}$ is always better than $\text{DPBERT}_{\text{BASE}}$ in conversations with different numbers of utterances, which indicates models with larger parameter scales still have stronger modeling capabilities, and it can capture more complex dialogue information.

5 Conclusion

In order to transfer the pre-trained language model to the dialogue text, we introduces three sub-tasks: utterance order sorting, sentence backbone regularization and sentiment shift detection. Hence, the pre-trained language model can capture utterance logic, factual associations and sentiment shift. First, data-related challenges include the consistency of emotion annotation, limited dataset size, and the absence of sentiment shift labels in some domains, which may affect the model's generalization ability. Second, the model design faces challenges in balancing multiple pre-training tasks (MLM, NSP, UOS, SBR, and SSD) and their relevance to sentiment analysis. The dialogue structure modeling might be oversimplified, potentially underutilizing contextual information. Third, the two-stage training strategy might lead to catastrophic forgetting, and the effectiveness of domain adaptation heavily relies on the similarity between source and target domains. Additionally, the model's robustness and interpretability need further investigation. In practical applications, challenges such as handling informal language, noise in real-world conversations, and computational resource requirements need to be addressed.

In future work, more complex dialogue structures can be explored to add into the pre-training task. For example, information such as phrases, entities, and person names in the utterance is also very important in the dialogue, and the correlations of these information are also part of the dialogue structure information. How to model the logical interaction and implicit sentiment change information in the dialogue are very critical for the task of dialogue sentiment analysis. At the same time, for dialogue texts with a long number of utterances, it is necessary to further improve the abilities of our model, such as understanding the long-distance logical dependencies and sentiment interactions.

Funding This work is partially supported by grants from the Key R\&D Projects in Liaoning Province award numbers (2023JH26/10200015), the Natural Science Foundation of China award numbers (62376051, 62366040, 62076046, 62066044, 61976036, 61702080) and the Fundamental Research Funds for the Central Universities award number (DUT24LAB123).

Data availability The data that support the findings of this study are public. It is available online.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval This article contains no studies with human participants or animals performed by any of the author.

References

1. Huddar, M.G., Sannakki, S.S., Rajpurohit, V.S.: Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN. Int. J. Interact. Multim. Artif. Intell. **6**(6), 112–121 (2021). https://doi.org/10.9781/IJIMAI.2020.07.004
2. Isnain, A.R., Marga, N.S., Alita, D.: Sentiment analysis of government policy on corona case using naive bayes algorithm. IJCCS (Indonesian Journal of Computing and Cybernetics Systems) **15**(1), 55–64 (2021)
3. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)
4. Wu, C., Cao, L., Chen, J., Wang, Y., Su, J.: Modeling different effectsr of user and product attributes on review sentiment classification. Appl. Intell. **54**(1), 835–850 (2024). https://doi.org/10.1007/S10489-023-05236-6
5. Xiong, H., Yan, Z., Wu, C., Lu, G., Pang, S., Xue, Y., Cai, Q.: Bart-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction. Int. J. Mach. Learn. Cybern. **14**(9), 3243–3255 (2023). https://doi.org/10.1007/S13042-023-01831-8
6. Hosseinalipour, A., Ghanbarzadeh, R.: A novel metaheuristic optimisation approach for text sentiment analysis. Int. J. Mach. Learn. Cybern. **14**(3), 889–909 (2023). https://doi.org/10.1007/S13042-022-01670-Z
7. Jiang, W., Zhou, K., Xiong, C., Du, G., Ou, C., Zhang, J.: KSCB: a novel unsupervised method for text sentiment analysis. Appl. Intell. **53**(1), 301–311 (2023). https://doi.org/10.1007/S10489-022-03389-4
8. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Bansal, N., Pruhs, K., Stein, C. (eds.) Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007, pp. 1027–1035. SIAM, USA (2007). http://dl.acm.org/citation.cfm?id=1283383.1283494
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/JAIR.953
10. Kim, Y.: Convolutional neural networks for sentence classification. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of The ACL, pp. 1746–1751. ACL, Qatar (2014). https://doi.org/10.3115/V1/D14-1181
11. Shi, S., Zhao, M., Guan, J., Li, Y., Huang, H.: A hierarchical lstm model with multiple features for sentiment analysis of sina weibo texts. In: Tong, R., Zhang, Y., Lu, Y., Dong, M. (eds.) 2017 International Conference on Asian Language Processing, IALP 2017,

Singapore, December 5-7, 2017, pp. 379–382. IEEE, Singapore (2017). https://doi.org/10.1109/IALP.2017.8300622
12. Li, G., Wang, H., Ding, Y., Zhou, K., Yan, X.: Data augmentation for aspect-based sentiment analysis. Int. J. Mach. Learn. Cybern. **14**(1), 125–133 (2023). https://doi.org/10.1007/S13042-022-01535-5
13. Pérez-Rosas, V., Mihalcea, R., Morency, L.: Utterance-level multimodal sentiment analysis. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers, pp. 973–982. The Association for Computer Linguistics, Bulgaria (2013). https://aclanthology.org/P13-1096/
14. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, USA (2019). https://doi.org/10.18653/V1/N19-1423
15. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 5754–5764 (2019). https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html
16. Mehri, S., Razumovskaia, E., Zhao, T., Eskénazi, M.: Pretraining methods for dialog context representation learning. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 3836–3845. Association for Computational Linguistics, Italy (2019). https://doi.org/10.18653/V1/P19-1373
17. Li, L., Li, C., Ji, D.: Deep context modeling for multi-turn response selection in dialogue systems. Inf. Process. Manag. **58**(1), 102415 (2021). https://doi.org/10.1016/J.IPM.2020.102415
18. Xu, Y., Zhao, H., Zhang, Z.: Topic-aware multi-turn dialogue modeling. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 14176–14184. AAAI Press, Virtual Event (2021). https://doi.org/10.1609/AAAI.V35I16.17668
19. Zeng, H., Hong, Z., Liu, J., Wei, B.: Multi-task learning for multi-turn dialogue generation with topic drift modeling. In: Chen, L., Fernández-Manjón, B. (eds.) 2021 IEEE International Conference on Big Knowledge, ICBK 2021, Auckland, New Zealand, December 7-8, 2021, pp. 410–417. IEEE, New Zealand (2021). https://doi.org/10.1109/ICKG52313.2021.00061
20. Zhang, Z., Li, J., Zhu, P., Zhao, H., Liu, G.: Modeling multi-turn conversation with deep utterance aggregation. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 3740–3752. Association for Computational Linguistics, USA (2018). https://aclanthology.org/C18-1317/
21. Zhang, Z., Yang, J., Zhao, H.: Retrospective reader for machine reading comprehension. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 14506–14514. AAAI Press, Virtual Event (2021). https://doi.org/10.1609/AAAI.V35I16.17705
22. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S.M., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR arXiv:2303.12712 (2023) https://doi.org/10.48550/ARXIV.2303.12712
23. Anil, R., Borgeaud, S., Wu, Y., Alayrac, J., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T.P., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P.R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., al.: Gemini: A family of highly capable multimodal models. CoRR arXiv:2312.11805 (2023) https://doi.org/10.48550/ARXIV.2312.11805
24. Filippini, C., Spadolini, E., Cardone, D., Bianchi, D., Preziuso, M., Sciarretta, C., Cimmuto, V., Lisciani, D., Merla, A.: Facilitating the child-robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot. Int. J. Soc. Robotics **13**(4), 677–689 (2021). https://doi.org/10.1007/S12369-020-00661-W
25. Chen, R., Wang, J., Yu, L., Zhang, X.: Learning to memorize entailment and discourse relations for persona-consistent dialogues. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pp. 12653–12661. AAAI Press (2023). https://doi.org/10.1609/AAAI.V37I11.26489
26. Xu, R., Tao, C., Jiang, D., Zhao, X., Zhao, D., Yan, R.: Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 14158–14166. AAAI Press, Virtual Event (2021). https://doi.org/10.1609/AAAI.V35I16.17666
27. Zhang, Z., Zhao, H.: Structural pre-training for dialogue comprehension. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pp. 5134–5145. Association for Computational Linguistics, Virtual Event (2021). https://doi.org/10.18653/V1/2021.ACL-LONG.399
28. Medsker, L.R., Jain, L.: Recurrent neural networks. Design and Applications **5**, 64–67 (2001)
29. Hsu, C., Chen, S., Kuo, C., Huang, T.K., Ku, L.: Emotionlines: An emotion corpus of multi-party conversations. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA), Japan (2018). http://www.lrec-conf.org/proceedings/lrec2018/summaries/581.html
30. Chatterjee, A., Narahari, K.N., Joshi, M., Agrawal, P.: Semeval-2019 task 3: Emocontext contextual emotion detection in text. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M.,

Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, pp. 39–48. Association for Computational Linguistics, USA (2019). https://doi.org/10.18653/V1/S19-2005
31. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of The ACL, pp. 1532–1543. ACL, Qatar (2014). https://doi.org/10.3115/V1/D14-1162
32. Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L., Zimmermann, R.: Conversational memory network for emotion recognition in dyadic dialogue videos. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pp. 2122–2132. Association for Computational Linguistics, USA (2018). https://doi.org/10.18653/V1/N18-1193
33. Shen, W., Chen, J., Quan, X., Xie, Z.: Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 13789–13797. AAAI Press, Virtual Event (2021). https://doi.org/10.1609/AAAI.V35I15.17625
34. Zhu, Z., Cheng, X., Huang, Z., Chen, D., Zou, Y.: Enhancing code-switching for cross-lingual SLU: A unified view of semantic and grammatical coherence. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pp. 7849–7856. Association for Computational Linguistics, Singapore (2023). https://aclanthology.org/2023.emnlp-main.486
35. Wu, H., Xu, K., Song, L.: Structure-aware dialogue modeling methods for conversational semantic role labeling. IEEE ACM Trans. Audio Speech Lang. Process. **32**, 742–752 (2024). https://doi.org/10.1109/TASLP.2023.3331576
36. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 2787–2795 (2013). https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html
37. Zhang, Z., Fang, M., Ye, F., Chen, L., Namazi-Rad, M.: Turn-level active learning for dialogue state tracking. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pp. 7705–7719. Association for Computational Linguistics, Singapore (2023). https://aclanthology.org/2023.emnlp-main.478
38. Firdaus, M., Chauhan, H., Ekbal, A., Bhattacharyya, P.: MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pp. 4441–4453. International Committee on Computational Linguistics, Spain (2020). https://doi.org/10.18653/V1/2020.COLING-MAIN.393
39. Xie, H., Lin, W., Lin, S., Wang, J., Yu, L.: A multi-dimensional relation model for dimensional sentiment analysis. Inf. Sci. **579**, 832–844 (2021). https://doi.org/10.1016/J.INS.2021.08.052
40. Lee, L., Li, J., Yu, L.: Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. ACM Trans. Asian Low Resour. Lang. Inf. Process. **21**(4), 65–16518 (2022). https://doi.org/10.1145/3489141
41. Yu, L., Wang, J., Lai, K.R., Zhang, X.: Refining word embeddings using intensity scores for sentiment analysis. IEEE ACM Trans. Audio Speech Lang. Process. **26**(3), 671–681 (2018). https://doi.org/10.1109/TASLP.2017.2788182
42. Yu, L., Wang, J., Lai, K.R., Zhang, X.: Refining word embeddings for sentiment analysis. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pp. 534–539. Association for Computational Linguistics (2017). https://doi.org/10.18653/V1/D17-1056
43. Wang, J., Yu, L., Zhang, X.: Softmcl: Soft momentum contrastive learning for fine-grained sentiment-aware pre-training. In: Calzolari, N., Kan, M., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, pp. 15012–15023. ELRA and ICCL (2024). https://aclanthology.org/2024.lrec-main.1305
44. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Evaluation **42**(4), 335–359 (2008). https://doi.org/10.1007/S10579-008-9076-6
45. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. CoRR arXiv:1810.02508 (2018)
46. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. In: Kondrak, G., Watanabe, T. (eds.) Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers, pp. 986–995. Asian Federation of Natural Language Processing (2017). https://aclanthology.org/I17-1099/
47. Zahiri, S.M., Choi, J.D.: Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In: The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Technical Report, vol. WS-18, pp. 44–52. AAAI Press (2018). https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16434
48. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 562–570. Association for Computational Linguistics, Canada (2017). https://doi.org/10.18653/V1/P17-1052
49. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.: Context-dependent sentiment analysis in user-generated videos. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 873–883. Association for Computational Linguistics, Canada (2017). https://doi.org/10.18653/V1/P17-1081
50. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. In: Korhonen, A., Traum, D.R., Màrquez, L.

(eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers, pp. 2978–2988. Association for Computational Linguistics, Italy (2019). https://doi.org/10.18653/V1/P19-1285